BW-CAR | SINCOM

SYMPOSIUM ON INFORMATION AND COMMUNICATION SYSTEMS

1. Baden-Württemberg Center of Applied Research
Symposium on
Information and Communication Systems

SInCom 2014

12. December 2014
in Furtwangen

ISBN 978-3-00-048182-6

HOCHSCHULE FURTWANGEN UNIVERSITY | HFU

Message from the Program Chairs

The organization Baden-Württemberg Center of Applied Research (BW-CAR) intends to further develop the applied research at the Universities of Applied Science (UAS). In BW-CAR outstanding university transversal research competences will be bundled, young scientists will be promoted, new research areas will be developed and research services at the UAS will be made more visible.

The BW-CAR working group "Informations- und Kommunikationssysteme" (IKS) in cooperation with the working group "Technologien für Intelligente Systeme" organized the 1st BW-CAR Symposium on Information and Communication Systems (SInCom) at the University Furtwangen, a university of applied science. The IKS working group represents various disciplines and application areas with expertise for different aspects of communication technologies, communication methods and procedures for processing in information systems. The collection, analysis, processing, transfer and output of information are the key technologies of information and communication systems, which in turn are inherent components of any modern technical system. These key technologies are found in plants, mobile networks, smart grid, navigation systems, "Industrie 4.0", ambient assisted living, environmental engineering in macroscopic or in embedded systems such as smart phones, sensor networks, instrumentation, smart meters to smart sensors on a microscopic scale.

The SInCom 2014 aimed at young researchers for contribution in the field of

- Distributed Computing
- Communication Networks
- Algorithms and Signal Processing

Many thanks to all for the good contributions for the 1st BW-CAR Symposium on Information and Communication Systems (SInCOM). An expression of gratitude to the reviewers for their suggestions for improvement. Without them no high-quality conference proceedings could have been achieved. Finally we like to thank the Furtwangen University for the support during the symposium.

Furtwangen, 12.12.2014
Prof. Dr. Dirk Benyoucef and Prof. Dr. Christoph Reich

Organizing Committee

Prof. Dr. Dirk Benyoucef, Hochschule Furtwangen

Prof. Dr. Christoph Reich, Hochschule Furtwangen

Program Committee

Prof. Dr. Dirk Benyoucef, Hochschule Furtwangen

Prof. Dr.-Ing. Andreas Christ, Hochschule Offenburg

Prof. Dr. rer. nat. Thomas Eppler, Hochschule Albstadt-Sigmaringen

Prof. Dr.-Ing. Jürgen Freudenberger, Hochschule Konstanz

Prof. Dr. Thomas Greiner, Hochschule Pforzheim

Prof. Dr.-Ing Reiner Jäger, Hochschule Karlsruhe

Henrik Kuijs M.Sc., Hochschule Furtwangen

Prof. Dr. rer.nat. Roland Münzer, Hochschule Ulm

Prof. Roy Oberhauser, Hochschule Aalen

Prof. Dr. Albrecht Oehler, Hochschule Reutlingen

Prof. Dr.-Ing. Franz Quin, Hochschule Karlsruhe

Prof. Dr. Christoph Reich, Hochschule Furtwangen

Thomas Rüebsamen M.Sc., Hochschule Furtwangen

Prof. Dr. Peter Väterlein, Hochschule Esslingen

Prof. Dr. Dirk Westhoff, Hochschule Furtwangen

Content

Algorithm and Architecture for Semi-Global Stereo Matching and Depth Calculation of Road Scenes

Frank Schumacher, Thomas Greiner
Merses Center for Applied Resarch
Pforzheim University
Pforzheim, Germany
{frank.schumacher | thomas.greiner}@hs-pforzheim.de

***Abstract*—Semi-Global Matching (SGM) is currently one of the most promising approaches to perform stereo correspondence of road scenes in real-time. While local stereo algorithms cannot cope with textureless areas and produce many errors, global approaches have long and often not deterministic runtimes. The latter property is important e.g. for autonomous vehicles, because automotive systems require guaranteed reaction times. This can be accomplished by the regular computation structure of semi-global matching. Additionally, SGM can be realized as a high speed FPGA architecture to achieve high data throughput. In this work, we present a memory efficient cost computation and aggregation scheme for SGM. We evaluate our approach, using road scene stereo images and groundtruth data of the challenging Kitti[1] vision benchmark. The disparity maps show competitive quality results. Finally, they are used to extract actual depth values by using calibrated and rectified projection matrices, provided by the Kitti dataset.**

***Index Terms*—Stereo correspondence, semi-global matching, depth from disparity, autonomous driving**

I. Introduction

Determination of the stereo correspondence on real world images, e.g. road scenes, is more challenging than for artificial settings. The main reason is the presence of large homogenous areas, either with slanted planes such as the road surface, or without an actual depth such as the sky. This makes local stereo approaches fail because their local window sizes are too small to overcome those textureless areas. Global approaches however, reach acceptable quality results, but tend to require many iterations and long optimization passes. This is not tolerable for automotive applications, which often request a fast reaction time from input to output and additionally a frame rate of at least 20-25 frames per second.

Drawing advantages from both, Semi-Global Matching (SGM) [3] provides good quality results and potentially a high, deterministic data throughput. Its basic idea is to use locally computed matching costs of all possible correspondence candidates within the maximum disparity range to form a cost volume. Then, the matching costs of either all pixel pairs in the entire image area (dual pass aggregation) or the costs of pixel pairs of one image row (single pass) are aggregated in a SGM cost buffer along linear image paths. Chosen properly, the paths approximate global image properties and thus, homogenous areas can be handled. After aggregation, the pixel pair with the lowest aggregated matching cost is selected and the disparity is calculated as difference of their horizontal displacement. The number, lengths and directions of the paths can be varied to balance the accuracy, memory consumption and runtime of an SGM architecture.

The SGM cost volume and aggregation buffer are very large, particularly for dual pass parallel implementations, because these require simultaneous read/write operations at different memory locations. Therefore, either expensive chip devices with large on-board memory are used [4] or the external storage of the data values in an off-chip memory [5]. The latter requires interfacing and forms a throughput bottleneck in most systems, complicates pipelined raster scan processing and necessitates the integration of hard- and software modules to provide the physical connection and access patterns to the memory.

In this work, we describe an algorithm and FPGA architecture of a SGM stereo image processing implementation with Hough transform based disparity pre-selection and Census transform refinement. Due to disparity pre-selection and raster scan processing, the matching costs can be computed on the fly for each pixel. Additionally, the cost aggregation buffer size can be reduced and the system requires no external memory. To conclude, we show how to extract absolute depth values from the relative disparities, calculated for a real world road scene example of the Kitti stereo vision benchmark using calibration and rectification data.

II. Related work

The Kitti benchmark table [1], [2] lists various SGM variants, which achieve good robustness and quality. Currently, the lowest error rate of all SGM approaches achieves weighted SGM (wSGM) [6]. It combines SGM with the concept of adaptive support weights, one of the latest local matching cost approaches [7]. The weights, which are added to the path costs, depend on normal vectors of pixel patches with constant disparities. By this, the weights penalize false disparity changes and promote connected object surface pixels. The drawback of wSGM is a relatively long and non-deterministic runtime due to an irregular processing scheme.

[1]The Kitti vision benchmark suite [1], [2] provides stereo image pairs, taken in and around the city of Karlsruhe with a calibrated stereo camera setup, mounted on a recording platform (test car).

Currently, the fastest SGM approach with competitive quality results in the Kitti table is called rapid SGM (rSGM) [8] and trimmed for efficiency on multicore CPUs. It reaches high data throughput by cost volume compression and coarse plus fine grain parallelism. The cost volume reduction is achieved with disparity sub-sampling and a frame rate of 16 Hz is reached for VGA images at a disparity range of 128. This is not yet fast enough to process Kitti image pairs in real-time, but the cost volume reduction is a promising concept for speeding up SGM while keeping memory demand on an acceptable level.

To enable high speed parallel processing in FPGAs, the computation scheme must be kept as regular as possible. Hence, simple and regular pixel operations such as local matching cost computation are simple to realize in hardware. More complex functions, necessary for example to compute adaptive weights from surface normals of slanted planes or for compression schemes such as in rSGM, are mostly difficult to implement with FPGA logic, however. Therefore, architectures implementing SGM, use simple parallel path cost aggregation and the full size cost volume and cost buffer. These cost aggregation buffers are too large for most reasonable FPGA devices and must externally be buffered, for example in DRAM. Therefore, a memory controller is required to access the external RAM sequentially, which forms data rate bottlenecks in most FPGA designs. Hence, FPGA architectures for SGM, capable of processing real world scenes are still rare. In [5], the first FPGA SGM implementation is reported. It utilizes two image passes due to 8 SGM path directions. The aggregation cost buffers are stored in external RAM and the architecture achieves 25 frames per second (fps) at a resolution of 320×200 and a disparity range of $D = 64$ per pass, i.e. 680×400 images are processed with $D = 128$. [4] presented an improved architecture, achieving 30 fps at VGA resolution, but only with single pass SGM using 4 paths. They describe a configurable systolic array structure with adjustable rates of parallelism. To process e.g. 5 image pixels and 2 disparities simultaneously, 240 kByte Block-RAM for the cost volume and additionally 220 kByte for the aggregation buffer are necessary. Other architectures require even more memory [9] but their throughput is only slightly increased.

III. Proposed Algorithm

To overcome the drawback of the huge memory demand of regular SGM approaches, we propose the reduction of the cost volume and aggregation buffer using a two stage combination of a feature and an area based approach. This is followed by a single pass, four path SGM cost aggregation similar to [4]. Without sacrificing quality, decrease of memory footprint of up to 7 times can be achieved compared to uncompressed SGM while Kitti images with size 1242×375 and a disparity range of $D = 160$ are processed.

A. Feature based pre-selection and area based refinement

Fig. 1 shows a block diagram of our system. To find coarse disparity matching candidates in the stereo image pair, the absolute differences between pixel pairs are computed and used as features in a generalized Hough transform inspired accumulation scheme. By this, the best disparity candidates are extracted and pixel pairs with too high matching costs are dismissed. The results form coarse disparity blocks and a candidate index vector (CIV) whose entries point to the best disparity candidate pixel pairs. By using this simple features, the approach reaches very high data throughput, but achieves a comparable low disparity quality [10].

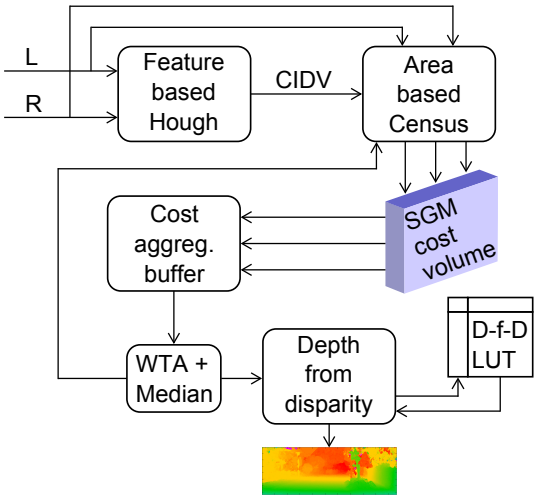


Fig. 1. Main blocks of algorithm including coarse disparity extraction with feature based Hough transform, area based Census transform for refinement of disparities, indexed by candidate index vector (CIV) and semi-global matching (SGM). The Look-Up table based mapping of disparities to depth values is explained in section V.

To enhance the quality of the disparities, which are found in the first stage, we utilize a second stage to apply the area based Census transform (CT) [11]. This transform is robust against illumination variations and shows good properties when combined with the Hamming distance as cost function in stereo correspondence. For efficiency purposes, not all pixel correspondence candidates are refined in our approach, but only those, which are pre-selected in the Hough transform stage. To account for possible inaccuracies at discontinuities and block borders, evaluations showed an increase of quality, when each pre-selected disparity is expanded in a small range of $\pm\delta$ to $2\delta + 1$ values. This technique allows the realization of an FPGA architecture with high data throughput and competitive disparity map quality for the Middlebury images [12]. However, the approach is a local one and as such, not suitable for real world scenes containing large homogenous areas. Its quality result for the Kitti dataset is not sufficient, see entry *HT+CT* in Table I. Therefore, we propose the usage the pre-selection Hough transform method to construct an efficient Census based cost volume and aggregation buffer for a semi-global matching approach.

B. Memory reduction for semi-global matching

In general, each location of the SGM path cost aggregation buffer $S(p, d)$ contains the summed up path costs $L_r(p, d)$ for

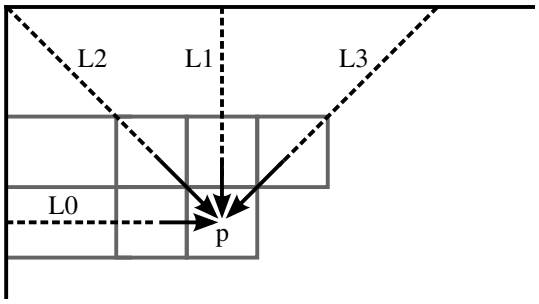


Fig. 2. SGM example with four paths.

a specific pixel p at a disparity d with r as path indicator, i.e. $S(p,d) = L_0 + L_1 \dots L_r$. Illustrated in Fig. 2, the current path cost value $L_r(p,d)$ for a specific pixel p is recursively computed from the minimum of the costs from previous pixel $p-1$ along that path r and from the cost volume entry $C(p,d)$ of p at disparity d, i.e. $L_r(p,d) = C(p,d) + min[L_r(p-1,d), L_r(p-1,d+1)+P1, L_r(p-1,d-1)+P1, L_r(p-1,d=1 \dots D) + P2]$. The constant P1 penalizes small disparity variations (± 1) and accounts for slanted planes. P2 is larger than P1 and adds a penalty to preserve disparity jumps at discontinuities. The cost volume C and path cost aggregation buffer S can therefore be seen as 3D data structures, which store matching and aggregated costs for each image pixel pair. When finally all path costs for one pixel are aggregated, the lowest accumulated cost value is selected in a Winner-Takes-All (WTA) scheme and the corresponding disparity is computed as difference of the horizontal pixel coordinates. As final refinement, the disparities are processed with a small median filter to smooth the result while preserving disparity jumps at discontinuities.

To fill the cost volume, matching costs are mostly computed over the entire possible disparity range D, i.e. the cost volume entry for pixel coordinate (x_i, y_i) contains the local matching costs between left pixel $p_L(x_i, y_i)$ and right pixel range $p_R(x_i - D \dots x_i, y_i)$. Hence, this cost volume $C(x,y,d)$ must store $w_I \cdot h_I \cdot D$ costs of the entire image with size $w_I \times h_I$ when SGM is performed in dual pass, i.e. when paths originate before and after the current pixel. When paths come solely from behind the current pixel as e.g. in Fig. 2, single pass processing can be realized. In this case, only $w_I \cdot D$ costs of one image row must intermediately be stored. Especially dual pass variants require large intermediate buffer structures, but also the storage of all correspondence candidates in a single row means a significant effort.

We used the disparity pre-selection technique already to reduce the computational load for the calculation of the area based Census transform. We propose to utilize the same method to reduce the cost volume and aggregation buffer size in a single pass SGM approach. To guarantee the SGM algorithm integrity, i.e. to provide a cost value for each (x,d) in the current image line, we propose the cost determination method

$$C(x,d) = \begin{cases} C'(x,d), & \text{if } d \in C'(x,d) \\ P2, & \text{otherwise} \end{cases} \quad (1)$$

with reduced cost volume C'. If, for a specific (x,d) combination, a cost value is contained in C', this vale is returned. Otherwise, the saturation constant value $P2$. Hence, C' contains those matching cost values computed on the fly for a stereo pixel-pair, which have been pre-selected in the Hough stage, expanded and refined in the Census stage. The size of the reduced cost volume is $|C'| = K = (\kappa_1 + \kappa_2) < D$. Parameter κ_1 represents the expansion width of the disparity candidates and hence, the length of the candidate index vector of the first stage. It determines the refinement of the pre-selected coarse disparity values and reflects the threshold for sensitivity and dismissing of disparity candidates. Parameter κ_2 is the feedback range of previous disparities: while calculating the disparity for a pixel $p(x,y)$, the disparity of the pixel $p(x,y-1)$ of the previous row is also taken into account and expanded with $\pm\frac{\kappa_2}{2}$. The selection of parameters κ_1, κ_2 determines the cost buffer reduction and also the quality result of the stereo correspondence approach.

Setting κ_1 and κ_2 properly can reduce the sizes but also achieve a competitive quality result, balanced according to the target application, see Table I, which shows error percentages of results for Kitti dataset *training* images according to the error metric in [13] for error threshold $\tau = 3$. The following approaches are compared:

- SAD: Sum of absolute differences used in OpenCV [14]
- ZSAD: Zero-mean SAD with SGM [5]
- SGBM: Semi-global block matching used in OpenCV
- CT: SGM as in [5] but instead of ZSAD cost volume, a 9×9 Census transform is used. Compared to our approach, all cost values are computed and aggregated.
- *Prop:* Proposed algorithm using pre-selected disparities. Parameters: 11×11 Census transform, 3×3 median filter, cost volume size $|C'| = K = \kappa_1 + \kappa_2 = 9 + 13 = 22$ covering a disparity range of $D = 160$.

Our approach outperforms other published versions in terms of disparity quality. Hence, the disparity dismissing scheme of the first stage sorts wrong values out, which contribute to erroneous values in other approaches. For the FPGA architecture realization, Table II shows also a clear reduction of memory occupation, compared to other published approaches.

IV. Architecture

The stages of the algorithm are realized as FPGA hardware modules. Main focus is on high data throughput, low latency and efficient memory usage. The latter requires the aforementioned cost and aggregation buffer reduction. While the algorithm calculates disparities for Kitti images in the range of $D = 1 \dots 160$, the FPGA architecture uses the reduction scheme and hence, instead of D candidate values per pixel, only $K = 22$ values are required, which is a reduction of more than 7 times. Practically, the cost volume can be implemented

as 2D-look-up tables for each pixel in the current row, obeying the single pass rule: Each image pixel is processed only once. One of such table contains the matching costs between a left pixel x_i and right pixels displaced by disparity $d = 1 \dots D$. Using the reduction scheme, the size of a look-up table is K, see Fig. 3 as example. Additional FPGA logic is required to check the input index d_n and to deliver either a look-up value if $d_n \in C'$ or $P2$ according to Eq. 1.

The same technique can be used to implement the SGM cost aggregation buffer. Here, the aggregated costs along the SGM paths are stored for each pixel and for each possible disparity candidate. Using the reduction scheme, we achieve a lower memory consumption compared to other published approaches: [4] requires 240 kByte for cost volume and 220 kByte for aggregation buffer. [9] requires in total 1250 kByte of memory. Our approach uses 54.6 kByte for cost values and 80 kByte for aggregation buffer.

The architecture processes one input pixel per clock cycle in a raster scan manner from top left to bottom right. Both, the feature based Hough transform and Census transform use pixel pipelining. While the Hough transform operates on a single image row, the Census transform is a neighborhood operation, which uses a 2D-window. For efficient realization, we utilize row buffers as FIFOs in FPGA on-chip block-RAM [15].

Using all the presented efficiency measures, our architecture achieves a data throughput of 199 frames per second (Fps) for Kitti images with size $w_I \times h_I$ = 1242 × 375 pixels.

TABLE I
COMPARISON OF ERROR RATES OF DIFFERENT APPROACHES FOR GROUNDTRUTHS "ALL" AND "NON-OCCLUDED". AS IMAGE, KITTI TRAINING DATASET SCENE *000158* IS SELECTED. ALSO THE AVERAGE ERRORS OF ALL 194 KITTI TRAINING IMAGES ARE SHOWN.

	0-193		**000158**	
	All	Noc	All	Noc
HT+CT	*34.6*	*33.1*	*26.18*	*23.94*
SAD	27.1	25.4	18.79	16.3
ZSAD (SGM)	23.27	21.48	16.62	14.06
SGBM	20.73	18.88	15.13	12.52
CT (SGM)	20.6	18.73	15.73	13.14
Prop.	*18.06*	*16.15*	*13.94*	*11.3*

TABLE II
COMPARISON OF HARDWARE RESOURCE OCCUPATION, MEMORY DEMAND AND REACHABLE FRAMES PER SECOND.

Appr.	Fps	$w_I \times h_I$	D	LUTs	RAM
[5]	25	680×400	128	60,000	150 kByte (a)
[4]	30	640×480	128	20,220	240 kByte (b)
[9]	30	1024×768	96	116,268	1250 kByte (c)
Prop.	199	1242×375	160	109,072	(54.6 + 80) kByte (d)

(a) External DDR RAM for cost aggregation buffer required.
(b) Cost aggregation buffer requires additional 220 kByte, either on-chip or off-chip RAM. LUT-occupation is relatively low, because only FPGA system infrastructure and cost calculation unit is regarded. 2 disparities and 5 pixels are computed simultaneously.
(c) 8 disparities and 2 pixels are computed in parallel.
(d) 54.8 kByte for cost volume and approx. 80 kByte for SGM aggregation buffer. D=160 disparities, reduced to K=22 are processed in parallel.

V. DEPTH FROM DISPARITY

A. Mapping from pixels to meters

To obtain corresponding depth values Z_n from stereo image pair disparities d_n in a standard stereo setup as illustrated in Fig. 4, the depth triangulation equation can be used:

$$Z_n = \frac{b_x \cdot f}{d_n} \tag{2}$$

with b_x as baseline distance and f as focal length of the camera system. Normally, f is a constant for a camera - lens setup and given in the unit mm. Depending on the remaining parameter dimensions, a conversion factor, computed from chip pixel size and resolution must be utilized to come from pixel units to meters. However, the Kitti stereo cameras use varifocal lenses with a focus range of 4-8 mm. Hence, there is no focal length parameter as stereo system constant. The focus value differs from image pair to image pair. As remediation, the Kitti dataset provides detailed calibration data for the camera setup [2] and the rectified projection matrices $P^i \in \mathbb{R}^{3\times4}$ obtained from calibration procedures before a test drive and for each stereo image pair. This can be used to compute the depth from disparity for all image pairs. With the relation $y_n^i = P^i x_n$, 3D-points $x_n = (X_n, Y_n, Z_n, 1)^T$ in homogenous camera coordinates can be projected to points $y_n^i = (u_n^i, v_n^i, 1)^T$ in the respective camera image planes. The matrix P^i contains the rectified intrinsic camera parameters [16]:

$$P^i = \begin{pmatrix} f_u^i & 0 & c_u^i & -f_u^i \cdot b_x \\ 0 & f_v^i & c_v^i & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{3}$$

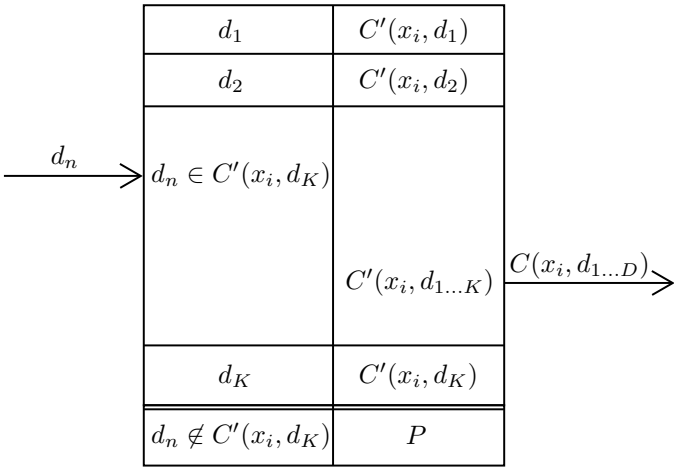


Fig. 3. Example for a reduced cost volume C', which contains K valid correspondence costs for disparity candidates, indicated by candidate index vector from Hough stage. The volume is realized with a look-up table. If a candidate d_n is contained in the table, the corresponding value is put out. If not, a default penalty value P is put out. By this, the entire cost volume function C can be emulated, providing cost values for all disparities $d = 1 \dots D$ with a table size $K + 1 < D$.

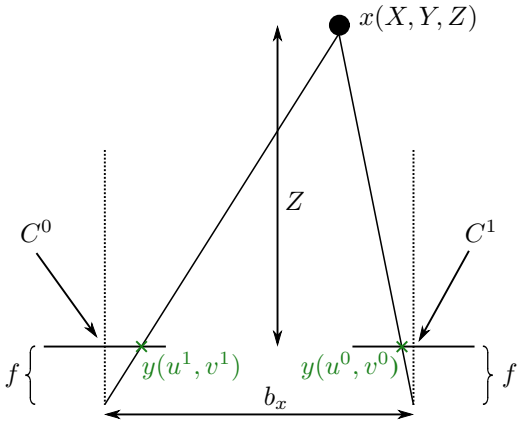


Fig. 4. Standard stereo setup with identical cameras, C_0, C_1 as center of camera coordinate systems, parallel optical camera axes, f as focal length, b_x as baseline distance and y^0, y^1 as mapping of real world point $x(X,Y,Z)$ to left and right camera image with epipolar constraint $v^0 = v^1$ and corresponding disparity $d = |u^0 - u^1|$.

with $i = 0$ for the left and $i = 1$ for the right camera. The left camera image plane center is defined as coordinate system origin. Therefore, the horizontal distance of the stereo camera system – the baseline b_x – is 0 in P^0 and 0.54 m in P^1. To put disparity and depth values into relation, a point, which lays on the optical axis of the left camera, can be projected to both images: $x = (X, Y, Z, 1)^T$ with $X = Y = 0$. With the projection matrices P^0 and P^1 for the left and right camera, $y^{0,1} = P^{0,1}X$ is

$$y^0 = \begin{pmatrix} u^0 \\ v^0 \\ 1 \end{pmatrix} = P^0 x = \begin{pmatrix} c_u^0 \cdot Z \\ c_v^0 \cdot Z \\ Z \end{pmatrix} \cdot \frac{1}{Z} \tag{4}$$

and

$$y^1 = \begin{pmatrix} u^1 \\ v^1 \\ 1 \end{pmatrix} = P^1 x = \begin{pmatrix} c_u^1 \cdot Z - f_u^1 \cdot b_x \\ c_v^1 \cdot Z \\ Z \end{pmatrix} \cdot \frac{1}{Z} \tag{5}$$

Using the epipolar constraint $v^0 = v^1$, the rectification properties $f_{u,v}^0 = f_{u,v}^1$, $c_{u,v}^0 = c_{u,v}^1$ and the fact, that a disparity is the horizontal distance $d = u^0 - u^1$, d becomes

$$d = u^0 - u^1 = \frac{f_u \cdot b_x}{Z} \text{ and } Z = \frac{f_u \cdot b_x}{d} \tag{6}$$

As d and f_u are given in pixels and the unit of b_x is meters, the unit for Z becomes automatically meters. This shows, that Eq. (2) holds also, using the calibrated focal length f_u in u-direction, given in pixels. This value may vary from image to image but the KITTI dataset provides the projective matrix values for each image set. Knowing this, a table can be filled to look-up the depth value from all possible disparity values in $O(1)$ time. As length of the table, D entries are enough, because the stereo correspondence algorithm uses a maximum disparity range of D.

B. Results

Exemplarily, the depth map for the KITTI dataset training image pair *000158* is composed, see Fig. 5. For this, the

Fig. 5. Rectified Kitti stereo vision benchmark example image pair *000158* of dataset *training*. Top is the left image and bottom the right.

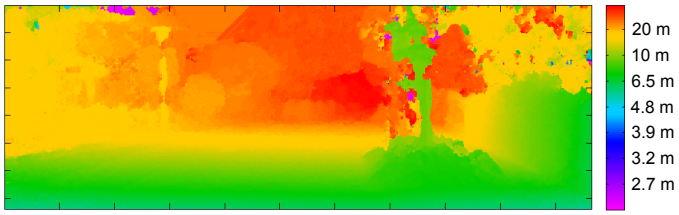


Fig. 6. Colored mapping of disparity to depth values for KITTI training image pair *000158*.

following rounded non-zero parameters for P_0 are extracted from the calibration files: $f_u = f_v = 707$, $c_u = 601$, $c_v = 183$. For P_1: $f_u = f_v = 707$, $c_u = 601$, $c_v = 183$ and $-f_u \cdot b_x = -379$. All values are given in pixels, except the baseline. It is given in meters and has the value $b_x = 0.54\,m$ for the KITTI setup. Using the values of matrices P_0, P_1 and Eq. (6), the depth from disparity can be calculated. The corresponding depth map for image pair *000158* is shown in Fig. 6 including the color-to-distance mapping. The theoretical disparity range of $1 \ldots 160$ maps to a distance of 390m to 2.4m, but as illustrated the depth map, the majority of significant depth values are located between 5m and 20m.

VI. Conclusion

For real world scenes, semi-global stereo matching provides the best tradeoff between accuracy and potential data throughput. The regular scheme of the approach makes it suitable for real-time implementations which are required in most autonomous driving applications. We provide an algorithm for the realization in a high data throughput FPGA architecture. The calculation of actual depth values could be used to provide telemetry data to environment sensing systems such as traffic light and road sign recognition or to the driver itself as obstacle overlay projected on the windshield.

Acknowledgment

The work was financially supported by the European Union (EFRE) and State of Baden-Wuerttemberg, Ministry of Science, Research and Arts, Germany, in progress of research project MERSES.

REFERENCES

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[2] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets Robotics: The KITTI Dataset," *Int'l. Journal of Robotics Research (IJRR)*, 2013.

[3] H. Hirschmüller, "Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2005, pp. 807–814.

[4] C. Banz, S. Hesselbarth, H. Flatt, H. Blume, and P. Pirsch, "Real-time stereo vision system using semi-global matching disparity estimation: Architecture and FPGA-implementation," in *Proc. of the 2010 Int'l. Conf. on Embedded Computer Systems (SAMOS)*, 2010, pp. 93–101.

[5] S. Gehrig, F. Eberli, and T. Meyer, "A Real-Time Low-Power Stereo Vision Engine Using Semi-Global Matching," in *Proc. of the 7th Int'l. Conf. on Computer Vision Systems*, ser. LNCS, M. Fritz, B. Schiele, and P. H. Justus, Eds., vol. 5815. Springer Berlin Heidelberg, 2009, pp. 134–143.

[6] R. Spangenberg, T. Langner, and R. Rojas, "Weighted Semi-Global Matching and Center-Symmetric Census Transform for Robust Driver Assistance," in *Computer Analysis of Images and Patterns*, ser. LNCS, R. Wilson, E. Hancock, A. Bors, and W. Smith, Eds. Springer Berlin Heidelberg, 2013, vol. 8048, pp. 34–41.

[7] A. Hosni, M. Bleyer, and M. Gelautz, "Secrets of adaptive support weight techniques for local stereo matching," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 620–632, 2013.

[8] R. Spangenberg, T. Langner, S. Adfeldt, and R. Rojas, "Large Scale Semi-Global Matching on the CPU," in *Intelligent Vehicles Symposium*, 2014.

[9] W. Wang, J. Yan, N. Xu, Y. Wang, and F.-H. Hsu, "Real-time High-quality Stereo Vision System in FPGA," in *Int'l Conf. on Field-Programmable Technology (FPT)*, 2013, pp. 358–361.

[10] F. Schumacher and T. Greiner, "Extension and FPGA Architecture of the Generalized Hough Transform for Real-Time Stereo Correspondence," in *Conf. on Design and Architectures for Signal and Image Processing (DASIP)*, 2013, pp. 223–229.

[11] D. Hafner, O. Demetz, and J. Weickert, "Why Is the Census Transform Good for Robust Optic Flow Computation?" in *Scale Space and Variational Methods in Computer Vision*, ser. Lecture Notes in Computer Science vol. 7893, A. Kuijper, K. Bredies, T. Pock, and H. Bischof, Eds. Springer Berlin Heidelberg, 2013, pp. 210–221.

[12] F. Schumacher and T. Greiner, "Two stage Real-Time Stereo Correspondence Algorithm and FPGA architecture using a modified Generalized Hough Transform," *Int'l. Conf. on Systems, Signals and Image Processing (IWSSIP)*, pp. 27–30, 2014.

[13] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," *Int'l Journal of Computer Vision*, vol. 47, pp. 7–42, 2002.

[14] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[15] D. G. Bailey, *Design for Embedded Image Processing on FPGAs*, 1st ed. John Wiley & Sons, 2011.

[16] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, Cambridge, New York, 2003.

Merging Multiple 3D Face Reconstructions

Leonard Thießen, Pascal Laube, Georg Umlauf, Matthias Franz

Institute for Optical Systems, University of Applied Sciences Constance, Germany

***Abstract*—In this paper we present a method to merge multiple 3D face reconstructions into one common reconstruction of higher quality. The individual three-dimensional face reconstructions are computed by a multi-camera stereo-matching system from different perspectives. Using 4-Points Congruent Sets and Iterative Closest Point the individual reconstructions are registered. Then, the registered reconstructions are merged based on point distance and reconstruction tenacity. To optimize the parameters in the merging step a kernel-based point cloud filter is used. Finally, this filter is applied to smooth the merged reconstruction. With this approach we are able to fill holes in the individual reconstruction and improve the overall visual quality.**

I. Introduction

Face recognition is an important problem in biometric applications that is usually based on two-dimensional images. However, it has been shown that the recognition rate can be improved, if the recognition is based on 3D face reconstructions, see Hensler et al. [1]. This requires an accurate and fast reconstruction method, e.g. based on real-time multi-camera stereo-matching as presented in [2]. The algorithm is based on four synchronized cameras (Figure 1) which captures images of a face from different perspectives (Figures 2(a)-2(d)). With this system it is possible to generate high-resolution depth images (Figure 2(e)) in real-time. This reconstruction also contains detailed information about the reconstruction tenacity (Figure 2(f)). A bad matching pixel correspondence in the computation of the depth map will result in a low tenacity value for the reconstructed 3D point. As a result, the algorithm yields one 3D reconstruction of the captured face.

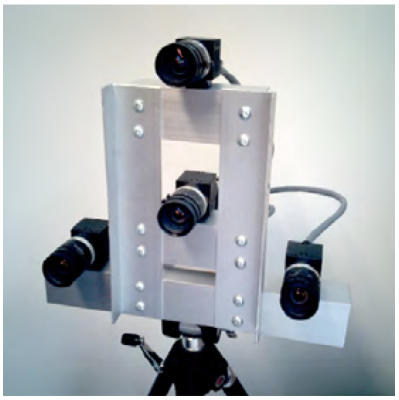


Fig. 1. Multi-camera stereo-matching system.

The main focus of [1] and [2] was on recognition rate and reconstruction speed, reconstruction quality was not the main concern. Thus, the reconstructions may be noisy or have holes. On a single GPU, the reconstruction system can compute up to four 3D reconstructions per second, the reconstruction quality can be improved by merging several 3D reconstructions shown in Figures 2(g)-(j). We propose a process to merge these reconstructions to improve the overall visual reconstruction quality. Due to the lack of a ground truth geometry the noise level is used as an additional quality measure during the merging process.

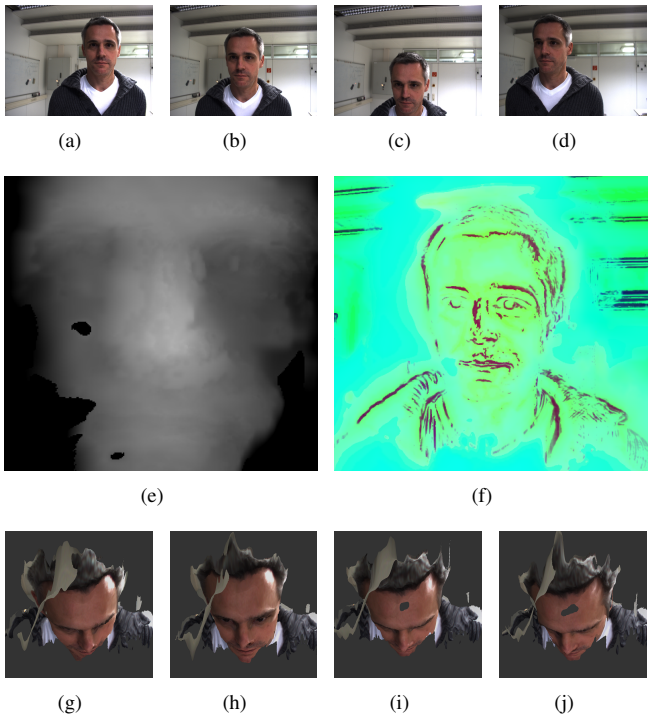


Fig. 2. Output of the multi-camera stereo-matching system. Figures (a)-(d) show the camera images yielding the depth map in Figure (e) with reconstruction tenacity (red: good tenacity, yellow: medium tenacity, cyan: bad tenacity) in Figure (f) and 3D reconstruction in Figure (j). Figures (g)-(j) show different 3D reconstructions from different perspectives.

II. Related Work

Registration and merging of 3D reconstructions are problems relevant to multiple research domains. For the scanning of large objects or terrains, Tang et al. [3] give a general overview of the techniques used to capture buildings. Their approach includes filtering and merging of a large number of scans. Bosse and Zlot [4] use a light detection and ranging sensor mounted on a vehicle to estimate vehicle motion between the

acquired merged scans. Local shape and constraints based on the vehicle motion are used to compute a 3D mapping. A bimodal 3D laser scanner is used in [5] to navigate an autonomous mobile robot. Range as well as reflectance data are combined to generate a navigable map. MacKinnon et al. [6] introduce quality metrics to detect regions which are likely to produce good quality when scanned. These regions are later on merged to generate a region map of optimal quality.

Scanning and reconstruction of smaller scale objects are done in [7]. The objects are scanned from different viewpoints and merged using the VRIP algorithm [8]. Lu et al. [9] use reconstructions from different angles of the human face for face recognition purposes but they do not describe how the different reconstructions are merged.

Thus, existing approaches either do not give a detailed outline of their merging techniques or do not use quality information in the process.

III. Merging 3D Face Reconstructions

The reconstruction system used captures a face from four different angles and uses multi-camera stereo-matching to compute a 3D reconstruction. The reconstructed geometry is represented as point cloud equipped with a measure to the quantify the local tenacity of the reconstruction. For details refer to [2]. Due to the speed of the reconstruction process several 3D reconstructions can be computed per second. These 3D reconstructions are usually from different perspectives since the person moves.

To merge these different 3D reconstructions we propose an approach consisting of four steps. First a coarse registration of the point clouds is done using 4-Points Congruent Set (4PCS) [10], see Section III-A. This is followed by a fine registration using Iterative Closest Point (ICP) [11], see Section III-B. Then the registered point clouds are merged to one 3D reconstruction using tenacity weighted interpolation, see Section III-C. In a last step the merged 3D reconstruction is filtered to erase noise.

A. Coarse Registration

The target of 3D registration is to align a data set $\mathcal{P}$ to a reference data set $\mathcal{Q}$. If $\mathcal{P}$ and $\mathcal{Q}$ are identical point clouds which only differ in position and orientation in space the result are two perfectly aligned point clouds with identical point coordinates. Thus, the actual result of a registration is the affine transformation for the alignment. The registration process is separated into a coarse registration step and the fine registration step. This is due to the fact that algorithms for fine registration are tuned to find local minima. Using these algorithms without initial coarse registration would lead to high computation times or most likely incorrect results.

For coarse registration we use 4PCS, see [10], [12]. This is a RANSAC-based algorithm [13] which performs well even for very noisy data. For RANSAC-based algorithms one has do define appropriate candidates for comparison. In 3D at least three points from each point cloud $\mathcal{P}$ and $\mathcal{Q}$ need to be compared. These randomly selected points define a corresponding pair of local coordinate frames. An affine transformation T can then be determined to align these frames. After applying T to $\mathcal{Q}$ the alignment is evaluated based on the number of points in $T(\mathcal{Q})$ that are within distance d to $\mathcal{P}$. This is the so-called Largest Common Point Set (LCP). This process is repeated until the desired size of the LCP or a certain iteration threshold is reached.

In 4PCS four coplanar points determine the pair of frames. A set of four coplanar points has the advantage that the ratios in the planar congruent set are invariant under affine transformations. A set $B = \{\mathbf{p}_1, \ldots, \mathbf{p}_4\}$ of four points is approximately coplanar, if the distance d_c between the lines $\mathbf{p}_1\mathbf{p}_2$ and $\mathbf{p}_3\mathbf{p}_4$ is small. Then, denote by $\mathbf{p}_E$ the midpoint of the shortest line perpendicular to $\mathbf{p}_1\mathbf{p}_2$ and $\mathbf{p}_3\mathbf{p}_4$, i.e. if $\mathbf{p}_1, \ldots, \mathbf{p}_4$ are coplanar, $\mathbf{p}_E$ is the intersection of these two lines. The ratios characterizing a frame are given by

$$r_1 = \frac{\|\mathbf{p}_1 - \mathbf{p}_E\|}{\|\mathbf{p}_1 - \mathbf{p}_2\|} \qquad \text{and} \qquad r_2 = \frac{\|\mathbf{p}_3 - \mathbf{p}_E\|}{\|\mathbf{p}_3 - \mathbf{p}_4\|}.$$

Finding B in $\mathcal{Q}$ is done by selecting three random points $\mathbf{p}_1$, $\mathbf{p}_2$, $\mathbf{p}_3$ and searching for $\mathbf{p}_4$ for which d_c is within a given tolerance. By selecting B with large diameter the algorithm becomes fast and globally robust.

To find a congruent frame in $\mathcal{P}$ the four parameters r_1, r_2 and the point distance $d_1 = \|\mathbf{p}_1 - \mathbf{p}_2\|$ and $d_2 = \|\mathbf{p}_3 - \mathbf{p}_4\|$ are used. After finding all point pairs in $\mathcal{P}$ with distances d_1 and d_2, the ratios r_1 and r_2 are computed, and the corresponding frames in $\mathcal{Q}$ and $\mathcal{P}$ are checked for congruence. The best matching pair of frames is finally selected based on LCP.

Selecting 200 pairs of frames has shown to be a good value for balancing run-time and registration error. Because points near the boundary of the scanned faces are particularly noisy we use only points in the center of the face, i.e. within a certain radius around the tip of the nose. Since the camera system is orthogonal to the captured face the tip of the nose can be found by evaluating z coordinates.

If the overlap of the two data sets is known, search can be limited for faster convergence. We assume an overlap between 50% and 60%.

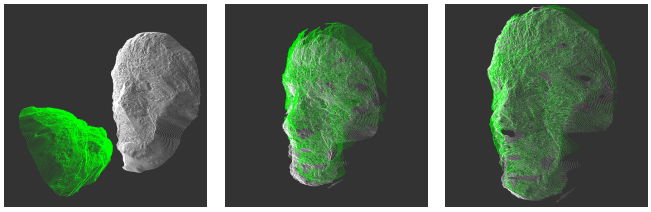
An example for the coarse registration using 4PCS is shown in Figures 3(a) and 3(b). We used the 4PCS implementation of [12].

B. Fine Registration

The fine registration using ICP is based on direct point neighborhoods, see [11], [14]. To register two point clouds $\mathcal{P}$ and $\mathcal{Q}$, for each point in $\mathcal{P}$ the nearest neighbor in $\mathcal{Q}$ is determined. The transformation T to align $\mathcal{P}$ and $\mathcal{Q}$ is computed by minimizing the squared distances between neighbor points. The algorithm is iterated until a specified error threshold is reached. Two error measures are used: The point-to-point distance, which is the Euclidean distance between two points, and the point-to-plane distance, which is the distance between a point and the tangent plane of its neighbor point. First we use the point-to-point distance up to a specified distance threshold. Then, point-to-plain distance is used until

the maximum number of iterations is reached. Repeating this ICP set-up two times yields sufficient registration results.

Figure 3(c) shows an example point cloud after registration with ICP. We used the ICP implementation of the trimesh2 library [15].



(a) (b) (c)

Fig. 3. Registration of two example point clouds (a), after applying 4PCS (b), and after applying ICP (c).

C. Tenacity Weighted Interpolation

Each 3D reconstruction has regions where the point positions are more or less reliable. This is due to lighting effects, relative camera positions and orientations, etc. This reliability of a reconstructed point $\mathbf{p}$ is measured by the tenacity $t(\mathbf{p}) \in [0,1]$, which is given as some weighted normalized cross-correlation of corresponding pixel neighborhoods in the four camera images, see [2]. Smaller values for t indicate higher point tenacity.

Having several 3D reconstructions of the same face, a rather frontal reconstruction is usually reliable for forehead, nose, and mouth regions and unreliable for the cheeks. Thus, we chose a rather frontal reconstruction as reference reconstruction $\mathcal{P}_0$ that is enhanced and enriched be the additional reconstructions $\mathcal{P}_1, \ldots, \mathcal{P}_k$. The merged reconstruction $\mathcal{R} = \{\mathbf{r}_1, \ldots, \mathbf{r}_M\}$ is generated by adding points from one $\mathcal{P}_i, i = 0, \ldots, k$, or from a tenacity weighted interpolation of points from several $\mathcal{P}_i$.

Initially, the merged reconstruction $\mathcal{R}$ is empty. Then, for every point in the reference reconstruction $\mathcal{P}_0 = \{\mathbf{p}_1, \ldots, \mathbf{p}_m\}$ a local merge step is computed. It is based on the point neighborhood N_i of $\mathbf{p}_i$ containing the n nearest neighbors of $\mathbf{p}_i$ from each of the additional reconstructions $\mathcal{P}_1, \ldots, \mathcal{P}_k$. Hence, N_i contains kn points. If $\mathbf{p}_i$ and all its neighbors have tenacity larger than $t_{\min}$, no point is added to $\mathcal{R}$ and further processing of $\mathbf{p}_i$ is skipped. This ensures a minimal overall point quality.

Denote by $\mathbf{p}_t$ a point from N_i with best tenacity and by $\mathbf{p}_d$ a point from N_i with smallest distance to $\mathbf{p}_i$. Furthermore, denote by d_t and d_d thresholds for the maximal distance of $\mathbf{p}_t$ and $\mathbf{p}_d$ to $\mathbf{p}_i$. N_i contains candidate points that represent the geometry better than P_i, if the set

$$\begin{aligned} P_R &= \{\mathbf{p}_t \mid (t(\mathbf{p}_t) < t(\mathbf{p}_i)) \wedge (\|\mathbf{p}_t - \mathbf{p}_i\| < d_t)\} \\ &\cup \{\mathbf{p}_d \mid (t(\mathbf{p}_d) < t(\mathbf{p}_i)) \wedge (\|\mathbf{p}_d - \mathbf{p}_i\| < d_d)\} \end{aligned}$$

is not empty. The point with best tenacity in P_R is added $\mathcal{R}$.

If P_R is empty we define a set of points for interpolation

$$\begin{aligned} P_I &= \{\mathbf{p}_t \mid (|t(\mathbf{p}_t) - t(\mathbf{p}_i)| < \delta_t) \wedge (\|\mathbf{p}_t - \mathbf{p}_i\| < d_t)\} \\ &\cup \{\mathbf{p}_d \mid (|t(\mathbf{p}_d) - t(\mathbf{p}_i)| < \delta_t) \wedge (\|\mathbf{p}_d - \mathbf{p}_i\| < d_d)\} \\ &\cup \{\mathbf{p}_i\}, \end{aligned}$$

where δ_t denotes a tenacity difference threshold. If P_I contains only one point, it is $\mathbf{p}_i$ which is added to $\mathcal{R}$. Otherwise the points in P_I are interpolated:

Linear two-point-interpolation For the set $P_I = \{\mathbf{q}_1, \mathbf{q}_2\}$ add to $\mathcal{R}$ the point $\mathbf{r}$ given by

$$\mathbf{r} = \lambda \mathbf{q}_1 + (1-\lambda)\mathbf{q}_2 \qquad \text{with} \qquad \lambda = \frac{t_{\mathbf{q}_1}}{t_{\mathbf{q}_1} - t_{\mathbf{q}_2}}.$$

Linear multi-point-interpolation For the set $P_I = \{\mathbf{q}_1, \ldots, \mathbf{q}_l\}, l \geq 3$, add to $\mathcal{R}$ the point $\mathbf{r}$ given by

$$\mathbf{r} = \frac{\sum_{i=1}^{l} (1 - t(\mathbf{q}_i))\mathbf{q}_i}{\sum_{j=1}^{l} (1 - t(\mathbf{q}_j))}$$

Overall there are the five parameters n, $t_{\min}$, d_t, d_d, and δ_t in the merge process that need to be optimized to achieve the best possible reconstruction. To compute an error measure for parameter optimization the ground truth geometry is required. However, this ground truth geometry is not available in our application setting. Therefore, a point cloud filter algorithm is used. If $\mathcal{R}_f = \{\mathbf{r}_1^f, \ldots, \mathbf{r}_M^f\}$ denotes the filtered reconstruction, the parameters are chosen such that $\mathcal{R}$ and $\mathcal{R}_f$ are close with respect to the error $e = \sum \|\mathbf{r}_i - \mathbf{r}_i^f\|^2$, i.e. the filtering has minimal effect on $\mathcal{R}$.

As point cloud filter we use the kernel-based filter method of Schall et al. [16]. For a point cloud $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_m\}$ a density function

$$\hat{f}(\mathbf{x}) = \frac{1}{mh^3} \sum_{i=1}^{m} \Phi\left(\frac{\mathbf{x} - \mathbf{p}_i}{h}\right)$$

is used to approximate the actual surface of a noisy point cloud, where Φ is the 3D Gaussian kernel of size h. This means, there is a likelihood function $L(\mathbf{x})$ that gives the probability that a point $\mathbf{x} \in \mathbb{R}^3$ is on the surface. L is an accumulation of local likelihood functions aligned to the local geometry at $\mathbf{p}_i$. This local geometry is represented by an anisotropic 3D Gaussian whose covariance is aligned to the local weighted principal component analysis at $\mathbf{p}_i$. The eigenvectors of the weighted covariance matrix

$$C_i = \sum_{j=1}^{m} (\mathbf{p}_j - \mathbf{c}_i)(\mathbf{p}_j - \mathbf{c}_i)^T \frac{\|\mathbf{p}_j - \mathbf{p}_i\|}{h}$$

approximate the tangent plane and surface normal at $\mathbf{p}_i$, where $\mathbf{c}_i$ is the weighted centroid of points $\mathbf{p}_j$ inside the kernel. The eigenvector corresponding to the smallest eigenvalue of C_i gives the (normalized) normal $\mathbf{n}_i$; the other two span the tangent plane. Thus, L is defined as

$$L(\mathbf{x}) = \sum_{i=1}^{m} \Phi_i(\mathbf{x} - \mathbf{c}_i) \left[h^2 - [(\mathbf{x} - \mathbf{c}_i)\mathbf{n}_i]^2\right].$$

Filtering the point cloud is now done by using the mean-shift method to move all points to positions of high probability. Using gradient-ascent maximization an iterative scheme

$$\mathbf{p}_i^0 = \mathbf{p}_i \qquad \text{and} \qquad \mathbf{p}_i^{k+1} = \mathbf{p}_i^k - \mathbf{m}_i^k$$

with

$$\mathbf{m}_i^k = \frac{\sum_{j=1}^{m} \Phi_j(\mathbf{p}_i^k - \mathbf{c}_j)[(\mathbf{p}_i^k - \mathbf{c}_j)\mathbf{n}_j]\mathbf{n}_j}{\sum_{j=1}^{m} \Phi_j(\mathbf{p}_i^k - \mathbf{c}_j)}$$

is applied. Iteration is stopped if

$$\|\mathbf{p}_i^{k+1} - \mathbf{p}_i^k\| < 10^{-4}h.$$

h is in the interval of one to ten times the average sampling density of the point cloud.

The final step in the merge process is a smoothing step on $\mathcal{R}$ using this kernel-based method.

IV. Results

To demonstrate the effectiveness of the proposed method we compare one reference reconstruction $\mathcal{P}_0$ of a male head to the merging results with three additional reconstructions $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3$. Examples are shown in Figures 2(g)-(j). The optimal parameters for the merging process were determined manually. A major influence on the overall visual quality of the reconstruction is the size n of neighborhoods N_i. If n is large the chance to find neighboring points with good tenacity increases. However, these points can be spatially far away, leading to visible holes in the reconstruction. In our tests, smaller neighborhoods led to better values for e. The reconstruction tenacity $t_{\min}$ has the biggest impact on the error e. Smaller values for $t_{\min}$ result in sparser point clouds with higher quality and small e. Distance thresholds d_t and d_d have similar effects on the reconstruction. Large distance thresholds result in holes in the point cloud and reduce the overall appearance. Small distance thresholds lead to merged reconstructions $\mathcal{R}$ mostly consisting of points from $\mathcal{P}_0$. Smaller distance thresholds as well as a tenacity difference threshold δ_t between 5% and 10% have a positive effect on appearance. Applying the kernel-based filter in a last step further smooths the surface and improves the visual quality.

Figure 4(a) shows a reconstruction with non-optimized parameters. It contains visible gaps and cracks that have not been part of the initial reconstructions. Parameters have then been stepwise adjusted to minimize e. For the given example in Figure 4(b), we use $t_{\min} = 0.5$. The smallest error e was achieved for parameters $n = 1, d_t = 0.000005, d_d = 0.000005$ and $\delta_t = 1.0$. Because point coordinates are based on pixel distance of the camera images, d_t and d_d are relative pixel distances. With these parameters almost all gaps and cracks have been removed from the merged reconstruction. By applying the filter to the point cloud, blurred and noisy regions 4(c) are smoothed and the contours become well defined. Evaluation of overall quality improvement is shown in Figures 5 and 6 with $t_{\min} = 0.4$. The reference reconstruction and the merged reconstruction are colored according to tenacity. One can see that the algorithm succeeds in filling the holes in the reference reconstruction and in expanding regions with high tenacity. With $t_{\min} = 0.3$ the merged reconstruction contains up to 25% more points than the reference reconstruction while maintaining or improving the average tenacity.

V. Conclusion

We present a method for merging face reconstructions to improve the overall reconstruction quality for use in face recognition. By endowing the merging process with thresholds and a tenacity-based interpolation as well as with an error measure for optimizing the merging parameters, we could increase the overall visual reconstruction quality.

The initial reconstructions that are merged later on contain color information which at the moment is lost in the interpolation step. For future work, point color has to be included in the process. Color information could be used when deciding point neighborhood as well as for lifelike visualization. The presented method and the resulting parameters have shown to be optimal for reconstructions generated by the stereo-matching approach in [2]. To prove the general applicability of our algorithm data sets of other 3D face reconstruction systems need to be evaluated. The overall algorithm has not been optimized for real-time application yet. Especially the filtering process is computationally expensive and could be improved by splitting the 3D-separable Gaussian function into three 1D functions as described in [17].

References

[1] J. Hensler, K. Denker, M. Franz, and G. Umlauf, "Hybrid face recognition based on real-time multi-camera stereo-matching," in *ISVC 2011*, G. B. et al., Ed. Springer, 2011, pp. 158–167.

[2] K. Denker and G. Umlauf, "An accurate real-time multi-camera matching on the gpu for 3d reconstruction," *Journal of WSCG*, vol. 19, pp. 9–16, 2011.

[3] P. Tang, D. Huber, B. Akinci, R. Lipman, and A. Lytle, "Automatic reconstruction of as-built building information models from laser-scanned point clouds: A review of related techniques," *Automation in construction*, vol. 19, no. 7, pp. 829–843, 2010.

[4] M. Bosse and R. Zlot, "Continuous 3d scan-matching with a spinning 2d laser," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2009, pp. 4312–4319.

[5] S. Frintrop, E. Rome, A. Nüchter, and H. Surmann, "A bimodal laser-based attention system," *Computer Vision and Image Understanding*, vol. 100, no. 1, pp. 124–151, 2005.

[6] D. MacKinnon, V. Aitken, and F. Blais, "Adaptive laser range scanning using quality metrics," in *Instrumentation and Measurement Technology Conference Proceedings, 2008. IMTC 2008. IEEE*, 2008, pp. 348–353.

[7] D. Huber and M. Hebert, "Fully automatic registration of multiple 3d data sets," *Image and Vision Computing*, vol. 21, no. 7, pp. 637–650, 2003.

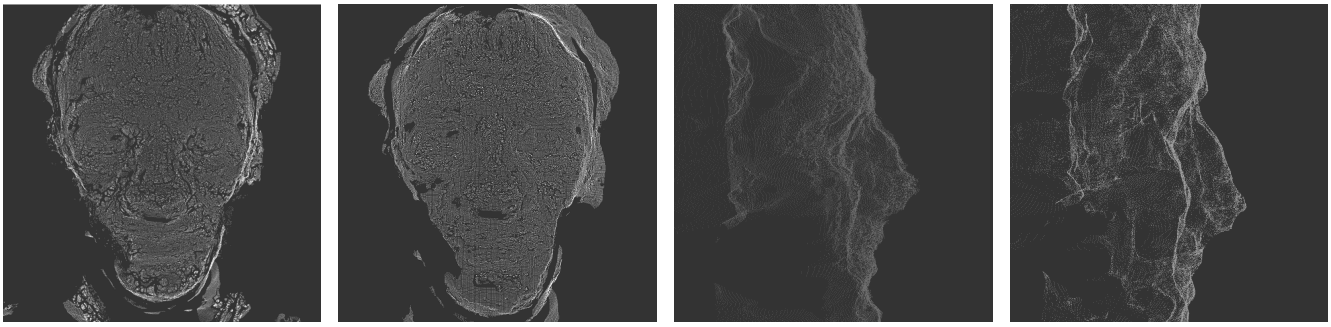
[8] F. P. Preparata and M. I. Shamos, *Computational Geometry: An Introduction*. Springer, 1985.

[9] X. Lu, A. K. Jain, and D. Colbry, "Matching 2.5d face scans to 3d models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 31–43, 2006.

[10] H. Alt, K. Mehlhorn, H. Wagener, and E. Welzl, "Congruence, similarity, and symmetries of geometric objects," *Discrete & Computational Geometry*, vol. 3, no. 1, pp. 237–256, 1988.

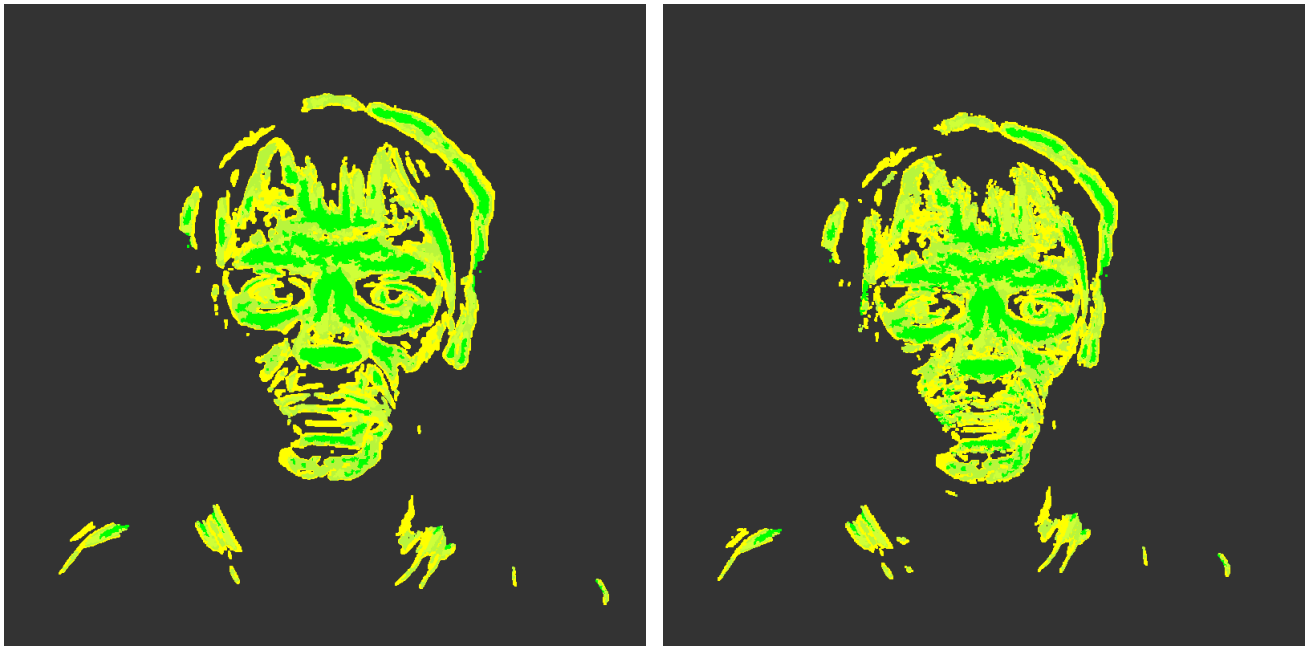
[11] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Robotics-DL tentative*. International Society for Optics and Photonics, 1992, pp. 586–606.

[12] D. Aiger, N. J. Mitra, and D. Cohen-Or, "4-points congruent sets for robust surface registration," *ACM Transactions on Graphics*, vol. 27, no. 3, pp. #85, 1–10, 2008.
[13] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Comm. of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[14] S. Rusinkiewicz and M. Levoy, "Efficient variants of the icp algorithm," in *Third International Conference on 3D Digital Imaging and Modeling*, 2001, pp. 145–152.
[15] S. Rusinkiewicz, "trimesh2," http://gfx.cs.princeton.edu/proj/trimesh2/.
[16] O. Schall, A. Belyaev, and H.-P. Seidel, "Robust filtering of noisy scattered point data," in *Proceedings of the Second Eurographics / IEEE VGTC Conference on Point-Based Graphics*, 2005, pp. 71–77.
[17] C. Lampert and O. Wirjadi, "An optimal nonorthogonal separation of the anisotropic gaussian convolution filter," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3501–3513, 2006.



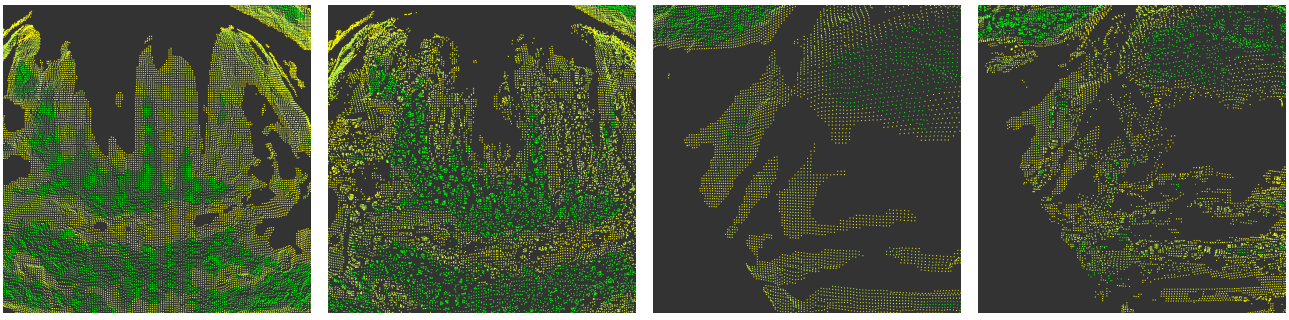
(a) (b) (c) (d)

Fig. 4. Merged reconstruction with non-optimized (a) and optimized (b) parameters. Closeup of the nose of a joined reconstruction with optimized (c) parameters and after filtering (d).



(a) (b)

Fig. 5. Tenacity colored images of a reference reconstruction (a) and the the respective merged reconstruction (b). The color gradient ranges from bright green (high tenacity) to yellow (bad tenacity).



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 6. Closeup on tenacity colored images of a reference reconstructions (a), (c), (e), (g) and the respective merged reconstructions below in (b), (d), (f), (h). The color gradient ranges from bright green (high tenacity) to yellow (bad tenacity).

Smoothie:a solution for device and content independent applications including 3D imaging as content

Razia Sultana
Offenburg University of Applied Sciences, Germany
Badstrasse 24, 77652 Offenburg
razia.sultana@hs-offenburg.de

Andreas Christ
Offenburg University of Applied Sciences, Germany
Badstrasse 24, 77652 Offenburg
Christ@hs-offenburg.de

Abstract— Network landscape of recent time contains many different network technologies, a wide range of end-devices with a large scale of capabilities and power, and an immense quantity of information represented in different data formats. Research on 3D imaging, virtual reality and holographic techniques will result in new user interfaces (UI) for mobile devices and will increase their diversity and variety. A lot of effort is being made in order to establish open, scalable and seamless integration of various technologies and content presentation for different devices including those that are mobile, considering the individual situation of the end user. Till today the research is going on in different parts of the world but the task is not completed yet. The goal of this research work is to find a way to solve the above stated problems by investigating system architectures to provide unconstrained, continuous and personalized access to the content and interactive applications everywhere and at anytime with different devices. As a Solution of the problem considered, a new architecture named "Smoothie" is proposed.

Keywords— Mobile learning; content formatting; device independency; virtual reality; 3D imaging; collaborative learning;

I. INTRODUCTION

Since the early attempts to support human communications by technological media such as the telegraph or telephone, a long time has passed. Many new technologies of communication have been developed and people became accustomed to them. Nowadays, we have so many ways to pass messages to each other that it becomes a complex task to maintain all these different systems. Additionally new methods to communicate not only with human beings but also with machines arise. This includes a range of applications from simple Web-based software up to the completely voice-controlled household. The speed of development brings benefits together with problems.

As first problem, the growth of communicative device use could be mentioned. The use of communicative device is growing every day. According to International Data Corporation (IDC) in 2017 the expected growth of smart connected device shipment is 58.1% compare to 2013 [1]. Cisco says today's 341 million Internet-of-Things (or machine-to-machine) connections are poised to grow to two billion in the coming years as wearable and smart infrastructure come online. By 2018, connected mobile devices will monthly flood the network with 15 times more data than all Internet traffic in 2000 [2]. Cisco IBSG predicts there will be 25 billion devices connected to the Internet by 2015 and 50 billion by 2020. That means 6.58 connected devices per person. Those devices will or may have different operating systems along with variety of capabilities.

As second problem, different formats of data could be mentioned. According to Internet Assigned Numbers Authority (IANA) on 11th June 2014 there are round about 1500 registered media types. Among those 1500 media types 69 are dedicated to text, 77 are for video, 21 for models (for example 3D models), 47 for images, and 142 for audio [3]. Approximately, 2500 file extensions are being supported by 1500 media types where 300 are 3D graphics formats, more than 200 CAD file formats, more than 500 audio file formats, around 600 image file formats, approximately 400 video file formats and around 300 text file formats [4]. In one hand so far there is no application or device found that is able to deal with all or at least most used data formats. On the other hand use and need of use of those varieties of data is raising everyday [2]. In this regard a ubiquitous data presentation method has to be mentioned named 3D. 3D representation is a long step ahead compare to 2D representation in performance, cost etc. Right now more than 300 different 3D data formats are available in the market [4]. It is impossible for any 3D enabled communicative device to deal with all of most used 3D data formats. Initially .obj later .step and .iges was designed and used for 3D asset exchange purposes. Now

.dae is a ubiquitous digital asset exchange format that makes interchange of file format for interactive 3D applications easier. Still the requirement of displaying 3D content independent of device remains unresolved.

The above mentioned problem becomes severe when 3D on mobile devices including wearable computers comes under consideration. Technological development in terms of hardware and software of mobile devices made it possible to have 3D as virtual reality, augmented reality or mixed reality being displayed on our mobile device at hand, which is anytime any where present and ready to use for business, learning, and entertainment or for a mixed purpose to meet group or individual interest.

Now the next questions to be answered are- do we need dedicated devices to be able to use application from different sectors; for example is it possible to use very same device for 3D games, reading text, view image, listen to audio, to see a 3D model of a product and see health report as 3D data for medical purpose? Are different devices along with 3D dedicated hardware able to communicate with each other in an understandable way; for example is it possible to send a 3D model from a handheld device to be printed by a 3D printer?

Right now within one word it is possible to answer all the above mentioned questions and the answer is "No". This barrier is created intentionally or no big step was taken so far globally to establish a device and content independent communication due to business purposes. We have many standards, protocols and specifications such as Zigbee, NFC, and Bluetooth etc to make communicative devices able to be connected and exchange data. But we do not have any standard or specification to make it sure that the sent data will be understood by the destination device independent of device type and capability.

Hence, the problems namely 1) heterogeneity of devices 2) heterogeneity of data formats including 3D data and 3) ever increasing use of both (1 & 2) in an unpredictable combination by the world population is introduced.

II. Drawback of Existing Solutions

There are several existing projects where target was to establish an underlying media independent communication such as Mobile People Architecture (MPA) [5], iceberg architecture [6], Integrated Personal Mobility Architecture (IPMoA) [7] Identification, Classification, Adaptation and Tagged XML (ICAT) [8]. Most of them were designed to serve a pre-decided commercial purpose for pre-decided types of devices. In addition, none of those projects except ICAT took independency of delivered data or content under consideration. As a major limitation of ICAT, lack of 3D content support has to be mentioned. There are existing projects where 3D content were enabled either as virtual reality or as augmented reality such as The invisible train, Augmented maps, Mobile augmented systems, Virtuoso etc. Most of those projects were designed for high-end devices but some of those considered mobile devices as well. There are supporting tools available as well to design theoretically a crossplatform application but practically all of those tools are dependent upon operating system and supported functionalities of the target device. There is no project found that considered not only independency of the device but also independency of the content regardless of the purpose of the communication.

III. Introduction to the Proposed Solution

The goal of this research work is to find a solution to enable the automation of the content adaptation process in heterogeneous devices. In order to realize such a system three major requirements have to be fulfilled:

- Identification of the connected device
- Generation, structuring and storage of generalized content
- Transformation process from general content to optimized and device dependent content

As a solution of above mentioned problem a new architecture named Smoothie is presented here. Figure 1 shows the most simplified version of the system overview of Smoothie.

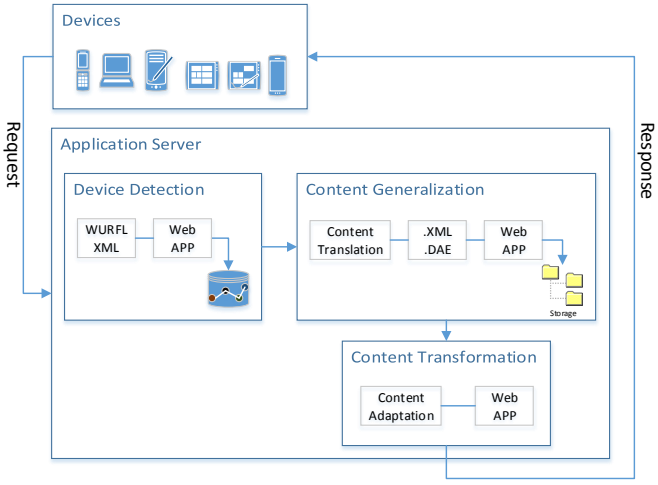


Figure 1. System overview of Smoothie

First, it is detected whether the user is connecting to the system via mobile device or by desktop device by analyzing the HTTP-request header coming from the end client´s device. Wireless Universal Resource File (WURFL) non commercial version is selected for the

description of the features of mobile devices and browsers because it is an XML configuration file which contains information about capabilities and features of many mobile devices in the wireless world. Also, the repository of device in WURFL is updated every day by contributors in the world. So it is an up to date specification that brings reliability in device data manipulation. The proposed architecture works with a combination of WURFL and a local database saved in the server.

Regarding content, providers do not have to care about the different mobile devices and to provide an optimized version of their file for every device. Content will be saved at first in its original format for safety reason then will be translated in a generalized format and stored in a XML database. In the prototype instead of XML database file system is used just for proof of concept. A pre-decided generalized format is XML or XML extended format such as DAE.

Based on the identified device capabilities generalized content format will be transformed into device dependent manner and displayed to the end client. So on the fly data conversion and serialization is happening in server side. For the prototype web browser is used along with the concept RWD (Responsive Web Design), as a container of the content to be presented at end client´s side. Any time it is possible to extend the serialization by implementing other available pipeline to process other output format such as PDF, WML etc. Web App is responsible for any kind of condition check and taking decision.

Even though the goal of this research work is to provide architecture of a system which is device and content independent, it should be kept in mind that there is no solution without limitation. Figure 2 below depicts the scope of the architecture named Smoothie. The creation of content and the processed content after being displayed at the end user device, how user is dealing with the content does not lie under the scope of this research work. Besides the mobile devices lie under the lifetime of telecommunication system from 3rd generation on, are considered under the scope of Smoothie due to necessary speed of data transfer capability.

IV. Implementation of Smoothie

From Figure 1 it is visible that major tasks of Smoothie can be divided in to three parts namely A) identification of device, B) Preparing generalized content and C) Transformation or adaptation of content including 3D data

A. *Identification of device*

The very first responsibility of the system i.e. Smoothie is to detect the connected device. Wireless Universal Resource File (WURFL) was selected for the description of the features of mobile devices and browsers, because the WURFL model is an XML configuration file that contains information about the capabilities and features of many mobile devices in the wireless world [9]. Also, the repository of devices listed in WURFL is updated every day by contributors in the world. Therefore, it is an up-to-date specification that brings reliability in device data manipulation. Under the scope of this research, the proposed system works with a combination of WURFL and a local database. Figure 3 below depicts the process

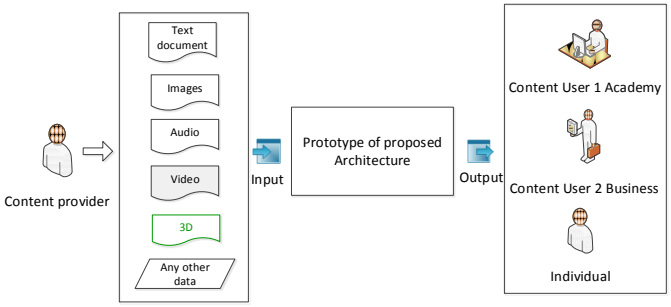


Figure 2 Pictorial representation of scope of the architecture

B. *Preparing generalized content*

As soon as the author uploads content, independent of its format or purpose, will be saved in the file system named Original Files Repository. The reason behind saving all content in its original form is safety, if anything goes wrong for example data gets corrupted or lost during any phase of the transformation process, the system can anytime start it over from the very beginning on with the originally uploaded data. As generalized form of content, Extensible Markup Language (XML) and for 3D content, an xml extended format namely Digital Asset Exchange (DAE) have been chosen. After preparing the generalized content the system will save them in Generalized Content Repository and the metadata of the content will be saved in SqlDB.

C. *Transformation of content*

Content adaptation is used as a synonym of transformation of content under the scope of this research work. In order to optimize the content presentation on different devices, the generalized content has to be adapted or translated in to a device dependent manner. To prove the concept of this research, Browser is used as a container for the content to be displayed at the end user´s device. Either web browser or mobile

browser is responsible to be the container based on the type of the client device.

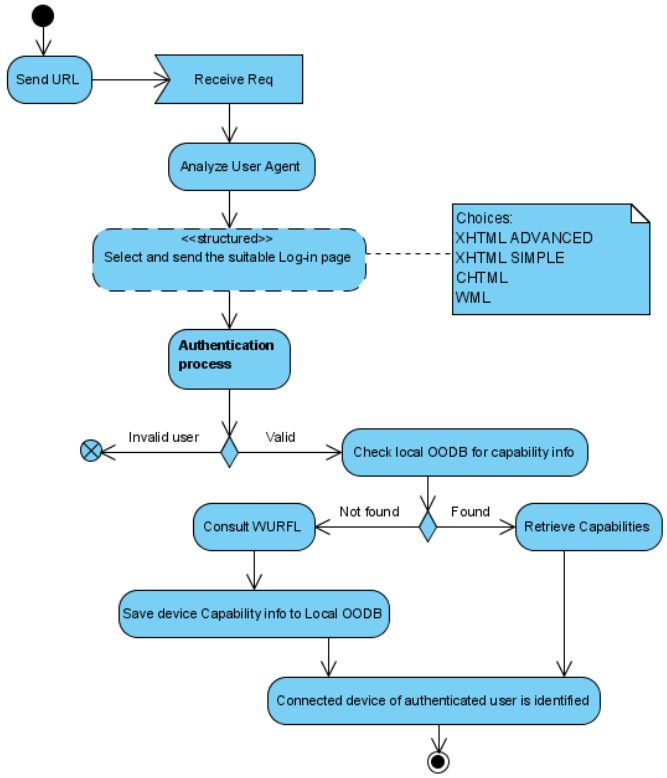


Figure 3 Identification of device

Content adaptation will take place if a user is requesting for an uploaded content. As soon as the request is received by the system, end user´s device identification and authentication processes will take place. If the mentioned processes are successful, metadata of the requested content will be retrieved from the database. Based on the ID which is a part of the link of the metadata different framework or library will be called for data transformation and visualization. For example for any kind of generalized document file apache cocoon will be appointed for serialization, for 3D generalized data ThreeJS will be appointed for rendering, any other data formats will be retrieved from the repository and at the end the content will be visualized on the browser of the end user´s device.

Initially according to W3C recommendation there were three defined ways for content adaptation [10] namely Server side, Proxy side and Client side.

Prototype of Smoothie implements server side, and client side content adaptation techniques; and easily extensible to use Hybrid content adaptation technique based on the capabilities of the end device along with it´s connectivity and the type and volume of the requested data. For example, a document file type transformation is taking place at server side; a 3D data adaptation is taking place at client side and for a file having bigger volume such as video file, Hybrid technique of content adaptation could be used.

V. SUMMARY OF TEST RESULT AND ANALYSIS

Smoothie was tested with several data types (or Internet media types) having registered in Apache.org mime type list [11]. All those are registered as described in RFC 4288. This is the document that defines procedures for the specification and registration of media types for use in MIME and other Internet protocols. The registry could be found in http://www.iana.org/assignments/media-types/. So far test has been conducted by using

- 4 different operating systems namely Windows 8 & 8.1, Android 4.1 & 4.4, iOS7.1 and LinuxMint 16 & Ubuntu 12.04
- 4 different browsers namely Mozilla Firefox 27 & 30; Google Chrome 33 &35; Safari 6.0; and Internet Explorer (IE) 10 & 11
- 4 different types of devices namely desktop (Linux), laptop (windows), tablet (android 4.4), and mobile phone (Android 4.1, iOS7.1)
- 3D data formats namely 3ds, obj, dae,stl,x3d,ply, lwo, ma, blend, fbx, wrl
- document and image formats namely doc, pdf, txt, jpeg, docx, odt, rtf
- audio and video data formats namely wma, mp4

Among above stated test parameters most of them were successful (in terms of data visibility) except iOS7.1, IE 10 and Safari in Windows machine, for 3D data due to lack of default WebGL support. Work around is possible to refine the functionality of Smoothie. For example use of Fallback in Smoothie. It facilitates an image-based fallback displaying a 360° view of the model where WebGL is not available. According to the idea provided by Sketchfab, a set of automatically generated screenshots of a 3D model can be taken, and put them together to create the equivalent of a turntable. It is generated as soon as the content author uploads the model. The fallback uses the set rotation axis of the model, so it might not always provide the ideal 360° view, depending on the initial view setup for the model. [12]. It cannot be guaranteed that the very same device which was able to display a certain kind of data once will always be able to show the data or vice versa.

VI. CONCLUSION

Nowadays enormous content is being created everyday and there are plenty of communicative devices available out there. As a part of the world population it is desired by everybody, independent of place, age, sex and social status, to have access to information. This innocent desire is being resisted by creating intentional barrier. The proposed architecture Smoothie is the first step towards the betterment of the world by opening a door to be able to communicate with everybody and anybody so that different schemes such as Education For All (EFA), Anytime Anywhere Communication, etc., can literally come true. Smoothie is an integrable architecture, could be integrated to any other application

- Where the application has to support varieties of devices along with varieties of data formats and those are not predictable
- Independent of purpose, such as business, education or individual interest.

It is equally applicable for private use, academic purposes as well as business purposes to meet requirement either for a group or for individual interest. The design goals of Smoothie were to establish a communication where

- Devices can access web content appropriate for their capabilities,
- Authors can create web content which is deliverable across different devices
- Content can be accessed from different kinds of device with different capabilities
- used tools for implementation are either open source or at least free to use so that the prototype could be open for everybody
- enhancement of the system to cope with future will not require huge effort or resource

To meet the design goal the complete process of Smoothie was divided into three major phases namely

Identification of device Preparing generalized content and Transformation of content

Identification of device is the very first responsibility of the system. In this phase end user device is being identified along with recognition of it´s capabilities with the help of WURFL. For later use, safety reason and to make the process faster collected device capabilities from WURFL are being saved in an Object Oriented Database. Among available open source object oriented databases DB4o was chosen for its replication through use of hibernate functionality, that synchronizes DB4o database with hibernate enhanced RDBMS (such as SQL) database. Another important feature of this phase is user authentication. The parameters of user authentication are dependent upon the application with which Smoothie will be integrated. Through user authentication, Smoothie is able to define the role and corresponding user rights over the data and the system of a valid user. User authentication and device identification are two different sub modules, correlated, not separable but easily extensible without requiring any change on the other sub module.

Preparing generalized content phase is independent from identification of device and Transformation of content phases. As soon as the content author uploads content this phase will start to be executed. Content author alias content provider may upload any kind of content/data in Smoothie independent of its format. Immediately after content upload Smoothie will save the contents in it´s original format in a file system for safety reason. As next step, based on the MIME Type of the uploaded content Smoothie will appoint either Apache Tika or Blender to prepare generalized content. Pre defined format of generalized content is XML for document type data and DAE for 3D data. Apache Tika and Blender are providing XML and DAE respectively as output. Generalized content is being saved in Generalized Content Repository and metadata of content is being saved in SqlDB, prepared to be transformed in device dependent manner based on necessity.

The third and last phase of Smoothie is transformation of content and it is dependent up on identification of device and preparing generalized content phases, even though it is a standalone module according to the design. Immediately after receiving a request the device has to be identified along with its capabilities then the requested content has to be retrieved from generalized content repository based on the metadata of the content saved in SqlDB. Afterwards, based on several parameters such as volume of requested data, connectivity, device capability etc. transformation of content phase will take place to provide the reply of a corresponding request in a device dependent manner. Depending upon the ID which is a part of the link of the metadata different framework or library will be called for data transformation and visualization. Prototype of Smoothie is using Server side and client side transformation of content but it is easily extensible to use Hybrid technique as well.

Above mentioned three phases of Smoothie are divided in three different modules, i.e. each of them is distinct but interrelated unit from which Smoothie is built up and it´s complex activity can be analysed and enhanced. In future if any further improvement in the quality, value, or extent of existing code is necessary in

any of those modules, it can be done without interrupting others. A pictorial representation of the divisions and relation of modules are shown below (in Figure 4).

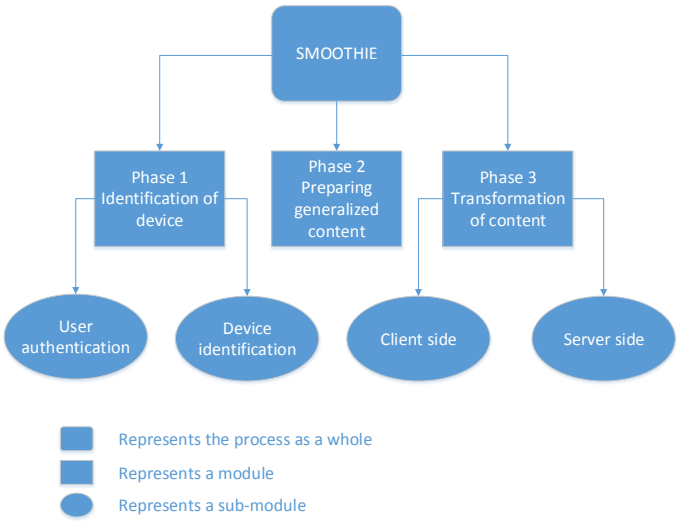


Figure 4 Pictorial representations of different modules of Smoothie

Smoothie was not designed to show what is technologically possible but what is technologically useful. It also does not want to describe what is coming next but what should come next. This research work provides a prototype to prove the concept but not a product. So naturally there are open issues and scope for further research and development. Some of those are listed below but not limited to those

- Test could be done by using other available browsers, operating systems, different data formats and different types of devices as many as possible.
- How to make loss less compression of generalized format of data, when it is necessary to be send through internet
- How to make the efficient use of the server capabilities so that the system is less dependent upon the capabilities of end client´s device. At this point facilities of cloud services could be introduced and implemented along with Smoothie.
- Import and Export functions of renowned open source software namely Blender is used to prepare generalized 3D data format. So, the proposed architecture is dependent upon the given functionalities of Blender. How to extend given functionalities of Blender so that any 3D data formats along with supporting files could be imported.
- Smoothie is concentrating now on displaying different formats of content but not on presentation of content with right style. It is dependent upon provided facilities of Apache Cocoon for serialization.
- Smoothie is dependent upon provided facilities of Apache Tika for document type content generalization.

Further research and developing time is necessary to overcome or improve the level of dependencies.

REFERENCES

[1] M. Framingham, "Tablet Shipments Forecast to Top Total PC Shipments in the Fourth Quarter of 2013 and Annually by 2015, According to IDC," IDC press release, 11 September 2013. [Online]. Available: http://www.idc.com/getdoc.jsp?containerId=prUS24314413. [Accessed 12 June 2014]

[2] Visual Networking Index (VNI), "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013–2018," CISCO, 2013. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html. [Accessed 20 June 2014]

[3] p. Ned Freed, s. Mark Baker and s. Bjoern Hoehrmann, "Media Types," IANA, 13 June 2014. [Online]. Available: http://www.iana.org/assignments/media-types/media-types.xhtml. [Accessed 20 June 2014]

[4] "The Source for File Extensions Information," File-Extensions.org, 5 February 2013. [Online]. Available: http://www.file-extensions.org/. [Accessed 12 June 2014]

[5] Mobi Thinking, "Global mobile statistics 2014 Part A: Mobile subscribers; handset market share; mobile operators," mobiThinking, May 2014. [Online]. Available: http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats/a. [Accessed 20 June 2014].

[6] H. Wang, B. Raman, R. Biswas, C. Chuah, R. Gummadi, B. Hohlt, X. Hong, E. Kiciman, Z. Mao, J. Shih, L. Subramanian, B. Zaoh, A. Joseph and R. Katz, "ICEBERG: An Internet-core Network Architecture for Integrated Communications," *IEEE Personal Communications: Special Issue on IP-based Mobile Telecommunications*, vol. 7, no. 4, pp. 10-19, 2000.

[7] B. Thai, R. Wan, A. Seneviratne and T. Rakotoarivelo, "Integrated Personal Mobility Architecture: A Complete Personal Mobility Solution," *Special Issue of MONET Journal on Personal Environment Mobility in MultiProvider and Multi-Segment Networks*, vol. 8, pp. 27-36, 2003.

[8] A. Christ and M. Feisst, "SW-Architecture for Device Independent Mobile Learning," in *Architectures for Distributed and Complex M-Learning Systems Applying Intelligent Technologies*, Hershey, Information Science Reference (an imprint of IGI Global), 2010, pp. 72-93.

[9] cientia Mobile Commercial Support & Licensing, "Welcome to WURFL," Scientia Mobile Inc., Last update 2014. [Online]. Available: http://wurfl.sourceforge.net/. [Accessed 24 January 2012].

[10] W3C Recommendation, "Mobile Web Best Practices 1.0 Basic Guidelines," July 2008. [Online]. Available: http://www.w3.org/TR/mobile-bp/. [Accessed 1 March 2014].

[11] Apache.org, "Mime Types," Apache.org, 12 March 2014. [Online]. Available: http://svn.apache.org/viewvc/httpd/httpd/branches/2.2.x/docs/conf/mime.types?view=annotate. [Accessed 20 June 2014].

[12] Sketchfab, "Sketchfab Blog," Sketchfab and users of Sketchfab, 11 August last update 2014. [Online]. Available: http://blog.sketchfab.com/post/47164072537/fallback-release. [Accessed 19 August 2014].

Applying a Traditional Calibration Method to a Focused Plenoptic Camera

Niclas Zeller, Franz Quint
Faculty of Electrical Engineering and Information Technology
Karlsruhe University of Applied Sciences
76133 Karlsruhe, Germany
Email: niclas.zeller@hs-karlsruhe.de,
franz.quint@hs-karlsruhe.de

Uwe Stilla
Department of Photogrammetry and Remote Sensing
Technische Universität München
80290 Munich, Germany
Email: stilla@tum.de

***Abstract*—This article presents a method to calibrate the optical imaging process of a focused plenoptic camera. At first, the concept of a focused plenoptic camera is presented, where the synthesis of images from the recorded light-field is described. It is shown that the synthesized image imitates the image of a real camera and thus traditional methods can be used for calibration. In this article a method approved for traditional cameras is applied to the recordings of a light-field camera. Based on experiments, the quality of the calibration method is evaluated. Amongst others it is shown that a pinhole camera with a conventional lens distortion model also holds true for a focused plenoptic camera.**

I. Introduction

A plenoptic camera or light-field camera is an optical camera which records, different from a traditional camera, not only the intensity of incident light on the image sensor but an image of the light-field of a scene.

During the last years plenoptic cameras gained more and more attention in the research fields of computer vision and photogrammetry. The main reason for revisiting this concept, which was already investigated more than hundred years ago [1], [2], is today's availability of fast graphic processor units (GPUs). Today's GPUs are capable to evaluate recorded light-field sequences with high frame rates ($\geq$ 30 fps). In [3] and [4] the first developed prototypes of plenoptic cameras are described.

Today, there exist different concepts of light-field cameras. While [4], [5] and [6] use a micro lens array (MLA) in front of the image sensor, the concept described in [7] relies on a 4×4 micro camera array. There are basically two different MLA-based concepts of a plenoptic camera. The "unfocsued" plenoptic camera as described in [4] and the focused plenoptic camera (or plenoptic camera 2.0) [8], [6]. Compared to the "unfocused" plenoptic camera the focused plenoptic camera has a higher spatial resolution. This results in a higher resolution of the synthesized image. In contrast, the "unfocused" plenoptic camera has a higher angular resolution.

For photogrammetric applications it is important to accurately know the relationship between a point in object space and the corresponding image point. To define this relationship, the intrinsic parameters of the camera have to be known precisely and have to be determined by a camera calibration process. In the last years some calibration methods for plenoptic cameras were proposed already. While [9] describes the calibration of a Lytro camera [4], [10] and [11] present calibration methods for a Raytrix camera [6]. Paper [11] mainly focuses on the calibration of the supplied depth information up to a range of about 10 m and investigates the depth resolution in this range, whereas [10] presents depth and image calibration of a Raytrix camera in a very short range. Therefore a more complex calibration setup is required.

In this article we apply a very simple calibration method [12], which is commonly used for traditional cameras, to the recordings of a Raytrix camera. Thus, we want to investigate the suitability of such traditional methods for plenoptic cameras.

This article is organized as follows. Section II briefly presents the concept of a focused plenoptic camera. In Section III we describe the calibration method [12] which was used in the experiment presented in Section IV. Section V illustrates the results of the experiments and Section VI draws conclusion.

II. Concept of a Focused Plenoptic Camera

A plenoptic camera records the light-field of a scene as a four dimensional (4D) function, by one single shot. Thus, a plenoptic camera gathers much more information about a scene than a traditional camera. In [13] Gortler et al. show that in free space it is sufficient to define the light-field as a 4D function since here the intensity along a ray does not change. Thus, the constant intensity along a ray can be defined by two position and two angle coordinates. Based on the 4D representation, the light-field of a convex object emitted in one direction can be described.

Since this article discusses calibration methods applied to a Raytrix camera, only the concept of this camera will be presented here. From a schematic point of view the only difference between a traditional camera and a focused plenoptic camera is the MLA in front of the image sensor. Thus, we will derive the concept of the focused plenoptic camera from a common thin lens projection.

Figure 1 shows the projection of an object which is in the distance a_L in front of the main lens to a focused image in the distance b_L behind the main lens. Here, the relationship

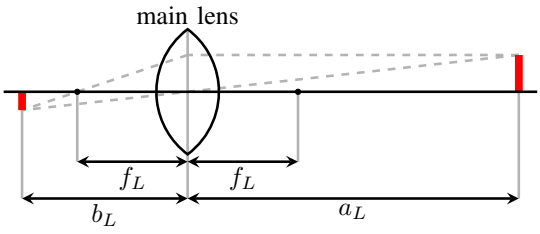


Fig. 1. Thin lens projection. A thin lens projects an object in distance a_L in front of the lens to a focused image in distance b_L behind the lens. The relationship between a_L and b_L dependents on the focal length f_L of the lens and is defined by the thin lens equation.

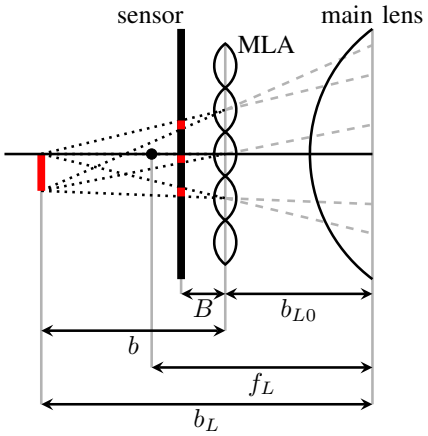


Fig. 2. Image projection inside a focused plenoptic camera. The virtual image, which would be formed in distance b_L behind the main lens, is focused on the image sensor by several micro lenses, which are placed in the distance b_{L0} behind the main lens. Out of the micro images of a point the distance b between MLA and the corresponding virtual main lens image can be estimated.

between the object distance a_L and the image distance b_L is defined by the thin lens equation given in eq. (1). Here f_L represents the focal length of the thin lens.

$$\frac{1}{f_L} = \frac{1}{a_L} + \frac{1}{b_L} \tag{1}$$

For a traditional camera the sensor would be placed in the image distance b_L behind the sensor and thus all object within the depth of field (DOF) around the object distance a_L occur focused in the recorded image.

In a Raytrix camera the sensor is placed closer than the image distance b_L to the sensor. Furthermore, a MLA is placed in distance B ahead of the sensor. Instead of placing the sensor in front of the image plane, a focused plenoptic camera can also be realized by placing the sensor behind the image plane as described in [5]. Figure 2 shows a schematic cross view of the interior of a Raytrix camera. The micro lenses of the MLA focus the virtual main lens image, which would occur behind the sensor, on the sensor. Thus, each micro lens forms a micro image on the sensor. One distinct feature of Raytrix cameras is that they have MLAs which consist of micro lenses with three different focal lengths. Each type of micro lenses focuses a different image distance b_L on the sensor. Thus, the DOF of the synthesized image is increased by a factor of three.

Within its DOF each micro lens can be considered as a pinhole. Thus, each pixel of a micro image represents one light ray (central ray of the corresponding micro lens). Since the optical center of each micro lens is known, for each ray additionally to the position coordinates (pixel position (x_I, y_I)) two angle coordinates (α_x and α_y) can be calculated.

A. Image Synthesis

One feature of a plenoptic camera is, that after capturing a light-field, images for different focus distances can be synthesized. This is done by calculating the image which would be formed on a sensor placed in a certain image distance b_L by the recorded light-field rays. To calculate the intensity of a pixel in the synthesized image, at first all rays of the 4D light-field which intersect the synthesized image plane at the corresponding position are searched. Out of the intensities of the selected rays a weighted average value is calculated. This average value represents the intensity of the synthesized pixel.

Because of the low angular resolution of a focused plenoptic camera there will occur artifacts in the synthesized image for regions which are not in focus on the selected image plane. Nevertheless, it is also possible to focus each pixel on a different image plane. Thus, if the image distance for each pixel is known, a totally focused image can be synthesized.

For an image point with a long image distance b_L (short object distance a_L) more rays are sampled than for an image point with a short image distance because it occurs focused in more micro images. Thus, in the synthesized totally focused image close objects have less spatial resolution than objects which are further away from the camera.

Here we will not go further into details of the image synthesis. For a detailed mathematical description we refer to [6].

B. Depth Estimation

It was already mentioned that if the image distance for each virtual image point is known, a totally focused image can be calculated. If a virtual image point is projected to at least two micro images, its distance to the MLA b can be estimated by triangulation of the corresponding rays. Thus, the depth map needed to synthesize a totally focused image also can be estimated from the recorded light-field function. Since the distance between MLA and sensor B usually is not known, a standardized value of the distance b, called virtual depth $v = \frac{b}{B}$, is estimated. For further descriptions on the depth estimation for plenoptic cameras we refer to the following papers: [6], [14], [15].

III. Image Calibration Method

Section II presented the concept of a focused plenoptic camera. Here it was shown that one can synthesize the image of a traditional camera from the recorded light field of a focused plenoptic camera.

In photogrammetry as well as in some fields of computer vision it is important to have a precisely defined relation between a pixel in the image and the corresponding point in the object space. Since the synthesized image imitates the recording of a traditional camera it is obvious to apply traditional camera calibration methods to the synthesized image.

An approved method for calibration of traditional cameras is the method described by Zhang [12]. In [12] the imaging process of the camera is simplified by using a pinhole camera model, like it is done in most of the common calibration methods. For our experiments we define the following intrinsic parameters for the pinhole camera model:

- f_x - focal length of the pinhole camera in x-direction (in pixels)
- f_y - focal length of the pinhole camera in y-direction (in pixels)
- (c_x, c_y) - image coordinates of the camera's principal point (in pixels)

By the definition of different focal lengths in x- and y-direction we consider the case that the pixels on the image sensor are not square but rectangular.

For the following definitions we use $\boldsymbol{x}_I$ as notation for an image point in Cartesian coordinates and $\tilde{\boldsymbol{x}}_I$ for the corresponding homogeneous coordinates. $\boldsymbol{x}_C$ defines a three dimensional (3D) point in camera coordinates and $\boldsymbol{x}_W$ in world coordinates.

$$\boldsymbol{x}_I = \begin{pmatrix} x_I & y_I \end{pmatrix}^T \tag{2}$$

$$\tilde{\boldsymbol{x}}_I = \begin{pmatrix} k \cdot x_I & k \cdot y_I & k \end{pmatrix}^T = k \cdot \begin{pmatrix} \boldsymbol{x}_I^T & 1 \end{pmatrix}^T \tag{3}$$

$$\boldsymbol{x}_C = \begin{pmatrix} x_C & y_C & z_C \end{pmatrix}^T \tag{4}$$

$$\boldsymbol{x}_W = \begin{pmatrix} x_W & y_W & z_W \end{pmatrix}^T \tag{5}$$

Based on the intrinsic parameters, the intrinsic matrix $\boldsymbol{M}$ can be defined, which describes the transformation from 3D camera coordinates $\boldsymbol{x}_C$ to image coordinates $\boldsymbol{x}_I$, as given in eq. (6).

$$\tilde{\boldsymbol{x}}_I = \boldsymbol{M} \cdot \boldsymbol{x}_C = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \cdot \boldsymbol{x}_C \tag{6}$$

Since real optical lenses are never perfect, they add distortion to the projected image. In the presented calibration method we use the distortion model presented by Brown [16], as it is implemented in the OpenCV calibration method [17]. Eq. (7) to (10) define the distortion model, where x_I and y_I are the undistorted and x'_I and y'_I ($\boldsymbol{x}'_I = \begin{pmatrix} x'_I & y'_I \end{pmatrix}^T$) the distorted image coordinates in pixels. This distortion model considers radial as well as tangential distortion.

$$x_I^* = \frac{x_I - c_x}{f_x} \quad \text{and} \quad y_I^* = \frac{y_I - c_y}{f_y} \tag{7}$$

$$r = \sqrt{(x_I^*)^2 + (y_I^*)^2} \tag{8}$$

$$x'_I = \left[x_I^* \cdot \left(1 + k_0 \cdot r^2 + k_1 \cdot r^4 + k_2 \cdot r^6\right) + 2 \cdot p_0 \cdot x_I^* \cdot y_I^* + p_1 \cdot \left(r^2 + 2 \cdot (x_I^*)^2\right)\right] \cdot f_x + c_x \tag{9}$$

$$y'_I = \left[y_I^* \cdot \left(1 + k_0 \cdot r^2 + k_1 \cdot r^4 + k_2 \cdot r^6\right) + p_0 \cdot \left(r^2 + 2 \cdot (y_I^*)^2\right) + 2 \cdot p_1 \cdot x_I^* \cdot y_I^*\right] \cdot f_y + c_y \tag{10}$$

Radial distortion is a radial symmetric distortion component with its origin in the principal point (c_x, c_y). Thus, it can be defined by a function of the distance r to the principal point. In Brown's distortion model radial distortion is defined by a polynomial of r, where the coefficients of odd exponents are zero. The nonzero coefficients are k_0, k_1 and k_2, as given in eq. (9) and (10). Tangential distortion is a radial asymmetric distortion which comes from decentralized lenses within the lens system. Brown defines the tangential distortion by the coefficients p_0 and p_1 as given in eq. (9) and (10).

The transform from world coordinates $\boldsymbol{x}_W$ to camera coordinates $\boldsymbol{x}_C$ can be defined by a 3D rigid transform which combines a 3D rotation and translation. The rigid transform is defined in eq. (11), where $\boldsymbol{R}$ is the rotation matrix and $\boldsymbol{t}$ the translation vector.

$$\begin{aligned} \boldsymbol{x}_C &= \begin{pmatrix} \boldsymbol{R} & \boldsymbol{t} \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{x}_W \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \end{pmatrix} \cdot \begin{pmatrix} x_W \\ y_W \\ z_W \\ 1 \end{pmatrix} \end{aligned} \tag{11}$$

The presented calibration method considers only planar objects. The world coordinate system for a recorded object is defined such that the x-y-plane is equivalent to the plane of the object. Thus, the z-component of the world coordinate system is always zero ($z_W = 0$) and the transfom becomes independent of the third colum of $\boldsymbol{R}$, as given in eq. (12).

$$\boldsymbol{x}_C = \begin{pmatrix} r_{11} & r_{12} & t_x \\ r_{21} & r_{22} & t_y \\ r_{31} & r_{32} & t_z \end{pmatrix} \cdot \begin{pmatrix} x_W \\ y_W \\ 1 \end{pmatrix} \tag{12}$$

If we now combine eq. (6) and (12), the transform from a point on the planar object to a point on the image plane is defined as given in eq. (13).

$$\begin{aligned} \begin{pmatrix} k \cdot x_I \\ k \cdot y_I \\ k \end{pmatrix} &= \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} r_{11} & r_{12} & t_x \\ r_{21} & r_{22} & t_y \\ r_{31} & r_{32} & t_z \end{pmatrix} \cdot \begin{pmatrix} x_W \\ y_W \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \cdot \begin{pmatrix} x_W \\ y_W \\ 1 \end{pmatrix} \end{aligned} \tag{13}$$

After dividing eq. (13) by h_{33} eq. (14) results.

$$\begin{aligned} \begin{pmatrix} k^* \cdot x_I \\ k^* \cdot y_I \\ k^* \end{pmatrix} &= \begin{pmatrix} h_{11}^* & h_{12}^* & h_{13}^* \\ h_{21}^* & h_{22}^* & h_{23}^* \\ h_{31}^* & h_{32}^* & 1 \end{pmatrix} \cdot \begin{pmatrix} x_W \\ y_W \\ 1 \end{pmatrix} \\ &= \boldsymbol{H} \cdot \begin{pmatrix} x_W \\ y_W \\ 1 \end{pmatrix} \end{aligned} \tag{14}$$

The planar homography between world and image coordinates, as given in eq. (14), is defined by eight linear independent coefficients h_{ij}^*. For each perspective from which the planar object is recorded the rotation matrix $\boldsymbol{R}$ and the translation vector $\boldsymbol{t}$ change. Thus each perspective results in a new transformation matrix $\boldsymbol{H}$. The rotation matrix $\boldsymbol{R}$ is defined by three independent angles (α, β and γ). Besides, the translation vector $\boldsymbol{t}$ relies on three independent coefficients (t_x, t_y and t_z). Since the estimated planar homography, defined by $\boldsymbol{H}$, gives us eight linear independent conditions, the six coefficients of the rigid transform between world and camera coordinates can be calculated. Thus, for each perspective two conditions are left to estimate the intrinsic parameters. The intrinsic matrix mostly is defined by four independent coefficients as given in eq. (6). Hence, the planar object has to be recorded from at

least two perspective to receive a unique solution. In reality recordings from much more than two perspectives are taken to average measurement errors.

After the estimation of the intrinsic and extrinsic parameters the distortion coefficients are estimated. This is done based on the calculated undistorted projection of an object point $\boldsymbol{x}_W$ to an image point $\boldsymbol{x}_I$, as defined in eq. (13), and the corresponding recorded and distorted image point $\boldsymbol{x}'_I$. Thus, the distortion coefficients (k_0, k_1, k_2, p_0 and p_1) can be estimated from eq. (7) to (10) by linear regression.

Based on the calculated undistorted points the intrinsic and extrinsic coefficients are updated and the distortion coefficients are calculated again. This procedure is repeated until consistency is reached.

For a more detailed description on how the intrinsic, extrinsic and distortion coefficients are estimated we refer to [12] and [18].

IV. Experiments

This section presents experiments which were performed to evaluate the calibration method described in Section III when applying it to a focused plenoptic camera. In the presented experiments the OpenCV [17] implementation of the calibration method was used. Here calibration points are recorded by using a planar chessboard pattern. Each corner point between four adjacent chessboard fields is detected in the recorded image. Since for all those points the corresponding world coordinates on the pattern are known, the points can be used for calibration. For the experiment a pattern with 10×7 fields was used.

In the presented experiments we used a Raytrix R5 camera with the following two different lenses mounted to it:

1) 4 mm–12 mm zoom lens, set to 12 mm focal length
2) 35 mm fixed focal length

The main goal of the performed experiments was to distinguish if the synthesized imaging of a focused plenoptic camera can also be defined by a traditional camera model. Besides, we wanted to evaluate how the two different lenses, with different focal lengths and levels of distortion, are effecting the calibration results.

For the experiments, for both lenses three measurement series were recorded. In each series the chessboard pattern was recorded from 50 as different as possible perspectives. To each series the calibration method presented in Section III was applied multiple times. For each calibration the camera model was slightly changed as will be described in Section IV-A to IV-D. The calibration method was performed to each of the three measurement series to evaluate the consistency of the results.

A. Complete Calibration Model

In the first experiment for both lenses the calibration was performed using the complete calibration model with four intrinsic and five distortion coefficients.

B. Constant Aspect Ration ($f_x = f_y$)

For the second experiment the aspect ration was set to one ($f_x/f_y = 1$). Thus, during the calibration one intrinsic parameter less has to be estimated. If this assumption conforms the real camera model a more consistent result can be expected.

C. Constant Aspect Ration and Fixed Principal Point

One further assumption is made in the third experiment. Beside the constant aspect ration the principal point of the camera model is considered to be in the image center and is set constantly to that point. This reduces again the number of coefficients to be estimated. Furthermore, the principal point certainly lies somewhere around the image center.

D. Constant Aspect Ration, Fixed Principal Point and Only One Distortion Coefficient

The fourth experiment is only performed to the 35 mm lens since for the 12 mm lens no improvement is expected. Here the assumptions from the second and third experiment are still considered to hold true. Besides, the distortion model is defined by only the first radial distortion coefficient k_0. All other distortion coefficients are set to zero ($k_1 = k_2 = p_0 = p_1 = 0$). This experiment is performed since the distortion for the 35 mm lens seems to be very weak. Thus, by reducing again the degrees of freedom the remaining coefficients should be estimated more consistent if the assumption holds true.

V. Results

In this section the results for the performed experiments are presented. For evaluation at first some statistics are defined.

The root mean square (RMS) of the reprojection error σ_{Rep} gives a measure how good the estimated model represents the measured values. The reprojection error $e_{Rep}^{(i,j)}$ is the distance between an object point $\boldsymbol{x}_W^{(i)}$ projected on the image plane, based on the projection model, and the corresponding recorded image point $\boldsymbol{x}_I'^{(i,j)}$. Eq. (15) gives the definition of σ_{Rep}.

$$\sigma_{Rep} = \sqrt{\frac{1}{N \cdot M} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} \left(e_{Rep}^{(i,j)}\right)^2} \tag{15}$$

Here, $e_{Rep}^{(i,j)}$ represents the reprojection error of the i-th chessboard corner in the j-th image. N is the number of chessboard corners and M the number of recorded images.

Besides, we define the root mean square error (RMSE) of the estimated focal length σ_f and the principal point σ_c, as given in eq. (16) and (17).

$$\sigma_f = \sqrt{\frac{1}{2 \cdot S} \sum_{i=0}^{S-1} \left(f_x^{(i)} - \overline{f}_x\right)^2 + \left(f_y^{(i)} - \overline{f}_y\right)^2} \tag{16}$$

$$\sigma_c = \sqrt{\frac{1}{S} \sum_{i=0}^{S-1} \left(\boldsymbol{c}^{(i)} - \overline{\boldsymbol{c}}\right) \cdot \left(\boldsymbol{c}^{(i)} - \overline{\boldsymbol{c}}\right)^T} \tag{17}$$

In eq. (16) and (17) S represents the number of measurement series which were recorded ($S = 3$). Of course, from a series of three no real statistics can be calculated. Nevertheless, based

TABLE I. ESTIMATED INTRINSIC PARAMETERS FOR THE EXPERIMENT A USING A 12 mm FOCAL LENGTH

series no.	**1**	**2**	**3**
f_x [Pixel]	1058.8	1053.8	1048.1
f_y [Pixel]	1064.3	1060.9	1053.9
c_x [Pixel]	494.0	488.4	475.8
c_y [Pixel]	496.9	528.4	502.4

TABLE II. ESTIMATED INTRINSIC PARAMETERS FOR THE EXPERIMENT B USING A 12 mm FOCAL LENGTH

series no.	**1**	**2**	**3**
f_x [Pixel]	1065.9	1068.7	1059.7
f_y [Pixel]	1065.9	1068.7	1059.7
c_x [Pixel]	495.3	492.6	471.0
c_y [Pixel]	494.2	525.6	500.7

TABLE III. ESTIMATED INTRINSIC PARAMETERS FOR THE EXPERIMENT C USING A 12 mm FOCAL LENGTH

series no.	**1**	**2**	**3**
f_x [Pixel]	1064.1	1068.6	1058.8
f_y [Pixel]	1064.1	1068.6	1058.8
c_x [Pixel]	511.5	511.5	511.5
c_y [Pixel]	511.5	511.5	511.5

TABLE IV. CALCULATED STATISTICS FOR THE 12 mm FOCAL LENGTH

experiment no.	**1**	**2**	**3**
σ_{Rep} [Pixel]	0.804	0.807	0.808
σ_f [Pixel]	4.333	3.758	4.032
σ_c [Pixel]	15.727	17.368	–

on σ_f and σ_c the quality of different calibration results can be compared.

In the following two subsections the calibration results for the 12 mm and 35 mm lens will be presented separately.

A. *Calibration Results for* $f_L = 12\,\text{mm}$

Table I to III present the estimated intrinsic parameters which resulted from the three experiments performed for the 12 mm lens. Besides, Table IV shows the corresponding calculated statistics. Since for the third experiment the principal point was set to the image center prior to the calibration, σ_c is not meaningful.

If we compare σ_{Rep}, one can see that the models of all three experiments conform quite well to the measured data. For the model of the first experiment which uses all degrees of freedom, the RMS of the reprojection errors of course is minimum. This case uses the most complex model and thus, the measured data can be adapted best.

Nevertheless, fixing the aspect ratio to $f_x/f_y = 1$ or setting the principal point to a certain image coordinate improves the estimation of the other parameters, at least as long as the assumption conforms more or less to the real model.

For the third experiment σ_f is worse than for the second experiment. This indicates, that the selected principal point differs from the real one. By adjusting the principal point it should be possible to achieve a more consistent estimation of the focal length.

TABLE V. ESTIMATED INTRINSIC PARAMETERS FOR THE EXPERIMENT A USING A 35 mm FOCAL LENGTH

series no.	**1**	**2**	**3**
f_x [Pixel]	3259.1	3261.6	3284.0
f_y [Pixel]	3262.3	3261.0	3291.0
c_x [Pixel]	520.7	430.0	426.4
c_y [Pixel]	346.3	389.5	323.7

TABLE VI. ESTIMATED INTRINSIC PARAMETERS FOR THE EXPERIMENT B USING A 35 mm FOCAL LENGTH

series no.	**1**	**2**	**3**
f_x [Pixel]	3262.1	3261.1	3285.0
f_y [Pixel]	3262.1	3261.1	3285.0
c_x [Pixel]	521.1	430.0	423.5
c_y [Pixel]	347.0	389.4	327.0

TABLE VII. ESTIMATED INTRINSIC PARAMETERS FOR THE EXPERIMENT C USING A 35 mm FOCAL LENGTH

series no.	**1**	**2**	**3**
f_x [Pixel]	3257.5	3247.8	3250.5
f_y [Pixel]	3257.5	3247.8	3250.5
c_x [Pixel]	511.5	511.5	511.5
c_y [Pixel]	511.5	511.5	511.5

TABLE VIII. ESTIMATED INTRINSIC PARAMETERS FOR THE EXPERIMENT D USING A 35 mm FOCAL LENGTH

series no.	**1**	**2**	**3**
f_x [Pixel]	3255.7	3252.7	3275.4
f_y [Pixel]	3255.7	3252.7	3275.4
c_x [Pixel]	511.5	511.5	511.5
c_y [Pixel]	511.5	511.5	511.5

TABLE IX. CALCULATED STATISTICS FOR THE 35 mm FOCAL LENGTH

experiment no.	**1**	**2**	**3**	**4**
σ_{Rep} [Pixel]	0.336	0.336	0.352	0.358
σ_f [Pixel]	12.583	11.006	9.709	10.053
σ_c [Pixel]	51.461	51.609	–	–

B. *Calibration Results for* $f_L = 35\,\text{mm}$

Table V to VIII shows the intrinsic parameters which were estimated for the 35 mm lens and Table IX gives the corresponding calculated statistics. For the 35 mm lens the same behavior as for the 12 mm lens can be observed. Here, the most consistent estimation of the focal length was achieved when fixing the aspect ratio as well as the principal point. One reason therefore could be that for this lens the principal point conforms quite well to the image center. Another reason is that for long focal lengths a shift of the principal point has not as much effect as for short focal lengths. This can also be seen when comparing σ_c for the 35 mm and the 12 mm lens. For the 35 mm lens σ_c is much higher than for the 12 mm lens. Besides, one can see that the results for experiment C and D are almost the same. This means, for the 35 mm lens the distortion model can be reduced to only one parameter with only a small rise of σ_{Rep} and σ_f.

VI. CONCLUSION

In conclusion it can be said that traditional camera calibration methods, like the method of Zhang [12] can be used

to calibrate the synthesized images of a plenoptic camera. This can be seen by the small reprojection errors for all experiments. Nevertheless, for a focal length of the main lens which is long with respect to the image size (small field of view (FOV)) Zhang's calibration method seems to be inappropriate. Especially for long focal lengths, the intrinsic parameters are strongly correlated to the extrinsic orientation. Thus, errors in the extrinsic orientation will also affect the intrinsic parameters. One way to improve the calibration would be to used a 3D calibration object instead of a planer object. A 3D object brings more perspective distortion to the recorded image. Thus, extrinsic and intrinsic parameters can be separated more accurate. Another problem is that Zhang's method does not really minimize the squared error between recorded and calculated image points like it is done for a bundle adjustment. Instead is uses several optimization steps. Here it could be investigated, how the errors of estimation are propagated form one step to the next, to see how the error of any parameter effects the reprojection error.

REFERENCES

[1] F. E. Ives, "Parallax stereogram and process of making same," USA Patent US725 567, 04 14, 1903.

[2] G. Lippmann, "Epreuves reversibles. photographies integrales," *Comptes Rendus De l'Academie Des Sciences De Paris*, vol. 146, pp. 446–451, 1908.

[3] E. H. Adelson and J. Y. A. Wang, "Single lens stereo with a plenoptic camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 99–106, February 1992.

[4] R. Ng, "Digital light field photography," Ph.D. dissertation, Stanford University, Stanford, USA, July 2006.

[5] A. Lumsdaine and T. Georgiev, "The focused plenoptic camera," in *IEEE International Conference on Computational Photography (ICCP)*, San Francisco, CA, April 2009, pp. 1–8.

[6] C. Perwaß and L. Wietzke, "Single lens 3D-camera with extended depth-of-field," in *Human Vision and Electronic Imaging XVII*, Burlingame, California, USA, January 2012.

[7] K. Venkataraman, D. Lelescu, J. Duparre, A. McMahon, G. Molina, P. Chatterjee, R. Mullis, and S. Nayar, "Picam: An ultra-thin high performance monolithic camera array," *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH Asia 2013*, vol. 32, no. 6, pp. 1–13, 11 2013.

[8] A. Lunsdaine and T. Georgiev, "Full resolution lightfield rendering," Adobe Systems, Inc., Tech. Rep., 2008.

[9] D. Dansereau, O. Pizarro, and S. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1027–1034.

[10] O. Johannsen, C. Heinze, B. Goldlücke, and C. Perwaß, "On the calibration of focused plenoptic cameras," in *GCPR Workshop on Imaging New Modalities*, 2013.

[11] N. Zeller, F. Quint, and U. Stilla, "Kalibrierung und Genauigkeitsuntersuchung einer fokussierten plenoptischen Kamera," in *34. Wissenschaftlich-Technische Jahrestagung der DGPF (DGPF Tagungsband 23 / 2014)*, vol. 23, 3 2014.

[12] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 1, 1999, pp. 666–673.

[13] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proceedings of the 23rd annual conference on computer graphics and interactive techniques, SIGGRAPH*. New York, NY, USA: ACM, 1996, pp. 43–54.

[14] N. Zeller, F. Quint, and U. Stilla, "Calibration and accuracy analysis of a focused plenoptic camera," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. II-3, pp. 205–212, 09 2014.

[15] S. Wanner and B. Goldlücke, "Globally consistent depth labeling of 4d lightfields," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[16] D. C. Brown, "Close-range camera calibration," *Photogrammetric Engineering*, vol. 37, no. 8, pp. 855–866, 1971.

[17] G. Bradski, "The opencv library," *Dr. Dobb's Journal of Software Tools*, 2000.

[18] G. Bradski and A. Kaehler, *Learning OpenCV - Computer Vision in C++ with the OpenCV Library*. O'Reilly Media, 09 2008.

Sequential Decoding of Binary Block Codes Based on Supercode Trellises

Jens Spinner, Jürgen Freudenberger*, Sergo Shavgulidze†

*Institute for System Dynamics
HTWG Konstanz, University of Applied Sciences, Germany
Email: jens.spinner@htwg-konstanz.de, jfreuden@htwg-konstanz.de
Web: www.isd.htwg-konstanz.de
†Faculty of Power Engineering and Telecommunications
Georgian Technical University, Georgia
Email: sshavgulidze@gncc.ge

***Abstract*—This work presents a novel sequential algorithm for soft decision decoding of binary linear block codes. Ordinary sequential decoding of block codes is based on the trellis representation of the code. In contrast to Viterbi decoding, the sequential algorithm does not visit all nodes in the trellis and hence reduces the time complexity. On the other hand, the space complexity with sequential decoding may even be larger than with Viterbi decoding, because the algorithm requires the complete trellis as well as a stack to store interim results. This work proposes an algorithm that reduces the space complexity of sequential decoding of binary block codes. The representation of the code is based on two trellises of supercodes, where both trellises have a much smaller space complexity than the original trellis of the code.**

***Index Terms*—maximum–likelihood decoding; soft-decision decoding; sequential decoding; binary block codes**

I. Introduction

Soft-decision decoding of binary block codes has a long history, e.g. reliability-based decoding dates back to the early seventies [1], [2] and was improved in many publications [3], [4], [5], [6]. Such algorithms can offer a performance that is similar to maximum-likelihood (ML) decoding, but usually do not guarantee to find the ML codeword.

Most common, maximum-likelihood decoding of block codes is implemented by representing the code as a graph, the so-called trellis [7], and applying Viterbi's well-known algorithm [8]. Alternative implementations to ordinary Viterbi decoding were for example presented by Fuijwara *et al.* [9] for block codes, or by Fossorier and Lin [10] for convolutional codes.

Due to the tree structure of convolutional codes, sequential decoding algorithms are an efficient alternative to Viterbi decoding [11]. With sequential decoding the time complexity of the algorithm depends on the channel error probability. For moderate channel error probabilities, sequential decoding can significantly reduce the time complexity of the decoding procedure.

In general, binary block codes have no tree structure. Hence sequential decoding requires an efficient representation of the code. In [**?**], different sequential decoding strategies for binary block codes were proposed, where all algorithms perform a search in the trellis representation of the code.

In this paper we consider a different approach to represent the code. The proposed sequential algorithm is based on two trellises of supercodes. A supercode is a superset that contains all codewords of the code. For codes of high rates, the trellis representation of a supercode has fewer nodes than the trellis of the original code. Using two supercodes, we can represent the actual code as the intersection of these supercodes. The concept of supercode decoding was introduced by Barg *et al.* [12]. Trellis based supercode decoding was proposed in [13], [14], where Viterbi decoding is applied in both supercode trellises. More recently, in [15] the concept of supercode decoding was combined with the priority-first search algorithm [16]. The concept of supercode decoding was also applied to reduce the complexity of algebraic decoding of Bose-Chaudhuri-Hocquenghem (BCH) codes [17]. The corresponding algorithm is limited to hard-input and bounded minimum distance decoding.

In this work, we demonstrate that the concept of supercode decoding can be applied to sequential decoding of binary block codes as introduced in [18]. The representation of the code by means of supercode trellises reduces the space complexity of sequential decoding procedure. The new approach enables a trade-off between time and space complexity. The paper is organized as follows. In Section II, we briefly discuss the syndrome trellis of a binary linear block code and the sequential decoding algorithm in order to introduce our notation and the basic concept. In the following section, we introduce the new decoding algorithm. The simulation results given in Section IV show that the proposed decoding algorithm increases the time complexity only slightly.

II. Sequential Decoding of Binary Block Codes

In this section we describe the sequential decoding algorithm for binary block codes as presented in [**?**]. This decoding procedure is based on the code trellis [7]. A trellis $\mathcal{T} = (\mathcal{S}, \mathcal{W})$ is a labeled, directed graph, where $\mathcal{W} = \{w\}$ denotes the set of all branches in the graph and $\mathcal{S} = \{\sigma\}$ is

the set of all nodes. The set $\mathcal{S}$ is decomposed into $n+1$ disjoint subsets $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_n$ that are called levels of the trellis. Similarly, there exists a partition of the set $\mathcal{W} = \mathcal{W}_1 \cup \mathcal{W}_2 \cup \ldots \cup \mathcal{W}_n$. A node $\sigma \in \mathcal{S}_t$ of the level t may be connected with a node $\widetilde{\sigma} \in \mathcal{S}_{t+1}$ of the level $t+1$ by one or several branches. Each branch w_t is directed from a node σ of level $t-1$ to a node $\widetilde{\sigma}$ of the next level t. We assume that the end levels have only one node, namely $\mathcal{S}_0 = \{\sigma_0\}$ and $\mathcal{S}_n = \{\sigma_n\}$. A trellis is a compact method of presenting all codewords of a code. Each branch of the trellis w_t is labeled by a code symbol $\mathrm{v}_t(w_t)$. Each distinct codeword corresponds to a distinct path in the trellis, i.e., there is a one-to-one correspondence between each codeword $\mathbf{v}$ in the code and a path $\mathbf{w}$ in the trellis: $\mathbf{v}(\mathbf{w}) = \mathrm{v}_1(w_1), \ldots, \mathrm{v}_n(w_n)$. We denote code sequence segments and path segments by $\mathbf{v}_{[i,j]} = \mathrm{v}_i, \ldots, \mathrm{v}_j$ and $\mathbf{w}_{[i,j]} = w_i, \ldots, w_j$, respectively. The *syndrome trellis*, can be obtained using its parity-check matrix [7]. The syndrome trellis is minimal inasmuch as this trellis has the minimal possible number of nodes $\mid \mathcal{S} \mid$ among all possible trellis representations of the same code. With the syndrome trellis we also introduce a node labeling. The nodes of the trellis will be identified by $(n-k)$-tuples with elements from $\mathbb{F}_2$, with $\mathbf{0}$ referring to the all zero $(n-k)$-tuple. At level $t=0$ and level $t=n$ the trellis contains only one node, the all zero node $\sigma_0 = \sigma_n = \mathbf{0}$.

The sequential decoding procedure as presented in [18] is a stack algorithm, i.e. a stack is required to store interim results. The stack contains code sequences of different lengths. Let $\mathbf{v}_t$ denote a code sequence of length t, i.e.$\mathbf{v}_t = \mathrm{v}_1, \ldots, \mathrm{v}_t$. Each code sequence is associated with a metric and a node σ_t. The node s_t is the node in the trellis that is reached if we follow the path corresponding to the code sequence through the trellis. The metric rates each code sequence and the stack is ordered according to these metric values where the code sequence at the top of the stack is the one with the largest metric value. There exits different metrics in the literature to compare code sequences of different length. In the following, we consider the Fano metric which is defined as follows. Let v_i be the i-th code bit and r_i the i-th received symbol for transmission over a discrete memoryless channel. The Fano metric for a code bit v_i is defined by

$$M(r_i|v_i) = \log_2 \frac{p(r_i|v_i)}{p(r_i)} - R \tag{1}$$

where $p(r_i|v_i)$ is the channel transition probability, $p(r_i)$ is the probability to observe r_i at the channel output, and R is the code rate. The Fano metric of a code sequence $\mathbf{v}_t$ is

$$M(\mathbf{r}_t|\mathbf{v}_t) = \sum_{i=1}^{t} M(r_i|v_i) \tag{2}$$

where $\mathbf{r}_t$ is the sequence of the first t received symbols.

Algorithm 1. The sequential decoding starts in the first node σ_0 of the trellis. Calculate the metric values for $v_1 = 0$ and $v_1 = 1$. Insert both paths into the stack according to their metric values. In each iteration, remove the code sequence at the top from the stack. Verify whether the branches for $v_{t+1} = 0$ and $v_{t+1} = 1$ exist for the node σ_t corresponding to the top path. If a branch exists then calculate the metric and insert the code sequence into the stack. The algorithm terminates when a path approaches the end node σ_n. The estimated codeword is the top path in the final iteration.

We demonstrate the decoding algorithm in the following example.

Example 1. Consider for instance the code $\mathcal{C} = \{(0000), (1110), (1011), (0101)\}$ with parity-check matrix

$$\mathbf{H} = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix} .$$

The corresponding trellis is depicted in Fig 1a). We assume transmission over a binary symmetrical channel with error probability 0.1. Hence, we have

$$M(r_i|v_i) \approx \begin{cases} 0.3 & \text{for } r_i = v_i \\ -2.8 & \text{for } r_i \neq v_i \end{cases}$$

The following tables represent the stack for the received sequence $\mathbf{r} = (0010)$.

1st iteration

$\mathbf{v}_t$	$M(\mathbf{r}_t\|\mathbf{v}_t)$
0	0.3
1	-2.8

2nd iteration

$\mathbf{v}_t$	$M(\mathbf{r}_t\|\mathbf{v}_t)$
00	0.6
01	-2.5
1	-2.8

3rd iteration

$\mathbf{v}_t$	$M(\mathbf{r}_t\|\mathbf{v}_t)$
000	-2.2
01	-2.5
1	-2.8

4th iteration

$\mathbf{v}_t$	$M(\mathbf{r}_t\|\mathbf{v}_t)$
0000	-1.9
01	-2.5
1	-2.8

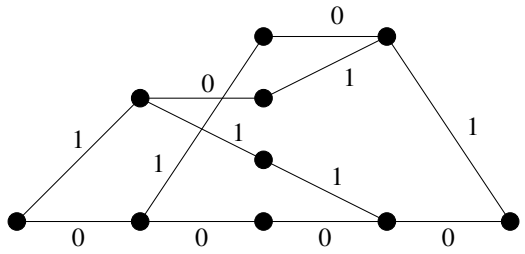
III. The Proposed Supercode Decoding Algorithm

A *supercode* $\mathcal{C}_i$ of the block code $\mathcal{C}$ is a code containing all codewords of $\mathcal{C}$. For a linear code $\mathcal{C}$ with parity-check matrix $\mathbf{H}$, we can construct two supercodes $\mathcal{C}_1$ and $\mathcal{C}_2$ such that $\mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_2$. Let $\mathbf{H} = \begin{pmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \end{pmatrix}$ be the parity-check matrix of the code $\mathcal{C}$, this means that $\mathbf{H}_1$ and $\mathbf{H}_2$ are two sub-matrices of $\mathbf{H}$. Then the sub-matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ define the supercodes $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively.

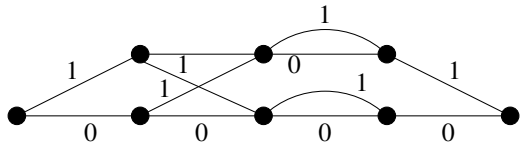
Example 2. Consider for example the code $\mathcal{C} =$ from Example 1. We obtain

$$\begin{aligned} \mathbf{H}_1 &= \begin{pmatrix} 1 & 1 & 0 & 1 \end{pmatrix} \\ &\Downarrow \\ \mathcal{C}_1 &= \{\underline{(0000)}, (1100), \underline{(1110)}, (0010), \\ &\quad \underline{(1011)}, (1001), \overline{(1011)}, \underline{(0101)}\} \end{aligned}$$

a) Trellis of code $\mathcal{C}$



b) Trellis of supercode $\mathcal{C}_1$



c) Trellis of supercode $\mathcal{C}_2$

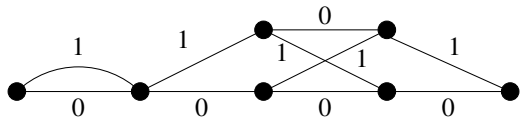


Fig. 1. Trellises of the example code and of its two supercodes.

and

$$\begin{aligned} \mathbf{H}_2 &= \begin{pmatrix} 0 & 1 & 1 & 1 \end{pmatrix} \\ &\Downarrow \\ \mathcal{C}_2 &= \{\underline{(0000)}, (1000), (0110), \underline{(1110)}, \\ &\quad \underline{(1011)}, (1101), (0011), \underline{(0101)}\} \end{aligned},$$

where the underlined vectors are the codewords of the code $\mathcal{C}$. The corresponding supercode trellises are depicted in Fig. 1b) and 1c).

Next we state the proposed sequential decoding algorithm. Any path stored in the stack is associated with a metric value as well as two states $\sigma_{t,1}$ and $\sigma_{t,2}$ which are the states in the trellis for supercode $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively.

Algorithm 2. The sequential decoding starts in the nodes $\sigma_{0,1}$ and $\sigma_{0,2}$ of the supercode trellises. Calculate the metric values for $v_1 = 0$ and $v_1 = 1$. Insert both paths into the stack according to their metric values. In each iteration, remove the code sequence at the top from the stack. Verify whether the branches for $v_{t+1} = 0$ and $v_{t+1} = 1$ exist for both nodes $\sigma_{t,1}$ and $\sigma_{t,2}$ corresponding to the top path. If a branch exists in both supercode trellises then calculate the metric for this path and insert the code sequence into the stack. The algorithm terminates when a path approaches the end nodes $\sigma_{n,1}$ and $\sigma_{n,2}$. The estimated codeword is the top path in the final iteration.

We demonstrate the decoding algorithm in the following example, where we consider the same setup as in Example 1.

Example 3. The following tables represent the stack for the received sequence $\mathbf{r} = (0010)$ for the proposed algorithm.

1st iteration

$\mathbf{v}_t$	$M(\mathbf{r}_t\|\mathbf{v}_t)$
0	0.3
1	-2.8

2nd iteration

$\mathbf{v}_t$	$M(\mathbf{r}_t\|\mathbf{v}_t)$
00	0.6
01	-2.5
1	-2.8

3rd iteration

$\mathbf{v}_t$	$M(\mathbf{r}_t\|\mathbf{v}_t)$
001	0.9
000	-2.2
01	-2.5
1	-2.8

4th iteration

$\mathbf{v}_t$	$M(\mathbf{r}_t\|\mathbf{v}_t)$
000	-2.2
01	-2.5
1	-2.8

5th iteration

$\mathbf{v}_t$	$M(\mathbf{r}_t\|\mathbf{v}_t)$
0000	-1.9
01	-2.5
1	-2.8

Note that the stack in the third iteration differs from Example 1, because the code sequence 001 exists in both supercode trellises but not in the actual code. This code sequence is deleted in the next iteration, because it cannot be extended in both supercode trellises.

As the previous example demonstrates, the time complexity of the proposed algorithm may be larger than with Algorithm 1. This results from code sequences that exist in the super codes, but are not valid in the actual code. Nevertheless, both algorithms result in the same codeword.

Theorem. *Algorithm 1 and Algorithm 2 result in the same estimated codeword.*

Proof: Both algorithms differ only with respect to the representation of the code. To prove the proposition it is sufficient to verify that both representations are equivalent. We first prove by induction that the estimated codeword corresponds to a valid path in both supercode trellises, i.e. it is a codeword in both supercodes. The base case is the initial step where the code bits 0 and 1 are inserted in the stack. Note that a linear code has no code bit positions with constant values. Hence, the transitions $\mathrm{v}_1 = 0$ and $\mathrm{v}_1 = 1$ exist in both supercode trellises. For the inductive step, we assume that a path for the code sequence $\mathbf{v}_t$ exists in both supercode trellises. It follows from Algorithm 2 that this path is only extended if the extended path exists in both supercode trellises. This proves the claim that the estimated codeword corresponds to a valid path in both supercode trellises. Now note that $\mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_2$, i.e. a path is only valid in both supercode trellises if and only if it is a valid codeword of the code $\mathcal{C}$. ■

IV. Simulation results

In this section we present some simulation results for a Bose-Chaudhuri-Hocquenghem (BCH) code [11], [19]. This code has length $n = 31$, dimension $k = 21$, and minimum Hamming distance $d = 5$. The trellis of this code has 14334 nodes. The two supercode trellises have only 750 and 611 nodes, respectively. Fig. 2 presents simulation results for

transmission over a binary symmetrical channel. With sequential decoding the number of visited nodes in the trellis (the number of iterations) depends on the number of transmission errors. Fig. 2 depicts the average number of visited nodes per codeword versus the bit error rate of the channel. Note that the time complexity with Algorithm 2 is at most 1.75 times larger than with Algorithm 1.

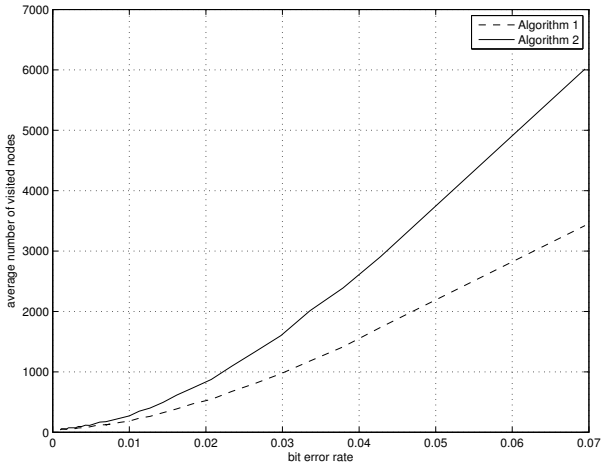


Fig. 2. Average number of visited nodes versus the bit error rate.

The final version of the paper will also present simulation results for transmission over an additive white Gaussian noise (AWGN) channel.

V. Conclusions

In this work we introduced the concept of sequential decoding of binary block codes based on supercode trellises. Note that a trellis of a binary block code of length n and dimension k contains up to $n2^{\min(k,n-k)}$ nodes [7], i.e. codes of rate $R \geq \frac{1}{2}$ have a space complexity of order $O(2^{n-k})$. The proposed algorithm is based on two supercodes which have only half the number of parity bits and hence a space complexity of order $O(2^{\frac{n-k}{2}})$. On the other hand, the algorithm increases the time complexity, because invalid code sequences are inserted into the stack. The presented simulation results show only a small increment in the number of decoding iterations.

References

[1] D. Chase, "Class of algorithms for decoding block codes with channel measurement information," *IEEE Transactions on Information Theory*, pp. 170–182, 1972.

[2] B. Dorsch, "A decoding algorithm for binary block codes and J-ary output channels," *Information Theory, IEEE Transactions on*, vol. 20, no. 3, pp. 391–394, May 1974.

[3] M. Fossorier and S. Lin, "Soft-decision decoding of linear block codes based on ordered statistics," *IEEE Trans. Inform. Theory*, vol. IT-41, pp. 1379–1396, Sep. 1995.

[4] C. Argon, S. McLaughlin, and T. Souvignier, "Iterative application of the Chase algorithm on Reed-Solomon product codes," *Proceedings IEEE ICC 2001*, pp. 320–324, 2001.

[5] M. Tomlinson, C. Tjhai, and M. Ambroze, "Extending the Dorsch decoder towards achieving maximum-likelihood decoding for linear codes," *IET Communications*, vol. 1, no. 3, pp. 479–488, June 2007.

[6] A. Gortan, R. Jasinski, W. Godoy, and V. Pedroni, "Achieving near-MLD performance with soft information-set decoders implemented in FPGAs," in *2010 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS),*, Dec 2010, pp. 312–315.

[7] J. K. Wolf, "Efficient maximum likelihood decoding of linear block codes using a trellis," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 76–80, Jan. 1978.

[8] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 260–269, Apr. 1967.

[9] T. Fuijwara, H. Yamamoto, T. Kasami, and S. Lin, "A trellis-based recursive maximum-likelihood decoding algorithm for binary linear block codes," *IEEE Trans. Inform. Theory*, vol. IT-44, pp. 714–729, March 1998.

[10] M. Fossorier and S. Lin, "Differential trellis decoding of convolutional codes," *IEEE Trans. Inform. Theory*, vol. IT-46, pp. 1046–1053, May 2000.

[11] S. Lin and D. J. Costello, *Error Control Coding*. Upper Saddle River, NJ: Prentice-Hall, 2004.

[12] A. Barg, E. Krouk, and H. C. A. Van Tilborg, "On the complexity of minimum distance decoding of long linear codes," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1392–1405, Jul 1999.

[13] J. Freudenberger and M. Bossert, "Maximum-likelihood decoding based on supercodes," in *Proc. 4th. International ITG Conference Source and Channel Coding*, Erlangen, Germany, Jan. 2004, pp. 185–190.

[14] J. Freudenberger and V. Zyablov, "On the complexity of suboptimal decoding for list and decision feedback schemes," *Discrete Applied Mathematics*, vol. 154, no. 2, pp. 294 – 304, 2006.

[15] Y. S. Han, H. T. Pai, P. N. Chen, and T. Y. Wu, "Maximum-likelihood soft-decision decoding for binary linear block codes based on their supercodes," in *submitted to the 2014 IEEE International Symposium on Information Theory*, 2014.

[16] Y. Han, C. Hartmann, and C.-C. Chen, "Efficient priority-first search maximum-likelihood soft-decision decoding of linear block codes," *Information Theory, IEEE Transactions on*, vol. 39, no. 5, pp. 1514–1523, Sep 1993.

[17] J. Freudenberger and J. Spinner, "A configurable Bose-Chaudhuri-Hocquenghem codec architecture for flash controller applications," *Journal of Circuits, Systems, and Computers*, vol. 23, no. 2, pp. 1–15, Feb 2014.

[18] L. Aguado and P. Farrell, "On hybrid stack decoding algorithms for block codes," *Information Theory, IEEE Transactions on*, vol. 44, no. 1, pp. 398–409, Jan 1998.

[19] A. Neubauer, J. Freudenberger, and V. Kühn, *Coding Theory: Algorithms, Architectures and Applications*. John Wiley & Sons, 2007.

Thermally Modulated MOG array for early detection of emissions of overheated cables

Jens Knoblauch, Navas Illyaskutty, Liwa Wu,
Christian Langen and Heinz Kohler
Institute for Sensorics and Information Systems
Karlsruhe University of Applied Sciences
Karlsruhe, Germany

Rolf Seifert and H. B. Keller
Institute of Applied Computer Sciences
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany

Abstract—**In the present work, we propose a novel, very sensitive multi metal oxide sensor array for early detecting and alarming of the emissions from overloaded isolation materials used in electrical cabinets. The principal sensor element containing four different sensing layers on one chip is operated thermo-cyclically, yielding conductance signatures which can be used to specifically identify gases and even distinguish from other interfering atmospheres. For validation of the sensor response on pyrolysis gases, a setup was built which allows (i) overloading of isolated electrical cables or (ii) heating of isolation polymers at defined temperatures. Gas sensor signals were analyzed using ProSens, a numerical tool designed for evaluation of conductance profiles. The obtained conductance profile shapes bear good capability for detection and identification of pyrolysis gas emissions at relatively low temperatures even before a color-change of the PVC-coating is visible.**

Keywords— metal oxide gas sensor; sensor array; pyrolysis; early fire detection; data analysis

I. INTRODUCTION

Developing sensor systems for the detection of fires in electrical cabinets at early stages of development is of great interest in the current scenario. Metal oxide gas sensors (MOG) of the TGS-type have been reported to be appropriate for detection of conventional fires and smoke [1]. This type of gas sensor could be utilized for early detection of fires in electrical cabinets and cable channels as well, as pyrolysis of cable materials leads to emission of distinct gas mixtures depending on insulation material composition. But isothermal operation of MOG can result in false alarms due to cross-sensitivities at similar conditions such as gases of conventional fires, solvent vapor, cigarettes smoke etc. However, surface gas processes involved in and responsible for huge changes of gas sensitive layer conductance are all specifically dependent on temperature and on the catalytic, the adsorption/desorption behavior and the diffusion properties of the in most cases porous gas sensing material. In the early work of Korotchenkov et al. [2] this aspect was investigated exposing pyrolysis gases of different materials on tin oxide thin film gas sensors. From investigations of the sensitivities dependent on thin-film temperature they found gas specific dependencies and proposed to operate the sensor element at two different fixed temperatures for better recognition of the gases and they concluded that additional thermal annealing of the active zone at 300-350°C is necessary for better recovering of the sensing layer after contact with pyrolysis gases. Later, in our earlier work, it was shown that operating MOG sensors thermo-cyclically and sampling conductance simultaneously can yield gas specific, well reproducible conductance over time profiles (CTPs) [3]. This enables identification of the reactive gas components involved and meets the addressed problem of better recovering of the gas sensing layer when operated in the range 100°C≤T≤420°C. Besides the fact, that it is still not clear, which kind of SnO2/additive – combinations are most efficient for reliable early fire detection [4], moreover, thermo-cycling of an array of sensors (Fig. 1) combining different material specific sensitivities can allow the analysis of complex gas mixtures by analyzing several gas and material specific CTPs simultaneously. Such a proposed four-fold sensor array (Fig. 1), some preliminary sensitivity data at exposure to pyrolysis gases and their numerical analysis using ProSens [5] are reported in this paper. Combined with a microcontroller based system, incorporating the basic functions needed for thermo-cyclic operation and data acquisition, as well as signal analysis the system will open various new applications as for example in the field of early fire detection.

Fig.1. Gas sensing element (4x4mm^2) comprising a four-fold sensor array which is thermo-cyclically operated for gas analysis.

II. EXPERIMENTAL

A. Preparation of sensors

The sensor chips (alumina substrate, Fig. 1) featuring four-fold inter-digitated-electrode (IDE) structures and a resistive temperature sensor (metal line in the middle of the chip) of Au or Pt on one side and a thin film Pt-heater on the reverse side were prepared initially by using DC sputtering, photolithography and plasma etching techniques. Then, micro-structured sensitive layers of SnO_2 and its various additive combinations [6] such as SnO_2+Pd, SnO_2+ZnO, or SnO_2+Pt were deposited on the Au or Pt electrodes in thick film form,

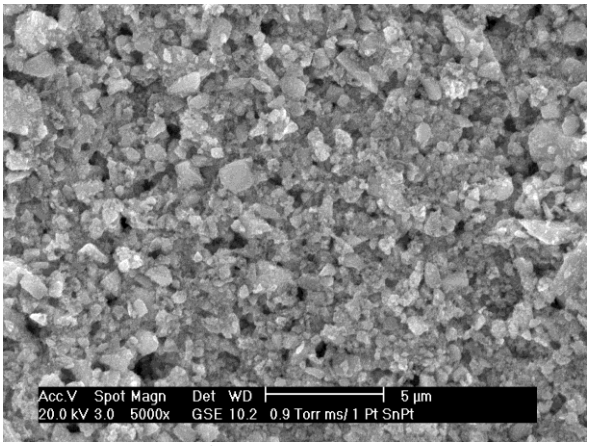


Fig. 2. ESEM image of Pt-added SnO_2 on Pt-IDE

employing micro-dispensing of SnO_2/additive-powder filled organic pastes. For this purpose pure SnO_2 powder was prepared following a sol-gel route. The gel was ground in liquid form for 6 h at 200 rpm by ball milling in zirconia vessels, then dried and pre-sintered at 450°C for 5 h to get fine pale yellow SnO_2 powder. 1 wt. % of PdO, ZnO and PtO (related to SnO_2) were admixed with the SnO_2 powder by adding $Pd(NO_3)_2{*}2H_2O$, $Zn(NO_3)_2{*}6H_2O$ and $Pt(NO_3)_2$ dissolved in ethanol, respectively. The mixture was milled for 12 hours to get a homogenous distribution of the fine powders with the dissolved salt, dried, mixed with organic carriers and then the paste was dispensed. After sintering at a maximum temperature of 700°C and prior to gas sensing studies, the as prepared layers (Fig. 1) were subjected to surface morphology studies and were analyzed using an environmental scanning electron microscope (Fig. 2), ESEM XL 30 FEG (Philips). Properly dispensed and sintered sensor chips were mounted on a TO8 header using a micro welding technique.

B. Experimental setup and mode of operation

An automated standard gas sensor test setup was used for characterization of the gas sensor arrays with respect to their specific conductance behaviour towards propene (C_3H_6) and carbon monoxide (CO) at different concentrations. Flow-through technique was used to test the gas sensing properties; using synthetic air as gas carrier with the desired concentrations of analytes admixed (total gas flux: 100 ml/min) [7].

For validation of the sensor response on pyrolysis gases, a setup was built which allows (i) overloading of isolated electrical cables or (ii) heating of isolation polymers at defined temperatures. Details of construction are described elsewhere [8]. A quartz glass tube is used as reactor and was equipped with either (i) copper electrodes or (ii) a thermocouple which is well placed next to the sample material for temperature measurement and control. The pyrolysis gases are transported by an adjustable primary air stream (150ml/min) through the heatable quartz glass tube and then diluted by a secondary air stream (0-400 ml/min) to adjust the pyrolysis gas concentration low. The gas mixture is directed into a measuring chamber which can house up to three sensor arrays and a humidity sensor. Pyrolysis experiments were conducted with PVC isolated litz copper wires (LiY – 0.14 mm^2, 1.2 mm outer diameter, 2 A current rating, yellow).

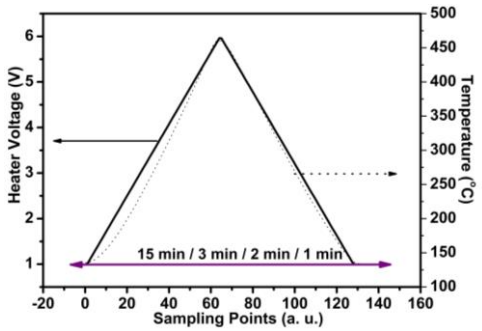


Fig. 3. Triangular heater voltage profile applied to operate the sensor array at thermo-cyclic mode (left ordinate) and corresponding temperature change monitored by IR camera (right ordinate).

For gas sensitivity studies the sensor arrays were run in thermo-cyclic operation mode. Instead of having a fixed working temperature, the sensor temperature was varied between 100°C and 450°C, by means of either corresponding triangular heater voltages (Fig. 3) or an active control loop employing the integrated temperature sensor for accurate control, while sampling conductance of the sensitive layers resulting in conductance-over-time-profiles (CTPs) with typically 128 sampling points per cycle. This is beneficial as conductance is highly estimated by adsorption/ desorption, surface chemical reactions, diffusion, which are all considerably temperature dependent, however, in an individual manner. By cycling temperature it is possible to separate these effects to some extent, resolving the otherwise singular information of sensor conductance into a profile with distinct shape and amplitude depending on kind of gas components and their concentrations and thereby enabling a more detailed analysis of the sensor response. In addition, CTP shape is highly dependent

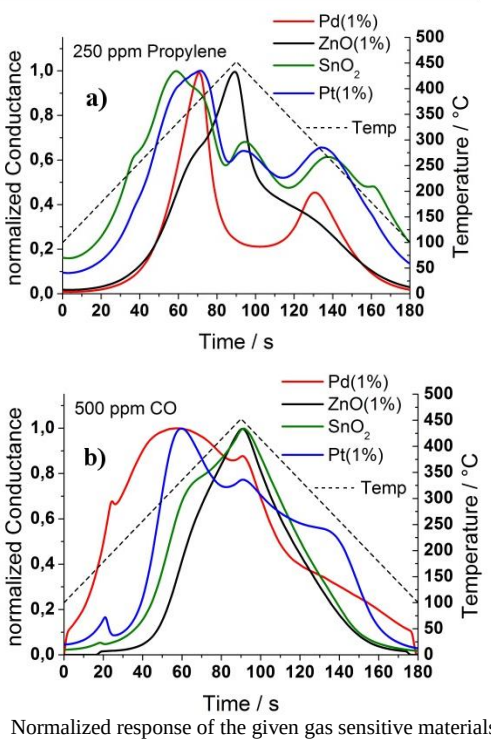


Fig. 4. Normalized response of the given gas sensitive materials on Pt-IDE to (a) 250ppm propylene and (b) 500 ppm CO with according set temperature

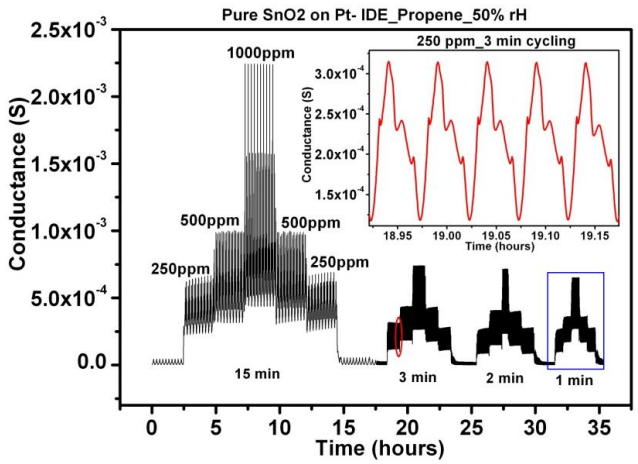


Fig. 5. Overall conductance measurement from SnO_2 layer on Pt-IDE to propene at different cycle times 15, 3, 2 and 1 minutes and gas concentrations 250, 500 and 1000 ppm. Out of 20, 5 profiles at 250 ppm at 3 min cycling (oval in red) were given in the inset to show the reproducibility of the profiles with stable baseline and maximum conductance at repeated temperature cycles.

on sensitive material composition and electrode material [6]. By extracting features from the CTP shape an identification of the incident gas is possible, while the CTP integral is typically related to gas concentration [9]. When applying this technique on the four-fold array, four individual CTPs are acquired, enhancing the ability for identification and therefore improving the analysis of complex gas mixtures.

III. Results and Discussion

A. Tests with model gases

Screening tests with model gases were conducted to gain insight about general sensitivity and stability of the layers prepared. Normalized CTPs, so as to focus on CTP shape, are given in Fig. 4 for 250 ppm Propylene and 500 ppm CO, respectively. A cycle time of 3 min was chosen. It is obvious that the four layers react quite differently to the gases, even though the amount of material added is quite low (1wt.%). From these data it is even possible to separate between the two gases by just looking at one individual layers response. The peak that is common to most of the layers at 200°C when exposed to CO can be attributed to catalytically promoted reactions at the Pt-electrode/SnO_2 interface, as changing electrode material from platinum to gold eliminates this feature. For evaluation of stablity and sensitivity of the proposed method a whole sequence of sensing tests at different concentrations of propene and cycle times is presented for a pure SnO_2-layer representatively for the other layers (Fig. 5). It shows the excellent repeatablity of the CTPs aquired, the good reproducibility with concentration changes, but also a considerable conductance drop especially when reducing cycle time from 15min to 3min. This indicates, that the CTPs are extensively influenced by non-steady state effects like diffusion etc. as well.

As a side note it should be mentionend, that the minimum cycle time to be considered is limited by two factors. One is the actually achievable temperature rate. Due to the thermal capacity of the sensor element it is not possible to reach cycle times lower than one minute without active cooling. Cycle time can be further reduced when using micro machined structures, i.e. silicon micro hotplates. Second, CTPs with short cycle times may lose features seen in longer cycles, as surface chemical reactions may not be fast enough and the influence of diffusion effects on the CTP may depend on the temperature rate as well.

B. Application to early fire detection

Pyrolysis gases were generated facilitating both heating methods described in the experimental part. First, generation of emissions was conducted by overloading the wire sample with excess electrical currents. The loading current was increased stepwise starting from 2 A up to 16 A while monitoring for visible changes in color. No change in sample morphology was visible up to 14 A, whereas discoloring occurred at 16 A which turned into severe charring. For comparison, these tests were complemented by emission experiments and sensitivity studies with externally heated polymers. For this purpose, the reactor temperature was increased stepwise, starting from room temperature up to 200°C. Optical sample monitoring was not possible through the thermal insulation of the reactor in this case and was therefore conducted offline after completion of the experiments. Evidently, the sample showed no visible change up to temperatures of 150°C and even at 170°C we can only see a slight change of the sample by shrinking of insulation diameter. At 200°C discoloring takes place, the sample turns brown.

A general overview of CTPs measured on one of the sensitive layers (SnO_2/(1%)Pd) while conducting the current/heat treatment of the isolated litz (PVC) is representatively shown in Fig. 6 and compared with the CTPs of model propylene gas. The absolute values presented here vary greatly depending on experimental parameters. Several CTPs were recorded at a reactor temperature of 170°C, while changing the dilution level to set defined relative concentrations. At 200°C the CTP acquired is quite similar to the one measured at 14 A current excitation, while overall conductance is higher by about a factor of ~10 compared to 170°C. The charring of the sample observed at 16 A is clearly detected by the sensor. As referencing of pyrolysis gas composition by chemical analysis was not possible in these

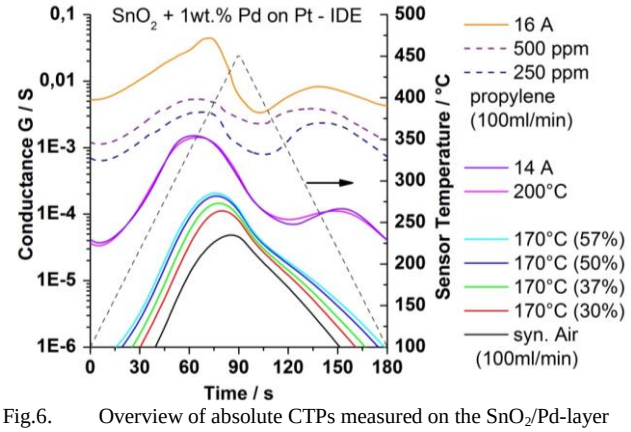


Fig.6. Overview of absolute CTPs measured on the SnO_2/Pd-layer for pyrolysis and model gas experiments. At 170°C relative pyrolysis gas concentration, as estimated by dilution with synthetic air, is given in brackets

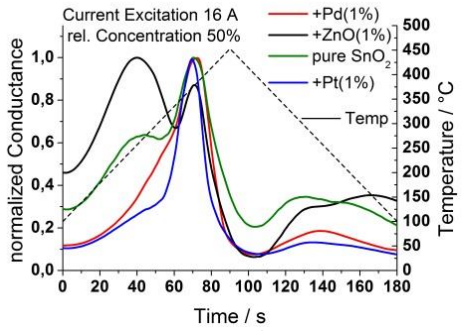


Fig. 7. Normalized sensor response of the layers SnO2/Pd, SnO2/ZnO, SnO2, and SnO2/Pt at 16 A current level with 1:1 dilution by synthetic air

preliminary experiments, CTPs of measurements with propylene are given as an orientation in measured signal amplitudes. Also the huge dynamic of the sensor signal is visible as conductance varies over several orders of magnitude. These data may be compared with those reported in [4] where resistance response Ra/Rg of isothermally operated MOG-layers to pyrolysis gases of heated PVC of about three and six were reported at 180°C and 200°C, respectiviely.

Fig. 7 shows the CTPs measured on different sensitive materials in the scenario of current excitation. The CTPs are normalized to get a better understanding of profile shape rather than amplitude, as basic layer resistance and sensitivity differ between materials. At an excitation current of 16 A the individual profile shapes differ greatly. While a main CTP-peak is visible for all materials at 350°C additional individual features can be observed in both rising and falling slope. Considering the increase in conductance seen in Fig. 6, this correlates with the observed charring of the sample, hinting to a distinct composition of the emitted gas mixture.

Based on the concentration-dependent measurement data yielded by these experiments, three data sets of one of the layers (SnO_2/(1%)Pd) of the sensor array (0%, 27,5% and 57%, rel. concentration at 170°C) were chosen for establishing and evaluation of the mathematical calibration model using ProSens (Fig. 8a) [5]. Based on this calibration model two further data sets (35% und 50%) were analyzed for testing. To identify an unknown gas sample ProSens calculates a so-called theoretical CTP and compares this curve with the measured CTP. A "small" difference between the two curves means that the unknown sample is the gas under consideration. Using this concept for gas identification, false alarms and misleading conclusions can be avoided. Fig. 8b shows the measured CTP and theoretical CTP for 50% rel. concentration. It can be clearly seen that the difference between the two curves is really very small. An analogous result can be found analyzing the sample at 35% concentration.

While showing very promising sample identification, the concentration analysis using ProSens is also very good. In both cases the relative error between dosed value and analyzed value is smaller than 8%, as can be seen in Table 1.

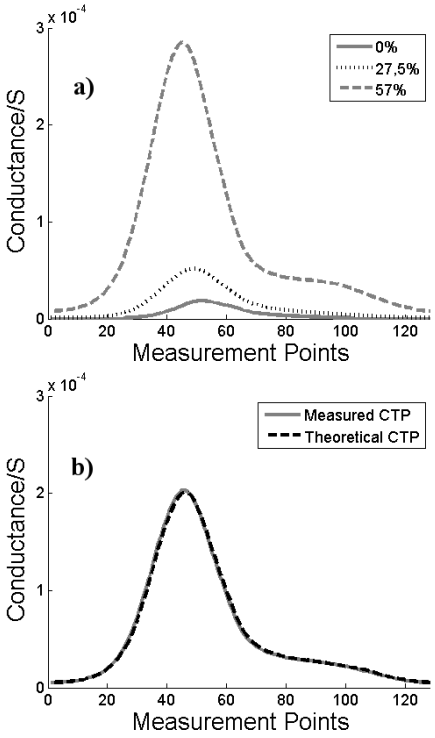


Fig. 8. Concentration-dependent CTP data of a) PVC at 170°C measured on the SnO_2/Pd-layer and b) comparison with theoretical CTP calculated with ProSens.

TABLE I. ANALYSIS RESULTS

Dosed Values	Analyzed Values	
35%	37,54%	(7,3%)
50%	46,46%	(7,1%)

IV. CONCLUSION

Thermo-cyclic method of operation of MOG arrays is an elegant method enabling the identification and analysis of gas mixtures by numerical analysis of specific CTPs. For the possibility of detecting developing fires in electrical cabinets at early stages, this method can be used to obtain characteristic patterns from gas emissions produced by overheated cable insulation materials. We could find that, with our test setup, emissions of heated PVC based insulation materials show well reproducible CTPs, even at temperatures where no color change of the sample could be observed. These results look promising considering the aim of early fire detection with high sensitivity. The acquired CTPs were numerically analyzed employing the ProSens algorithm and the results showed good identification capabilities and concentration estimation, which can lead to better incident identification and a more robust detection with low probability of generating false alarms. For field applications a microprocessor-based sensor system was devised, featuring the necessary means for thermo-cyclic operation and data analysis. Future steps focus on the optimal selection of four gas sensing materials on a single chip for early fire detection with high-quality discrimination of harmless atmospheres (cross sensitivity), the investigation of other insulation materials, as well as referencing the emitted gas

mixtures with a HT-FTIR system to get a better understanding about gas components emitted and their relation to CTP generation.

ACKNOWLEDGMENT

The authors would like to thank Dr. Matthias Schwotzer (Institute of Functional Interfaces, Karlsruhe Institute of Technology (KIT), for contributing with the ESEM-image and the German Federal Ministry of Education and Research (BMBF) for financial support of this work (Project-Nr. 16N12262).

REFERENCES

[1] D. Gutmacher, C. Foelmli, W. Vollenweider, U. Hoefer, J. Wöllenstein, "Comparison of gas sensor technologies for fire gas detection", Procedia Engineering 25, 1121–1124, 2011

[2] G. Korotchenkov, V. Brynzari, S. Dmitriev; SnO2 thin film gas sensors for fire-alarm systems, Sensors and Actuators B 54, 191-196, 1999

[3] K. Frank, V. Magapu, V. Schindler, H. Kohler, H.B. Keller, R. Seifert, "Chemical Analysis with Tin Oxide Gas Sensors: Choice of Additives, Method of Operation and Analysis of Numerical Signal", Sensor Letters 6, 908–911, 2008

[4] Ye Zhao, Shunping Zhang, Guozhu Zhang, Xiaoshuang Deng, Changsheng Xie; Highly sensitive porous metal oxide films for early detection of electrical fire: Surface modification and high throughput screening, Sensors and Actuators B: Chemical 191, 431-437, (2014)

[5] R. Seifert, H. B. Keller, K. Frank, H. Kohler, "ProSens - an Efficient Mathematical Procedure for Calibration and Evaluation of Tin Oxide Gas Sensor Data", Sensor Letters 9/1, 7-10, 2011

[6] N. Illyaskutty, J. Knoblauch, M. Schwotzer and H.Kohler, "Thermally modulated multi sensor arrays of SnO2/additive/electrode combinations for gas identification", submitted to Sensors and Actuators B, 2014

[7] A. Jerger, H. Kohler, F. Becker, H.B. Keller, R. Seifert, New applications of tin oxide gas sensors. II: Intelligent sensor system for reliable monitoring of ammonia leakage, Sensors and Actuators B 81, 301–307, 2002.

[8] J. Knoblauch, N. Illyaskutty, and H.Kohler, " Early Detection of Fires in Electrical Installations by Thermally Modulated SnO2/Additive-Multi Sensor Arrays", submitted to Sensors and Actuators B, 2014

[9] R. Seifert, H.B. Keller, H. Kohler, "SimSens – a New Mathematical Procedure for Simultaneous Analysis of Gases with Resistive Gas Sensors", submitted to Sensors and Actuators B, 2014

Line Encoding for 25 Gbps over one Pair Balanced Cabling

Katharina Seitz, MSc - Prof. Dr.-Ing. Albrecht Oehler
Reutlingen University
Reutlingen, Germany

***Abstract*—In this paper, research projects with 30 meter balanced cabling and data rates up to 25 Gbps over one single pair are described. The project aim is to achieve 100 Gbps via a four pair balanced cabling channel. In the following, spectral characteristics of the used prototype twisted pair are presented. Therefore, the insertion loss of the single cable in comparison to the insertion loss of the cable in combination with an equalizing amplifier, as well as the group delay of the cable and the cable connected to the equalizing amplifier is shown. Furthermore, a carrierless Pulse Amplitude Modulation with 32 different levels (PAM-32) as an approach for a possible line encoding is presented. Finally, research measurements of the data transmission with a data rate up to 25 Gbps via shielded twisted pair is shown.**

Keywords—Balanced cabling; twisted pair, 100 Gbps data rate; prototype cables; line encoding.

I. INTRODUCTION

The consumer's request is to get more data within a shorter period of time. Therefore the data rates are crucial for the consumers and it is important to get the optimum information out of the existing network as well as knowing the requirements for the future. The request of higher data rates advances innovations and new technologies.

Although it is possible to realize higher data rates with fiber optic cables, there are some applications where copper cables are necessary or easier to realize. Therefore, maximum data rates of copper cables are examined here.

Currently, it is possible to transmit 10 Gbps over 100 m balanced cabling [1]. The corresponding cable category is the cable category 6_A. Under development at IEEE 802.3bq, there is a new technology that transmits 40 Gbps (40 GBASE-T) over four pair balanced cabling with a length of 30 meter, requiring at least category 8.1 [2][3].

The aim of this research project is to achieve the data transmission of 100 Gbps over a four pair balanced cabling system. In ITG-Fachbericht 232 channel capacity of different balanced cabling channels are provided [4]. This analysis shows a sufficient channel capacity for the system of this investigation.

The paper is divided into three parts. The first part covers the characteristic behavior of the prototype cable concerning the insertion loss and the group delay. It describes the behavior of the single cable and thereafter the insertion loss and the group delay of the cable connected with the equalizing amplifier.

The second part covers an approach for line encoding to realize a data transmission of 25 Gbps over a one pair balanced cable. The presented line encoding is called carrierless Pulse Amplitude Modulation.

The third and last part deals with the first measurement results of a data transmission of 25 Gbps per pair. The transmission over a four pair balanced cable results in the desired data rate of 100 Gbps. To obtain these measurements, a twisted pair with individually shielded pairs with a length of 30 meter was used. This reflects a typical application in data centers.

II. INSERTION LOSS AND GROUP DELAY

This chapter characterizes the experimental copper cable concerning the insertion loss and the group delay. For the following measurements, the same prototype twisted pair was utilized. It has a length of 30 m and each pair of the four pair cable is single shielded. The used cable fulfils the cable category 7_A requirements of EN 50288-9-1, which are specified up to 1 GHz [5].

Each transmission is primarily influenced by attenuation. Due to the skin-effect the cable has a low-pass characteristic. Figure 1 shows the insertion loss of the prototype cable used for the measurements in the last part of this paper.

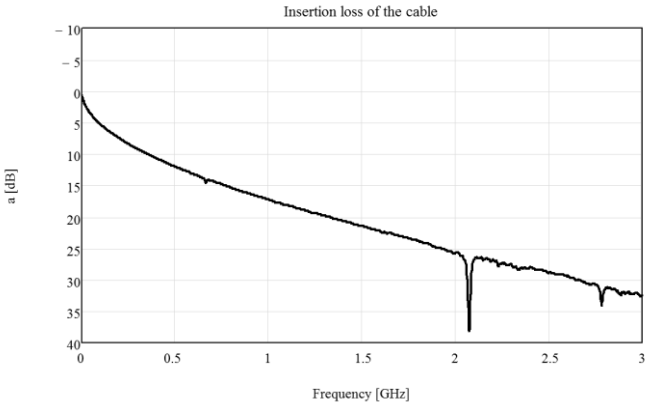


Fig. 1. Insertion loss of the cable

The measurement concerning the insertion loss of the cable shows this low-pass behavior of the cable. The attenuation is increasing smoothly as expected until about 2 GHz. At 2 GHz the attenuation is about 25 dB. Compared to other copper cables in the category of 7_A it has a higher usable bandwidth.

To counteract the insertion loss of the cable, a special equalizing amplifier was developed. The aim was to minimize the insertion loss by flattening the transmission characteristics of the used copper cable.

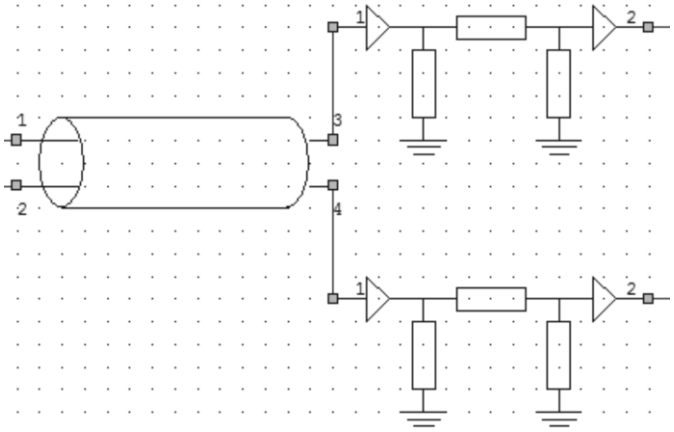


Fig. 2. A wires pair and it's compensator

Figure 2 shows the construction of the equalizing amplifier. It is divided into two identical compensation branches, one for each conductor. It consists of a preamplifier, a flattening attenuator and a post amplifier.

The result of the insertion loss measurement of the cable connected to the prototype equalizing amplifier is shown in Figure 3.

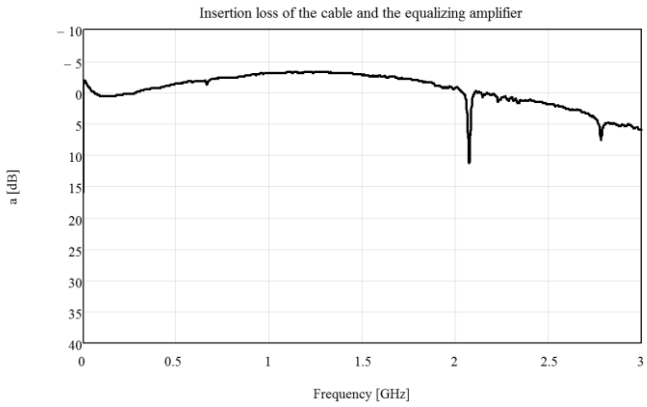


Fig. 3. Insertion loss of the cable and the equalizing amplifier

Figure 3 illustrates the positive effect of the connected equalizing amplifier. Now the deviation up to a frequency of 2 GHz is about 4 dB dynamic instead of 25 dB dynamic without the equalizing amplifier. Therefore it results in a flatten transmission over a 2 GHz frequency range.

Figure 4 shows the group delay of the prototype cable and the cable connected to the equalizing amplifier.

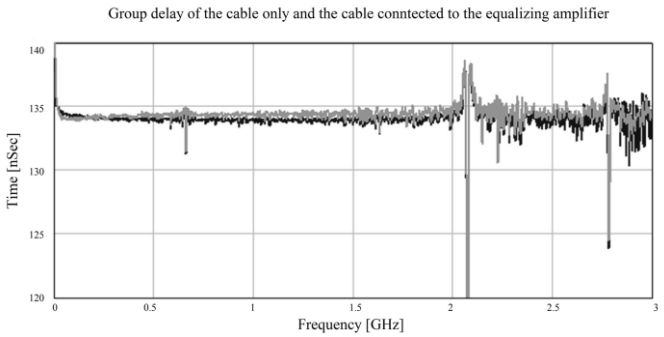


Fig. 4. Group Delay: black line: 30 m cable only, gray line: cable and equalizing amplifier

The Group Delay is steady up to a frequency of more than 2 GHz. This is similar to the already presented insertion loss measurements. The maximum deviation, up to a frequency of 2 GHz, is 4 nanoseconds.

III. LINE ENCODING

Line encoding is used to transfer information. Therefore a digital logic has to be transferred so that it is possible to use the physical attributes of the channel.

For the line encoding, a carrierless Pulse Amplitude Modulation (PAM) with 32 different relative levels of amplitude to transfer the information was selected. The PAM with 32 levels is a multilevel line encoding [1]. One symbol, represented by one of the 32 relative levels is equal to 5 bits. Therefore, the symbol period is five bits. The resulting symbol rate for a data rate of 25 Gbps is about 5 GBd.

The selected line encoding is based on a minimum symbol rate to achieve a minimum transmission bandwidth. For the transmission the PAM was used with a specially formed Raised-Cosine impulse. This impulse is defined as:

$$g_i(t) = \frac{\sin\left[\frac{\pi t}{T_S}(1-\alpha)\right] + \frac{4\alpha t}{T_S}\cos[\frac{\pi t}{T_S}(1-\alpha)]}{\frac{\pi t}{T_S}[1-\left(\frac{4\alpha t}{T_S}\right)^2]} \quad (1)$$

with $\alpha = 0.15$.

T_S defines the symbol period [6].

To realize a transmission of 25 Gbps over one twisted pair of a four pair channel, the Raised-Cosine impulse ($g_i(t)$) was factored with the 32 different levels of amplitude.

IV. Data Transmission

A. General

In this last part, the first results of the transmission with the data rate of 25 Gbps is presented using two different measurement setups. Both setups will be explained in detail in the following.

Both measurement setups consist of the Tektronix Arbitrary Waveform Generator AWG 70001A and the Tektronix Sampling Oscilloscope CSA 8000 or the Teledyne LeCroy Real-Time Oscilloscope SDA 820Zi-A.

The Waveform Generator generates the desired waveform to transmit the data. For the first measurement setup it is connected directly to the oscilloscope. This setup is used to receive the generated signal at the output of the arbitrary waveform generator. In the following the measurement of this setup is called transmitted signal.

The second measurement setup includes the copper cable and the equalizing amplifier, which is connected at the end of the cable. The measurement of the second setup is called received signal.

B. Measurement results

For the first measurements the PAM was used as a line encoding as previously described. The selected test sequence of signals included all 32 possible levels in decrementing order. That means the first symbol is the symbol with the largest positive amplitude, followed by the symbol with the largest negative amplitude. This process is repeated in decreasing order until the symbol with the smallest positive and negative amplitudes have been sent.

Figure 5 shows the result of the first measurement setup. This represents the transmitted signal at the output of the arbitrary waveform generator.

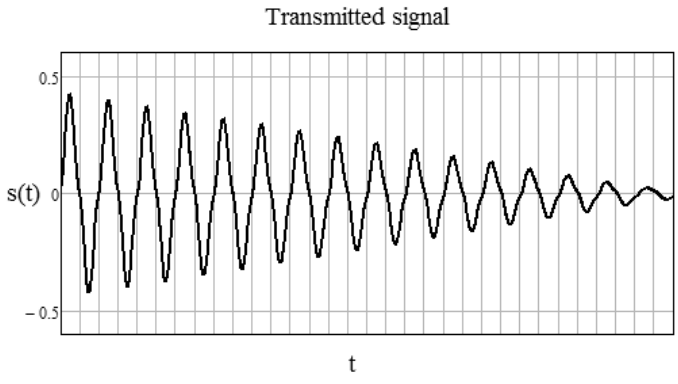


Fig. 5. Transmitted PAM 32 signal (200 ps per division and 500mV per division)

It illustrates that all 32 different levels of amplitude can be easily distinguished. The selected test sequences could be completely decoded.

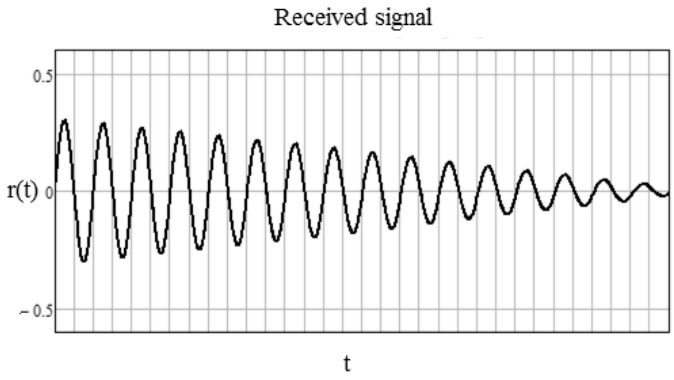


Fig. 6. Received PAM 32 signal (200 ps per division and 500mV per division)

Figure 6 shows the received signal. Similar to the transmitted signal, the received signal could be decoded with all 32 different amplitude levels [7].

The comparison of the different symbols of the transmitted and the received signals are shown in Figure 7.

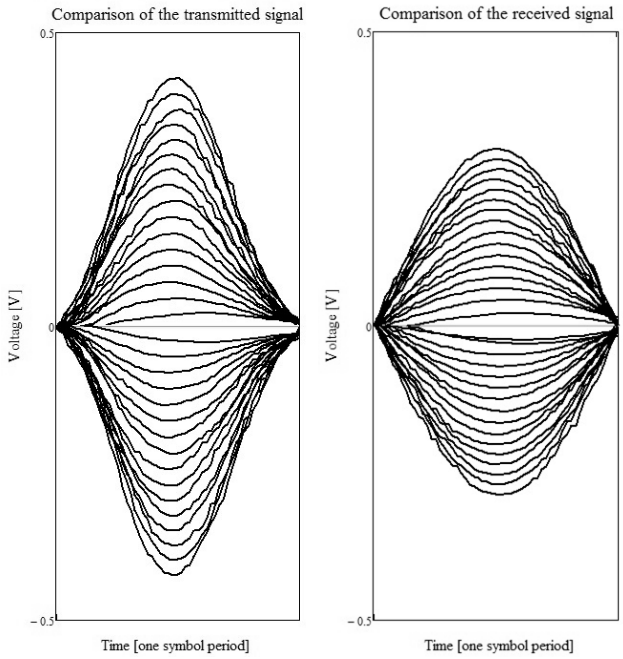


Fig. 7. Comparison of the transmitted and the received signal

Figure 7 illustrates the distinguished levels of amplitude of both the transmitted signal and the received signal. Thus a digital transmission with 31 decision levels is possible.

The voltage loss between the transmitted signal and the received signal is induced by the attenuation of the cable at a frequency of 2.5 GHz. This attenuation is about 3 dB.

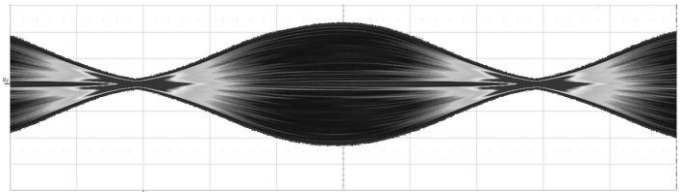


Fig. 8. Eye diagram of the transmitted signal (33.4 ps per division and 100 mV per division)

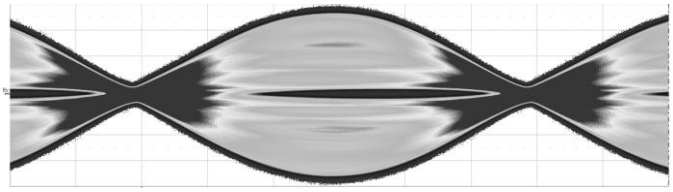


Fig. 9. Eye diagram of the received signal (33.4 ps per division and 50 mV per division)

Figure 8 shows the eye diagram of the transmitted signal, Figure 9 shows the eye diagram of the received signal. Both diagrams are not optimal because the number of effective bits used from the real time oscilloscope is limited.

V. Outlook

For the future, further measurements with other line encodings are in process. In addition, we would like to optimize the equalization of the signal, as well as the signal generation of the arbitrary waveform generator.

Furthermore, we will measure statistical sequences instead of the presented selected sequence to identify the bit error rate of the data transmission.

Additionally, we are working on eye diagrams with higher resolution.

VI. Conclusion

In this paper we first described the results of a transmission with a data rate of 25 Gpbs over one twisted pair of a four pair cable. Therefore, there was a total data rate of 100 Gbps over all four twisted pairs can be achieved.

Furthermore we characterized the spectral behavior of the experimental copper cable. We presented the insertion loss and the group delay of the single cable and compared it to the cable and the special produced equalizing amplifier.

A line encoding is introduced allowing a transmission rate of 25 Gbps over one twisted pair. The presented line encoding, the PAM, is capable of transmitting high data rates by minimizing the symbol rate and the transmission bandwidth as well as taking the noise floor into account.

Finally initial test sequences of the PAM with all different amplitude levels are presented. Conclusively, the transmission and the subsequent demodulation of transmitted signal sequence with a data rate of 25 Gbps are possible.

VII. Acknowledgments

This research project was supported by the German Federal Ministry for Economic Affairs and Energy on the basis of a decision by the German Bundestag.

Supported by:

Federal Ministry
for Economic Affairs
and Energy

on the basis of a decision
by the German Bundestag

References

[1] IEEE 802.3-2008-Section Four, Standard for information technology - Telecommunications and information exchange between systems, 2008

[2] ISO/IEC TR 11801-9901:2014-10, Information technology – Generic cabling for customer premises – Part 9901: Guidance for balanced cabling in support of at least 40 Gbit/s data transmission, 2014

[3] VDE-AR-E 2800-902, Informationstechnik – Symmetrische vierpaarige Übertragungsstrecke für Datenraten mit mindestens 40 Gbit/s, 2014

[4] ITG-Fachbericht 232, Kommunikationskabelnetze - Vorträge der 18. ITG-Fachtagung vom 13. Bis 14. Dezember 2011 in Köln, 2011

[5] DIN EN 50288-9-1 VDE 0819-9-1:2013-11, Mehradrige metallische Daten- und Kontrollkabel für analoge und digitale Übertragung – Teil 9-1: Rahmenspezifikation für geschirmte Kabel bis 1000 MHz – Kabel für den Horizontal- und Steigbereich, 2013

[6] The ATM Forum - Technical Committee, 155.52 Mb/s Physical Layer Specification for Category-3 Unshielded Twisted Pair- af-phy-0047.000, 1995.

[7] IWCS, presented at the 63rd IWCS Conference form November 9-12, 2014 in Providence, Rhode Island, USA, 2014

MOEMS based concept for miniaturized monochromators, spectrometers and tunable light sources

Ulrich Mescheder, Isman Khazi, Andras Kovacs, Alexey Ivanov
Institute for applied research, Faculty of Mechanical and Medical Engineering
Hochschule Furtwangen University
Robert Gerwig Platz 1, 78120 Furtwangen, Germany
mes@hs-furtwangen.de

***Abstract*— A novel concept for MOEMS based system including porous-silicon-based photonic crystals which forms the heart of miniaturized monochromators, spectrometers and tunable light sources is presented. The fabricated porous-silicon based 1D photonic crystal is tuned with the combination of fast micromechanical tilting and pore-filling of the porous silicon multilayer, thereby providing the wavelength tuning of ca. ± 20% around the working wavelength. The optical characterization of the photonic crystal for its spectral behavior is done using the commercial Essential Macleod simulation software. Experimental and simulation data for the visible and near-infrared wavelength range supporting the proposed approach are also shown.**

Keywords—MOEMS; MEMS; Porous silicon; Photonic crystals; Dual tunability; Silicon anodization; Tunable optical filter; Monochromator; Spectrometer; Biosensor

I. INTRODUCTION

Porous-silicon multilayers based photonic crystals opened vast avenues of new applications; they are well suitable for the formation of interference filters such as Distributed Bragg reflectors (DBR), rugate filters and Fabry-Perot interferometers (FPI) [1]. They can be molded into combination of layers with varied microstructures in depth resulting in different optical properties, thereby permitting the design of complex optical components [2]. On contrary to thin film deposited devices, porous-silicon based photonic crystals has the capability of forming multi layers of varied thickness and refractive index in a simple wet chemical process which is compatible to the MEMS technology, thereby making it possible to design filters which are compact, cost effective, precise, miniaturized, easy to fabricate [1] and can be easily integrated to form MOEMS based system for miniaturized spectrometers, monochromators and tunable light sources.

The optical filters especially tunable filters are used for various optoelectrical, biological sensing and chemical spectroscopy applications. In optical applications it is employed for wavelength selection, manipulation and reconfiguration of optical networks in Dense Wavelength Division Multiplexing (DWDM), and hence forms a vital component of such an optical fiber transmission system [3]; hyper spectral imaging [4] and sensor spectroscopy [5]. In the area of chemical and biological applications it helps in chemical analysis, food safety, quality monitoring and bio-component detection as being the fundamental component of micro spectrometer.

Tunable optical filters (TOF) based on FPI have been reported in [6], where the working wavelength is tuned by changing the gap between the involved mirrors, but an extreme precise control of the micromechanical movement is required. In the work of [7], a thermal actuation approach is employed to change the refractive medium inside the FPI, which provides tuning of the working wavelength but with the limitation of very low tuning range and a very low frequency response. PS based TOF was reported by Lammel et al [8], where the flip-up optical filter was tilted and tuned by two sophisticated thermal bimorph microactuators, but with the limitation of precise control over the tilt position. In a similar approach, Ruminski et al. [9] demonstrated spectral wavelength shifts of PS based photonic crystals resulting from tilting and irreversible pore-filling with polystyrene as optical reference.

In this proposed work, the working wavelength λ_c of the PS based 1D photonic crystal is tuned by a very novel approach of dual tunability, which involves the combination of fast mechanical tilting and reversible pore-filling of the photonic crystal by liquids or gases. This MOEMS based TOF system when compared to above approaches results in wavelength tuning of ca. ± 20% around the working wavelength. In this work, we also present extensive optical simulation results of the spectral behavior of the photonic crystals for optical applications. A fabrication method of PS based photonic crystal is also discussed and a MOEMS based system integration concept is also presented.

II. BACKGROUND

The first Bragg reflectors based on PS was first demonstrated in 1990 [10]. It is the most common optical filter which is also referred to as dielectric or distributed Bragg mirror. PS based 1D photonic crystals forming Bragg filters, rugate filters, microcavities or other optical components show a pronounced resonant peak of the stop band or a sharp resonant fall-off within the stop band. In case of DBR with layers of

alternating high and low refractive indices n_L and n_H, the position of the resonance peak (central working wavelength λ_c) is given by:

$$n_H d_H = \frac{\lambda_c}{4} = n_L d_L \quad (1)$$

where d_H, d_L are the thickness of high and low refractive index layer n_H and n_L respectively, satisfying quarter wave condition for center wavelength λ_c. In PSMLs, incident light beam is reflected from the different interfaces of the multi layers, which interfere constructively for a specific wavelength and at this wavelength there is maximum reflection and is forbidden in transmittance and hence also referred to as stopband.

The bandwidth of the stop band $\Delta\lambda$ around the central wavelength λ_c can be selected by the proper adjustment of n_H and n_L and is given for DBR by [6]:

$$\frac{\Delta\lambda}{\lambda_c} = \frac{4}{\pi}.\sin^{-1}\frac{n_H - n_L}{n_H + n_L} \quad (2)$$

The shift of the central wavelength λ_c to λ_θ in transmission or reflection spectrum of such a multilayer structure as function of incidence angle (θ) can be described with the Bragg´s law [6]:

$$\lambda_\theta = \lambda_c \sqrt{1 - \left(\frac{\sin\theta}{n}\right)^2} \quad (3)$$

$$\lambda_c = 2dn \quad (4)$$

where d is the thickness of a period of the two layers with low and high refractive index (d= d_H + d_L) and n is the effective refractive index of the porous layer.

According to equation (3), fast tuning of some hundreds of nm to shorter wavelengths (blue shift) of the center peak position λ_c can be achieved by a relatively large rotation (up to 20° - 40°) of the photonic crystal in respect to the incident light. However, the maximal rotation which maintains the optical filter properties has to be considered. With a micromechanical actuation, fast rotation frequencies of about 1 kHz are possible.

By pore-filling of the porous optical filter with different gases or liquids (organic or aqueous solutions), wavelength shift to longer wavelengths (red shift) of the central wavelength can be achieved. This shift is due to increase of the effective refractive index of the porous silicon by pore-filling. It is important to note that the response time for this tuning principle is limited by the transport processes in nanostructured layers [11] which is in the range of few seconds without heating or cooling.

III. Optical Simulation of PS based 1D Photonic Crystal

The simulations were done in the commercial Essential Macleod optical simulation software to simulate and characterize the multilayers for optical filter applications and to study the tunability based on tilting. The normal DBR

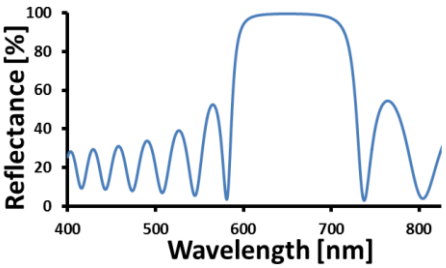


Fig. 1.(a) Simulated DBR spectrum for 20 alternating layers for p^+ at λ_c = 650 nm

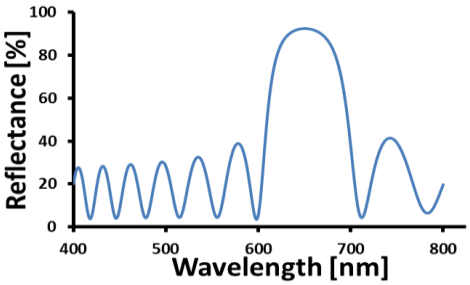


Fig. 1.(b) Simulated DBR spectrum for 20 alternating layers for p^- at λ_c = 650 nm

structures were simulated pertaining to refractive index (RI) contrast for heavily doped (0.01-0.02 Ω-cm) low resistive p^+ Si based fabricated PS photonic crystals and lightly doped (10-20 Ω-cm) high resistive p^- Si based PS photonic crystals, where the RI of the individual layer was computed using Bruggeman's effective medium approximation model after fabrication of the PS photonic crystals. Fig.1(a) and (b) shows the simulated DBR with 10 periods for p^+ and p^- respectively, with center wavelength λ_c at 650 nm. The bandwidth of p^- based DBR is narrower when compared to p^+ based DBR, because of low RI ratio for the former case when compared to high RI ratio in the latter case. Fig. 2 shows the influence of number of layers in the multilayer stack of the photonic crystal on the bandwidth and intensity of the stopband peak. As the number of layers increases, the stopband peak becomes narrower and its reflectance increases. Fig. 3 shows the influence of RI ratio of the alternating high and low RI layers on the spectral characteristics of the photonic crystal. As the RI ratio decreases the stopband peaks becomes narrower and its reflectance decreases.

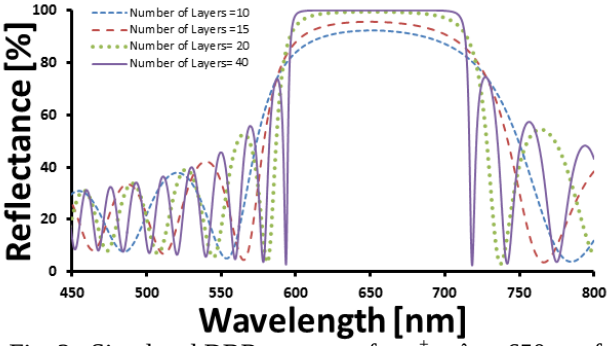


Fig. 2. Simulated DBR spectrum for p^+ at λ_c = 650 nm for different number of layers

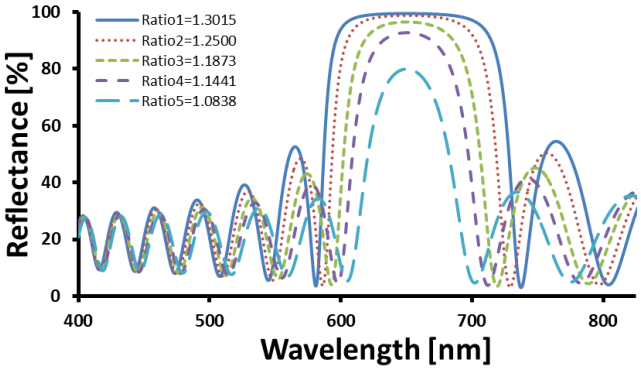


Fig. 3. Simulated DBR with 20 layers at λ_c =650 nm with different RI ratios $\left(\frac{n_H}{n_L}\right)$

Fig. 4 shows the simulation of Rugate based PSML with a narrow stopband peak. The rugate filter is simulated by continuous and periodic modulation of current density (e.g. sinusoidal), which results in periodic and continuous transition of RI between the high RI layer and low RI layer.

Apodization of the periodic and continuous modulation of the RI contrast eliminates the sidelobes. While the higher order harmonics are suppressed by index matching of the Si to the boundary conditions (in this case air and Si substrate) at the respective interfaces.

The narrowband tunability of the photonic crystal by implementing the very novel approach of tilting is also simulated. Fig. 5 (a) and (b) shows the tunability of the stopband peak on tilting the PS based photonic crystal with respect to normal incident light source for both p^+ and p^- photonic crystal respectively. At 0° position the stopband peak is centered at 650 nm. As the photonic crystal is tilted from 0° to 10° and further, the center wavelength shifts to shorter wavelength i.e. center wavelength experiences a blue-shift. The simulation results shows that the wavelength shift for the p^- is wider when compared to the p^+ photonic crystal at the same tilt angle.

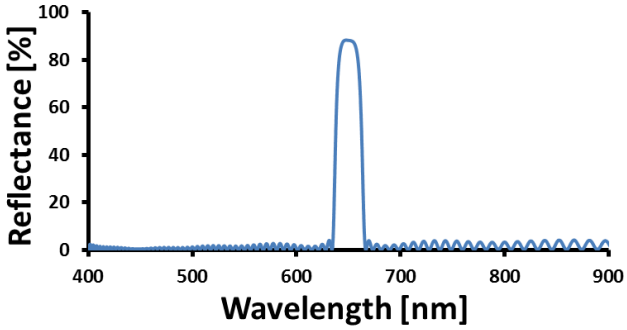


Fig. 4. Simulated Rugate filter with apodization and index matching at the boundaries with λ_c = 650 nm [12].

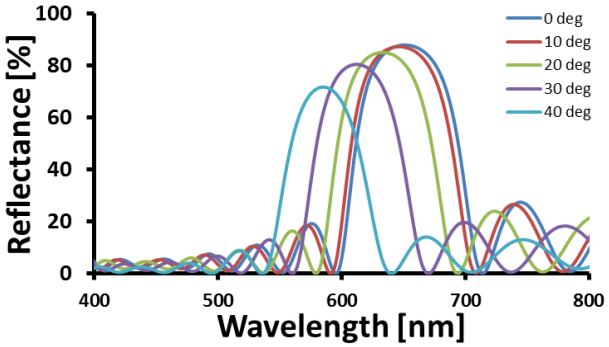


Fig. 5.(a) Simulated shift in the center wavelength position in case of p^+ PS based photonic crystals due to tilt technique.

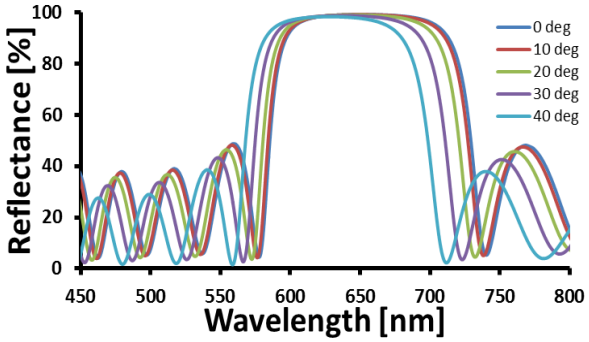


Fig. 5.(b) Simulated shift in the center wavelength position in case of p^+ PS based photonic crystals due to tilt technique.

IV. Fabrication of Photonic Crystal

The photonic crystals are defined in the surface of p-type boron doped one-side polished silicon wafers (10-20 Ω•cm). Backside (not polished side) is doped additionally with boron by ion implantation to achieve low sheet resistance of about 24 Ω/□ in order to provide good electrical contact of wafer's backside to the HF-electrolyte during the anodization process. The samples were anodized at room temperature in a double-tank cell where the wafer is forming a virtual anode (therefore this process is called anodization). Electrolyte mixture of 1:1 volume ratio of 50 wt. %HF and pure ethanol is used. For proof of concept two types of photonic crystals were realized – DBR and Rugate filters. The DBR filters comprised of 20 porous layers with alternating low and high refractive index. The Rugate filters were fabricated by sinusoidal modulation of refractive index with 16 and 32 periods. The time dependent current profiles for anodization were calculated based on experimentally determined dependencies on current density of the effective refractive index (calculated using the Bruggemann model [2] from porosity values) and of porous silicon formation rate. Current density for all filters fabricated in this work was set between 20 mA/cm² and 70 mA/cm². All photonic crystals were designed and fabricated to have a central wavelength λ_c in the visible spectrum.

V. Experimental Setup for Measuring the Tunability

For testing the properties of the tunable filters, a set-up with optical fibres for in- and outcoupling of light was used. Light reflected from the photonic crystal was guided to the spectrometer by a fiber. The entire setup was assembled on an optical breadboard with all components being firmly fixed to avoid vibrations. The fabricated p^- photonic crystal (DBR with λ_c =650 nm) was attached to a holder which was fixed on the rotational mount. Three different setup approaches were used to measure the tunability as shown in fig. 6, in order to provide different possibilities for fabrication of final miniaturized MOEMS based systems. In case of fig. 6 (b) and (c) to measure the influence of tilting of photonic crystal on the shift of the central wavelength, the rotational mount was rotated manually from 0° (normal incidence) to 50° in steps of 10°.

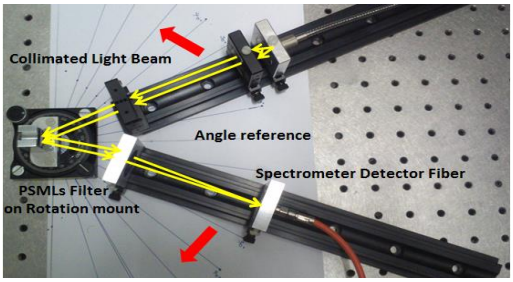


Fig.6(a) Fixed photonic crystal and movable collimated light source and detector.

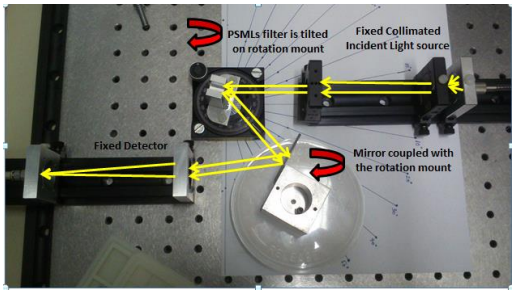


Fig.6(b) Fixed collimated incident light source and detector with movable photonic crystal and a compensation mirror.

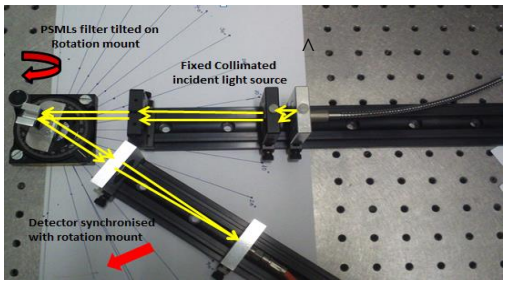


Fig.6(c) Fixed collimated incident light source with movable photonic crystal and detector.

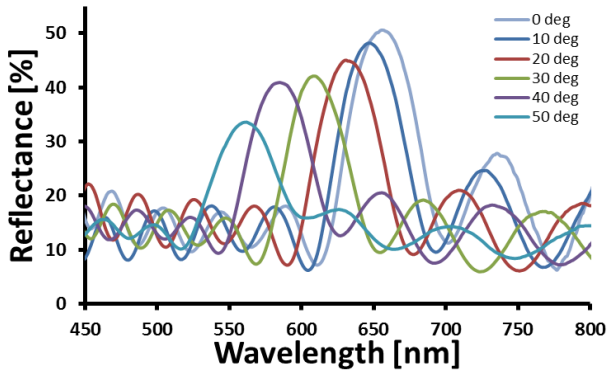


Fig. 7. (a) Experimentally measured center wavelength shift in the p^- PS based photonic crystals induced by tilting.

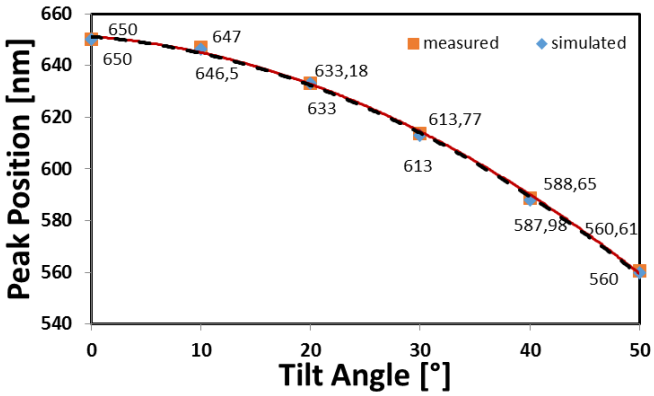


Fig. 7. (b) Comparison of simulated and measured wavelength shift positions due to tilting the photonic crystal [12].

The measured spectral shift of the central wavelength as function of tilt angle for the low doped photonic crystal is shown in fig. 7(a) and it was found in good agreement to simulation as shown in Fig. 7 (b). The experiment showed that the shift of the central wavelength as a result of tilting is instantaneous without any noticeable delay. Tunability by tilting worked well in a narrow wavelength range limited by tilting angles up to 50°. For higher tilting angles the integrity of the spectrum tended to fade away due to total internal reflection. This causes degradation of the peak (of high reflectance in case of DBR).

In order to measure the spectrum in case of pore-filling, a closed chamber with dedicated inlet and outlet orifices for vapor or liquid, an anti-reflection glass window and a holder for the porous Si photonic crystal (fabricated rugate filter) was constructed. Ethanol vapor was pumped into the closed chamber by a self-designed circulating system through the inlet orifice and left through the outlet orifice. When the photonic crystal is filled with ethanol vapor, capillary condensation within the mesoporous layers (pore size some nm) of the photonic crystal occurs and changes the refractive index contrast, thereby shifting the central wavelength to a higher wavelength (red shift). Fig. 8 shows a non-normalized spectrum of the shift in the center wavelength of the Rugate

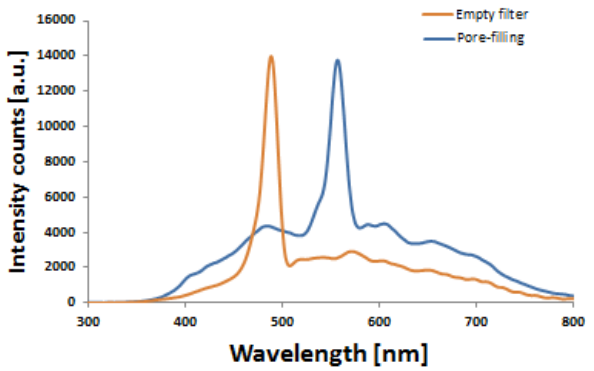


Fig. 8. Center wavelength shift in case of pore-filling of the PS based rugate photonic crystal with ethanol vapors.

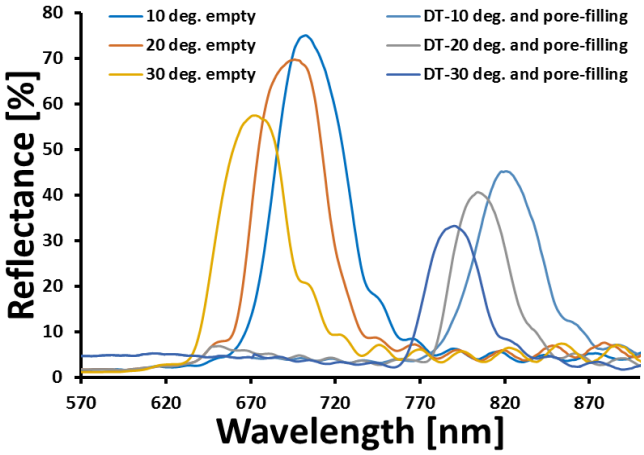


Fig. 9. Dual tunability spectrum induced by tilting and pore-filling of the photonic crystal with ethanol vapors [12].

filter in case of pore-filing of the photonic crystal with ethanol vapors. The shift of the central wavelength due to pore-filling is higher than the shift resulting due to the tilting (at least when limiting the tilt angle to +-20°).

It was also observed that spectral shift due to pore-filling is not instantaneous, but has a delay of few seconds depending on how quick the pores are filled with ethanol vapor. It has been found that adsorption is rather fast (faster than the detection limit of 1 s) whereas desorption without heating the multilayers can need several ten s.

In order to characterize the dual tunability, the spectrum of the fabricated rugate photonic crystal was measured for each tilting angle for two states. First, spectrum of the photonic crystal in the empty chamber (pores filled with air) was recorded, as shown in fig. 9. Afterwards, the chamber was filled with vapor, which resulted in capillary condensation of vapor in the pores of the photonic crystal. The right side of the spectrum in fig. 9, shows the wavelength shift due to both tilt and pore-filling techniques. However in the case of the pore-

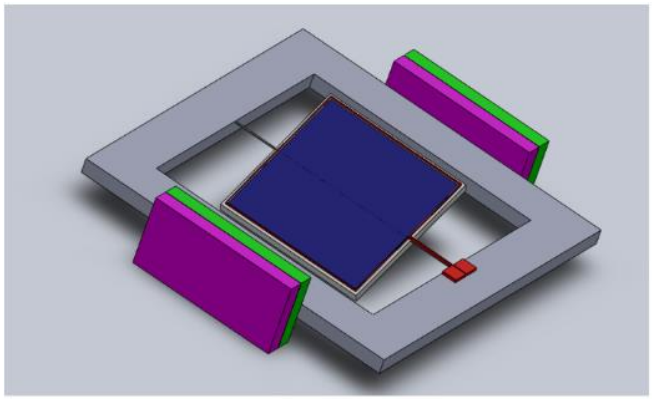


Fig. 10. Schematic set-up for on-chip actuator for tilting mechanism based on electro-magnetic actuation using two permanent magnets and a driving current flowing through a planar coil on the tilted stage in the center [12].

filling the reflectance intensity of the central wavelength decreased at the shifted wavelength position as the photonic crystal was optimized for air but not for the pore-filled state. However, for future application the optimization should be done for the pore filled state.

VI. System Concept

An important feature of the presented concept is the possibility of monolithical integration into a MEMS based actuator systems on a single chip. In principle, different actuation principles providing the needed tilting of the filters are possible. Piezoelectric, electrostatic and electro-magnetic principles were modelled and simulated. As a result, electro-magnetic principle was chosen for realization. The concept is shown in Fig. 10: The filters are defined in the surface of the suspended plate (shown in tilted situation) using SOI-technology. The two suspensions have a double function. They allow the tilt movement and contain the metal lines with which current is supplied to the planar coil on the plate. An electro-magnetic force based on the Lorenz-force is created by the current through this coil and the two outer magnets. This force tilts the mirror.

The simulation has proved that an electromagnetically actuated photonic crystal reflector suspended by square-shaped torsion beams can provide tilt angles of up to ±20° at frequencies up to KHz even for one layer metallization considering the maximal possible current density in Cu-lines. However, the needed thickness of the Cu-lines is about 10 µm which is much thicker as standard thin film technology. Therefore, an electroplating process has been developed which provide such a thickness for about 10 µm wide lines.

In the final optical setup the system is placed in a closed chamber with input and output channels for gas or liquid input/output and optical input/output.

VII. Conclusion

We presented an extensive simulation results for the optical characterization of PS based photonic crystals for optical filter applications. We also devised a very novel technique of dual-tunability to tune the working wavelength of the fabricated PS based 1D photonic crystals. With our presented approach we can get the tunability of +-20% around the working wavelength. The concept of MOEMS based system for miniaturized integration of photonic crystal with different setup possibilities and electromagnetic actuation technique is also presented.

References

[1] M. Thoenissen et al.: Dielectric filters made of porous silicon: advanced performance by oxidation and new layer structures. Mater. Res. Soc. Symp. Proc. (USA) vol.431, 1996 Pages 373-8.

[2] Andras Kovacs, Prasad Jonnalagadda, and Ulrich Mescheder: Optoelectrical Detection System Using Porous Silicon-Based Optical Multilayers. IEEE SENSORS JOURNAL, VOL. 11, NO. 10, OCTOBER 2011

[3] M.S. Borella, J.P. Jue, D. Banerjee, B. Ramamurthy, B. Mukherjee: Optical components for WDM lightwave networks. Proc. IEEE 85, 1997, Pages 1274–1307.

[4] A. Irajizad, F. Rahimi, M. Chavoshi, M.M. Ahadian: Characterization of porous poly-silicon as a gas sensor", Sensors and Actuators B: Chemical, Volume 100, Issue 3, 15 May 2004, Pages 341-346.

[5] M. Noro, K. Suzuki, N. Kishi, H. Hara, T. Watanabe, H. Iwaoka: CO2/H2O gas sensor using a tunable Fabry–Perot filter with wide wavelength range. In Proc. IEEE MEMS, Kyoto, Japan, January 19–23, 2003, Pages 319–322.

[6] Tuohiniemi M, Nasila A, Antila J, Saari H, Blomberg M: Micro-machined Fabry-Pérot interferometer for thermal infrared. In Sensors, 2013 IEEE. IEEE; 2013:1–4.

[7] Li S, Zhong S, Xu J, He F, Wu Y: Fabrication and characterization of a thermal tunable bulk-micromachined optical filter. Sensors and Actuators A: Physical 2012, 188:298–304.

[8] Lammel G, Schweizer S, Renaud P: Microspectrometer based on a tunable optical filter of porous silicon. Sensors and Actuators A: Physical 2001, 92:52–59.

[9] Ruminski AM, Barillaro G, Chaffin C, Sailor MJ: Internally Referenced Remote Sensors for HF and Cl2 Using Reactive Porous Silicon Photonic Crystals. Advanced Functional Materials 2011, 21:1511–1525.

[10] G. Vincent: Optical properties of porous silicon superlattices. Appl. Phys. Lett. 64, 1994, Page 2367.

[11] Kovacs A, Malisauskaite A, Ivanov A, Mescheder U, Wittig R: Optical sensing and analysis system based on porous layers. In The 17th International Confernece on Miniaturized Systems for Chemistry and Life Sciences (MicroTAS 2013), October 27-31, 2013, Freiburg (Germany). 2013:275–277

[12] Ulrich Mescheder, Isman Khazi, Andras Kovacs and Alexey Ivanov: Tunable optical filters with wide wavelength range based on porous multilayers. Nanoscale Research Letters 2014, 9:427

Actuation concept for miniaturized tactile systems

Rui Zhu, Ulrich Mescheder, Frederico Lima
Department of Mechanical & Medical Engineering, Institute of Applied Research (IAF),
Hochschule Furtwangen University
Furtwangen, Germany
Rui Zhu: zhu@hs-furtwangen.de

***Abstract*—**

This paper describes an ongoing research on miniaturized, MEMS (micro electro mechanical system) based devices to realize tactile displays. They allow information transfer via mechanical stimulation of receptors in the human skin. An overview on tactile actuators is given with specific attention to miniaturization potential and high frequency drive. Then, a new concept of actuation using electro active polymers is presented. Fabrication aspects and preliminary results to realize tactile arrays are shown. Furthermore, integration aspects are discussed.

Keyword: Tactile actuator, Braille, Tactile display, Miniaturization

I. INTRODUCTION

In the last decade, tactile devices have been developed to provide multidimensional informations to humans simultaneously to vision, hearing and olfaction. Tactile displays are devices that can provide human information through the tactile sense in the form of Braille letter, warning or hint with different structure combinations, even in two dimensional image using the same dots as the ones used for Braille texts. Thus, not only blind people can use it for reading, but it also can be applied to car navigation systems to enable drivers to receive information without taking their eyes off road. It is also expected to be useful in keypad onside cell phones, PDAs (personal digital assistants), computer interfaces and virtual reality applications. Further on, tactile displays can be used in vision and hearing limited environment. Therefore, a proper tactile actuator is a promising human interface [1].

So far, commercially available tactile devices are mainly based on traditional mechanical processing. Therefore, these products have large dimensions, are heavy and expansive. Researchers are paying attention in reducing weight and costs due to the rapid development of MEMS technology. It offers: batch fabrication, lowering the cost per actuator on a display; miniaturization, reducing the weight of the display and increasing its resolution; and integration, minimizing the assembly needed to fabricate actuators and electronics on the same substrate [2].

However, since the actuator is miniaturized, due to the scaling law, the deformation which is generated by the tactile actuator is decreased as the dimension of the actuator is decreased. Fortunately, the detectable tactile receptor threshold on fingers are not only depending on the stroke length and contact force, but it is also affected by the contact velocity of fingers moving across the tactile surface, time duration and the object vibration frequency. Especially the vibration frequency increases miniaturization capability, for example when the objects vibrating at 200 Hz, humans are able to perceive a 1.0 μm step on the surface, thus the actuator structure can be further reduced by adding vibration, which results in a new concept: dynamic tactile actuator. Meanwhile, the size and the distance between the bumps on tactile devices also depend on the sensitivity of human skin, especially the fingertip when a tactile device is designed for finger touching. Therefore, the problem to reduce the size and weight of tactile actuators is how to reduce thickness and achieve the most sensitive vibration frequency: 200 Hz [3].

The needs on actuation for tactile displays can be derived from Table I. It summarizes the sensitivity of different cells in the human fingertip to mechanical stimulation. However, for reliable information transfer, actuators must provide amplitudes well above the listed threshold values.

In this paper, the actuator mechanism and actuation principles of previous tactile devices are compared. Then, a new concept made out of a dielectric electro active polymer (EAP) and a hydraulic displacement amplifier is presented.

TABLE I. AFFERENT SYSTEM AND THEIR PROPERTIES IN FINGERTIP [4]

Afferent type	**SA1**	**RA**	**PC**	**SA2**
Receptor	Merkel	Meissner	Pacinian	Ruffini
Effective stimulus	Edges, points, corners, curvature	Skin motion	vibration	Skin stretch
Response to sustanied indentation	Sustained with slow adaption	None	None	Sustained with slow adaption
Frequency range	0-100 Hz	1-300 Hz	5-1000Hz	0-? Hz
Peak sensitivity	5 Hz	50 Hz	200 Hz	0.5 Hz
Threshold for rapid indentation or vibration	8 μm	2 μm	0.01 μm	40 μm
Threshold	30 μm	6 μm	0.08 μm	300 μm
Receptive field area (measured with rapid 0.5 mm indentation)	9 mm^2	22 mm^2	Entire finger or hand	60 mm^2
Spatial acuity	0.5 mm	3 mm	10+ mm	7+ mm

II. TACTILE ACTUATOR MECHANISM AND ACTUATION PRINCIPLES

Many researches exerted their utmost effort to develop and design various actuators with different mechanisms according to different actuation principles. The actuator mechanisms can be categorized to traditional mechanical and MEMS based actuators. Important designs are listed in Table II and the mechanism sketches are shown in Fig 1.

A. Direct tactor probe driving mechanism

An actuation component for this mechanism lies below a tactile tactor, which is driven up directly by an actuation component as shown in Fig 1 (a). In this case, the stroke of tactile tactor is the same as the deformation of actuation component in the direction of stroke (piston like movement). Thus, an actuation principle that can offer large deformation is required, but it also results in a large bulk of actuation component.

Y. Haga et al. developed the actuator using shape memory alloy (SMA) coil as actuation component. The response time reported is under 1 s at a driving current of 300 mA. And the height of the actuator is not less than 2 cm [5]. A SMA wire perpendicular to a tactor was developed by F. Zhao et al. The thickness of the actuator is reduced since the actuation component is miniaturized. Its vibration frequency could reach up to 50 Hz as well [6]. Meanwhile, other researchers employed piezo plates as actuation components. Since a large deformation is demanded and the strain of piezo material is small, large piezo plates have to be used to reach the required deformation. As a result, it is impossible to integrate all the piezo plates in the same layer and compels the designer to place them layer by layer. Nevertheless, the thickness of the actuator could be reduced to 1 cm, yet the vibration frequency is only 10 Hz [7]. Electro active polymers (EAP) were also employed as actuation principle, but this principle brought no considerable effort to reduce the size of actuator due to the large size of column EAP. The thickness of this type actuator is similar as the SMA coil actuator, and the vibration frequency is less than 2 Hz. [8, 9]

TABLE II. COMPARISON BETWEEN ACTUATOR MECHANICS

Category	Mechanics	Feature
Traditional mechanical manufacturable	Direct tactor probe driven	Slow response Large volume Complex structure Non-miniaturizable
	Cantilever tactor probe driven	Amplification mechanism included Fast response Large volume Complex structure Non-miniaturizable
MEMS technology manufacturable	Storage chamber	Small volume Very low response speed Miniaturizable
	Amplifier chamber	Volume and response time depend on the chosen of actuation principle Miniaturizable

B. Cantilever tactor probe driven mechanism

In order to miniaturize the actuation component or to amplify the deformation created by actuation component, a cantilever was introduced as amplification component to the tactile actuator, it is also the merit of this mechanism (Fig 1 (b)). The cantilever can be a lever with fulcrum or a bimorph bar.

F. Yeh et al. designed a cantilever type actuator with electromagnetic actuation component. The height of this actuator is very large (53 mm) [10]. Piezoelectric bimorph bar as the cantilever for deflection mechanism is also used in commercial product from METEC AG and the height of it is reduced to 12.7 mm [11]. The most impressive design is from T. Sakurai's group. They designed a flexible tactile display by the cantilever structure. The height of it can be estimated to be few millimeters and the response frequency is 2 Hz [12, 13]. Despite that, there is a critical problem of cantilever type actuator. The cantilever occupies a large area around the braille cell which force the gaps exist between Braille cells. Although it is beneficial to distinguish each individual Braille letter, it obstructs to design tactile display with same distance between all Braille dots. This makes this approach complicated to use for array like displays.

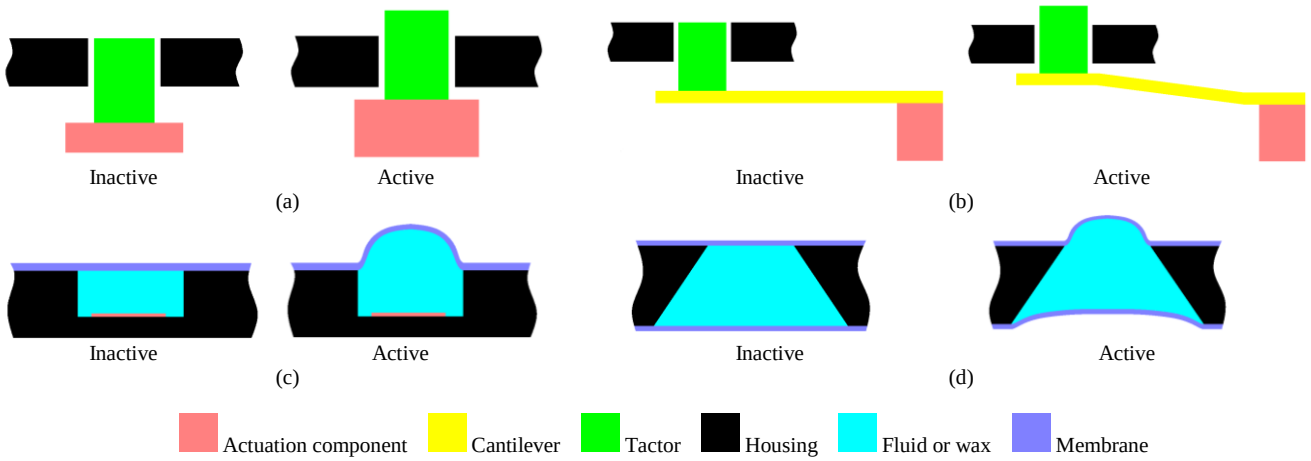


Fig. 1 Sketches illustrating actuator mechanisms
(a) Directly tactor probe driven (b) Cantilever tactor probe driven (c) Storage chamber (d) Amplifier chamber

C. *Storage chamber mechanism*

The cavity works as a space to store liquid or wax which could expand and shrink as the temperature rise and drop respectively, or by using other stimulating sources (Fig. 1 (c)).

S. R. Green et al. and J. S. Lee et al. used paraffin wax as the stored material inside each individual chamber. The height of the actuator can be reduced to few milimeters, the actuation circle time is about 1 min [14, 15]. M. Shikida et al. developed the actuator with fluorinert FC-72 (3M Chemicals) filled inside chamber. As the boiling point (56 ℃) of liquid is reached, the vapor is generated and activates the actuator [16]. S. Ukai et al. developed bubble driven actuator with the same principle, and the response frequency reaches 1 Hz [28]. Besides, UV light induces the cross link in liquid crystal polymer carbon nanotube which is also developed for this approach [17]. Due to the large size of the UV light actuation component, the volume of actuator cannot be reduced efficiently. Although some of the actuators are miniaturized, the response speed is limited due to transformation time of the stored material inside chamber.

Furthermore, a chamber connected to a valve which works as a switch for pressurizing was designed. Since a valve for each braille dot and a pressure source is required here, the volume of this actuator is extremely large [18]. To reduce the size, micro valves were developed too, but the leakage increased with their flow rate. Moreover, the micro valve can be damaged by large pressure. [19]

D. *Amplifier chamber mechanism*

This type of cavity works as an amplification component. It consists of a chamber with liquid encapsulated inside by two hyperelastic membranes, one from each side. Since the two membrane areas are different from each other, the deformation of the large membrane can be amplified and a larger deflection is noticed in the small membrane (Fig. 1 (d)).

Y. Matsumoto et al. has developed this principle, but as they used a piezo stack for actuation, the volume and weight of Braille cell is quite large, the vibration frequency can reach at least 200 Hz [20]. Meanwhile, M. Sadeghi et al. use the same principle with electrostatic force as actuation principle. Since the actuation component is of membrane type, the thickness of the tactile actuator is reduced to less than 1 mm. But the deflection of the bump is only 30 μm [21]. Afterwards, H. Seok et al. also designed an actuator which uses this principle and took an EAP as the driven component. The response frequency reaches up to 10 Hz [22, 23].

E. *Other strucutures*

A few other structures which cannot be classified based on the previous categories are presented here.

K. Matsuura et al. and R. Bansevicius et al. developed a tactile actuator using an electro rheological gel, because the stiffness of electro rheological gel can be controlled by electricity [2, 24]. Electromagnetic force is also taken as the driving principle to design a tactile actuator by J. Streque et al: A permanent magnet is placed above an electromagnet coil, and a vibration frequency of 350 Hz can be reached [25]. M. Matysek et al. designed a tactile display using out of plane actuation principle of multilayer dielectric EAP. With 50 actuation layers, a 300 Hz vibration frequency can be reached [26].

III. NEW CONCEPT OF EAP BASED TACTILE ACTUATOR

Owing to above discussion, one can find that the traditional mechanical processed mechanisms are not suitable for portable tactile devices because they are too large. Even if the amplification mechanism is employed, it is a challenge to reduce the thickness of a tactile device. Additionally, up to now there is no good approach to design and fully compact tactile display using cantilever mechanism.

Based on MEMS technology, the tactile actuator can be miniaturized efficiently when it is driven by some actuation principles such as thermal expansion, electro static forces, electro active polymer and electromagnetic forces. In spite of that, thermal expansion is the only one with large deformations from the ones mentioned above. Even though, due to the low response speed and temperature restrictions, it is not a good approach for tactile actuator. Therefore, in order to get large deformation and vibration frequency, a micro displacement amplification mechanism and reliable actuation principles are required.

As a result, an actuator structure with hydraulic amplification mechanism and multilayer dielectric EAP actuation component is proposed by the authors as shown in Fig. 2. The whole actuator manufacturing is based on silicon and MEMS technology. It consists of hydraulic amplifier, slider, dielectric EAP and substrate with orifice.

Amplifier concept: the cavity can be structured with different processes: dry etching by DRIE, RIE or wet etching by KOH among others. The amplification ratio can be varied by changing the area ratio between the two membranes. The fabrication process with KOH etching is described in the following section. The oil and membrane materials are Fomblin Y and PDMS (Sylgard 184), respectively.

Since the amplifier uses the displacement from the actuation component more efficient when a flat displacement (Fig. 3a) is applied compared to the rather curved displacement as shown in Fig. 3b. A slider is placed below the hydraulic amplifier to create flat deformation and then apply the deformation on amplifier. The simulation is done in COMSOL

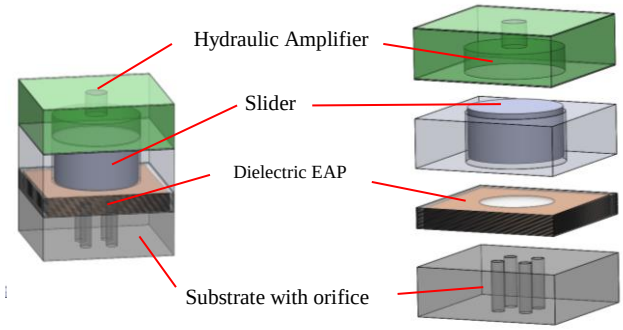
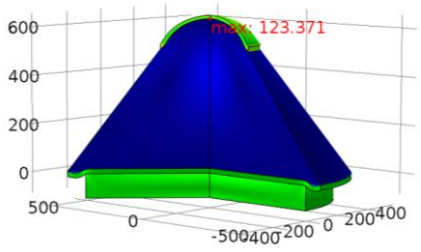
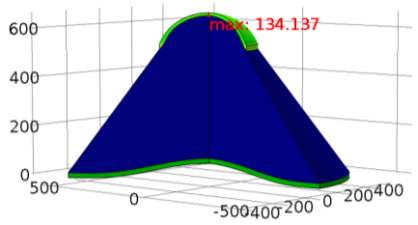


Fig. 2. Schematic of proposed tactile actuator



(a) Deformation of lower membrane is 11.4 μm

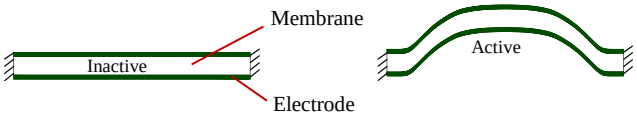


(b) Maximum deformation of lower membrane is 52.8 μm

Fig. 3. Hydraulic amplifier simulation in COMSOL with flat (a) and curved (b) deformation

multiphysics, the diameter of both lower and upper membranes are 400 μm and 1144 μm, respectively. The thickness of the membranes and the applied pressure from below are 20 μm and 10 KPa. Money-Rivlin model is used to consider the PDMS membrane properties in the simulation, and the parameters are material constant 1 (0 MPa), material constant 2 (0.1342 MPa), bulk modulus (1.214 GPa) [30]. Thus, with the same pressure, maximal deflection of the lower membrane 52.8 μm is amplified to 134 μm for curved displacement whereas for a flat displacement 11.4 μm is amplified to 123.4 μm showing the pronounced advantage of a flat deformation in comparison to a curved one. The size of the slider is slightly smaller than the lower membrane of the amplifier. At the bottom of actuator, there is a substrate which is used to force the dielectric EAP to deflect into the direction of the slider. Orifices were designed on the substrate to reduce air damping under deflection. Both structures can be realized by laser cutting.

The most important component is the actuation part, PDMS based dielectric EAP. The driving force of dielectric EAP is an electrostatically, increased by the dielectric constant of the material (PDMS) compared to actuators with air gap between the electrodes. It results in larger electrostatic forces and the soft polymer acts like support



(a) Electrodes on the deformation membrane

(b) Electrodes out of the deformation membrane

Fig. 4. Cross section schematic of dielectric EAP actuation principle

material for electrodes. It is beneficial to fabricate multilayer stack structure [27, 29].

As show in Fig. 4 (a), in previous research, the out-of-plane actuation principle is realized by coating the electrode on both sides of a clamped membrane. When a voltage is applied, the material between the electrodes is compressed which causes area expanding of the membrane and deflection. Assuming the use of a metal electrode, a membrane deformation in the range of few percentage points of the membrane diameter, its breakage would be caused. Graphite electrodes can be employed to prevent breakage, and also still increase the membrane's stiffness. Afterwards, more flexible net metal electrode and ion implanted electrodes have been researched to solve the problem, and yet the process is complex or expansive [31, 32]. The new actuation principle is shown in Fig. 4 (b), the electrodes are defined only in the non- deformed areas. As the voltage is applied, the material between electrodes is squeezed to the center and results in membrane deflection without stretching the electrodes. A simulation of this approach for single membrane is done in COMSOL Multiphysics as show in Fig. 5. The membrane material is PDMS, the deformable membrane radius is 750 μm and the membrane thickness is 20 μm, the ring width of electrode is 750 μm and the maximum deformation of the membrane is 30 μm under 1200 V.

Another merit of this type actuator is that the deformation depends on the material amount pressed out to the center at the boundary between electrode and free deformation region. In other words, the size of electrodes has only very little influence on the deformation of membrane. It is proofed by simulation as shown in Fig. 6, where the ring width of electrode is reduced to

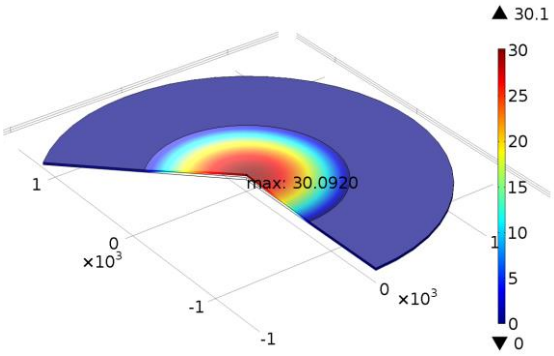


Fig. 5 Dielectric simulation

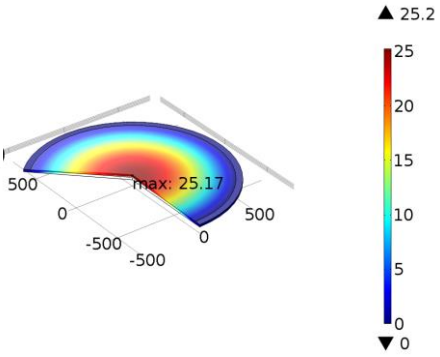


Fig. 6 Dielectric simulation

50 μm, the maximum deformation is 25 μm. It means, with 95.4% decreasing of electrodes area, the maximum deformation is reduced only by 16.7%. This behavior can be used to miniaturize the needed area of the actuator. Furthermore, a multilayer approach will increase the deformation of dielectric EAP as well.

IV. Amplifier Fabricaiton

The process flow is shown in Fig. 7. A 2.5 μm thickness silicon oxide layer is grown on a standard (100) silicon wafer in oven under 1050 ℃ for 20 h. Afterwards, squared windows with edges along <111> direction of silicon wafer are structured in silicon oxide layer by buffered HF (10:1) etching. The next step is to etch the cavities into the wafer by KOH etching. Benefit from the etching stop plane (111), a perfect oblique side wall with constant angle of 54.7° can be etched. The area ratio between the upper and lower membrane could be varied by changing the side length of opening structure. After the cavity is etched through, the left silicon oxide masking layer is removed by buffered HF etching. In the next step, a PDMS membrane is bonded on the large opening side of the chip. For proper bonding, the surface of the PDMS membrane is activated by O_2 plasma (100 W, 400 sccm O_2) for 30 s. Then, a thin film of wet PDMS is transferred on the other side of silicon chip and the cavity is filled with Fomblin Y afterwards. The last step is to close the cavity by a PDMS film with thin wet PDMS on the surface as glue, and place the encapsulated chip for 5 h inside an oven at 90 ℃ with 3 Kg weight on it.

Fig. 8 shows two sides of a fabricated amplifier. One actuated array of amplifiers is shown Fig. 9. In which, the side length of amplifiers are 750 μm.

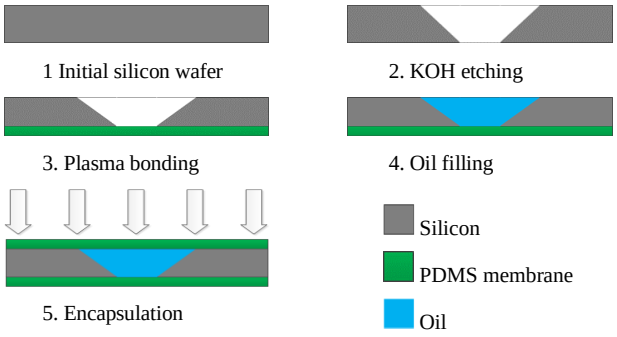


Fig. 7. Process flow of chip encapsulation

(a) (b)

Fig. 8. Fabricated chip, chip size is 25 mm ×25 mm (a) back side of amplifier, (b) front side of amplifier

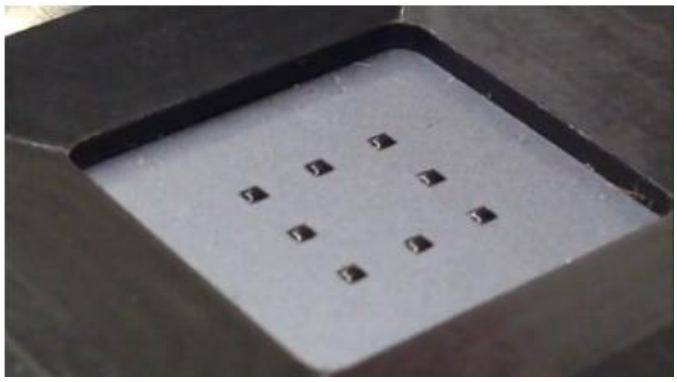


Fig. 9. Macroscopical actuated array of amplifiers

V. System Integration

In the previous sections, the actuation principle and the basic design of a single tactile pixel have been presented. In this section, the system integration is discussed. For a tactile display consisting of an array of tactile actuators each actuator has to be controlled individually. Thus, a tactile display can be used to transfer informations from the environment to humans. It can be used as applications such as Braille letters, 2D (or even 3D) pictures and so. It works similar as monitor for tactile sensing.

The complete system consists basically of three main units: central processing, voltage amplification and tactile display. The data to be displayed on the tactile system is received through a serial peripheral interface bus of a hardware single board computer [33]. The data consists of a binary $m \times n$ matrix and is treated in the central processing unit (CPU). The CPU sends the data to the amplification unit using a parallel bus. Each tactile actuator is driven by a single bus, the signal coming from the CPU is amplified and can physically deform a membrane. Therefore, using a tactile display is possible to reproduce the desired matrix. The system integration can be seen as a block diagram in Fig. 10.

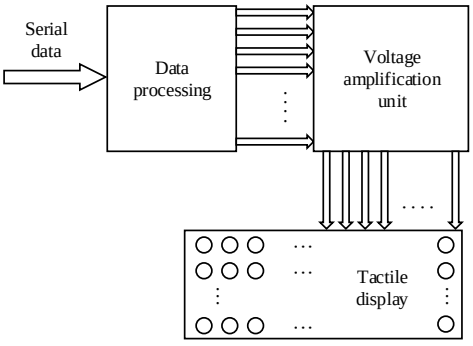


Fig. 10. System integration block diagram

References

[1] Jan B.F. van Erp, "Tactile displays for navigation and orientation:perception and behabiour," Mostert & Van Onderen, Leiden, The Netherlands, 2007

[2] K. Matsuura, T. Yakoh, T. Aoyama, H. Anzai, K. Sakurai, K. Isobe, “Smooth tactile display in mouse using electro-rheological gel,” Industrial Electronics, Proceedings of IEEE International Symposium, vol. 2, pp. 424-429, 2002

[3] L. M. Brown, “Tactons: structured vibrotactile for non-visual information display,” University of Glasgow, 2007

[4] S. Yantis, H. Pashler, “Steven’s handbook of experimental psychology, ” 3 Edition, John Wiley & Sons, Inc., 2002, Volume 1, chapter 13,

[5] Y. Haga, W. Makishi, K. Iwami, and K. Totsu, “Dynamic braille display using SMA coil actuator and magnetic latch,” Sensors and Actuators A, 2005, pp. 316–322

[6] F. Zhao, K. Fukuyama and H. Sawada, “Compact Braille display using SMA wire array,” The 18th IEEE International Symposium on Robot and Human Interactive Communication Toyama, Japan, pp. 28-33, September 2009

[7] H. Hernández, E. Preza, and R. Velázquez, “Characterization of a Piezoelectric Ultrasonic Linear Motor for Braille Displays,” Electronics, Robotics and Automotive Mechanics Conference, pp. 402-407, 2009

[8] K. Ren, S. Liu, M. Lin, Y. Wang, and Q.M. Zhang, “A compact electroactive polymer actuator suitable for refreshable Braille display,” Sensors and Actuators A, 2008, pp.335-342

[9] P. Chakraborti, H.A. Karahan Toprakci, P. Yang, N. Di Spigna, P. Franzon, and T. Ghosha, “A compact dielectric elastomer tubular actuator for refreshable Braille displays,” Sensors and Actuators A, 2012, pp.151–157

[10] F. Yeh, H. Tsay, and S. Liang, “Applied CAD and ANFIS to the Chinese Braille display optimization,” Displays, 2003, pp. 213-222

[11] Http://web.metec-ag.de/braillemodul%20p20.html, 2014, Sept. 16th.

[12] M. Takamiya, T. Sekitani, Y. Kato, H. Kawaguchi, T. Someya, T. Sakurai, “An organic FET SRAM with back gate to increase static noise margin and its application to braille sheet display,” IEEE J. of Solid-State Circuits, vol. 42, pp. 93-100, January 2007

[13] Y. Kato, S. Iba, T. Sekitani, Y. Noguchi, K. Hizu, X. Wang, K. Takenoshita, Y. Takamatsu, S. Nakano, K. Fukuda, K. Nakamura, T. Yamaue, M. Doi, K. Asaka, H. Kawaguchi, M. Takamiya, T. Sakurai, and T. Someya, “A flexible, lightweight braille sheet display with plastic actuators driven by an organic field-effect transistor active matrix,”, IEEE International Conference on Services Computing, July 2005

[14] S. R. Green, B. J. Gregoryt, and N. K. Gupta, “Dynamic braille display utilizing phase-change microactuators,” IEEE Sensors 2006, EXCO, Korea, pp. 22-25, October 2006

[15] J. S. Lee, S. Lucyszyn, “A micromachined refreshable braille cell,” J. of Microelectromechanical Systems, vol. 14, no. 4, 2005, pp. 673-682

[16] M. Shikida, T. Imamura, S. Ukai, T. Miyaji, and K. Sato, “Fabrication of a bubble-driven arrayed actuator for a tactile display,” J. of Micromecanics and Microengneering, 2008, pp.1-9

[17] C.J. Camargo, N. Torras, H. Campanella, J.E. Comrie, E.M. Campo, K. Zinoviev, E.M. Terentjev, and J.Esteve, “Light-actuatred CNT-doped elastomer blisters towards Braille dots,” IEEE Transducers, Beijing, China, pp. 1594-1597, June 2011

[18] X. Wu, S.-H. Kim, H. Zhu, C.-H. Ji, and M. G. Allen, “A refreshable braille cell based on pneumatic microbubble actuators,” J. of Microelectromechanical Systems, vol. 21, no. 4, 2012, pp. 908-916 L. Yobas, D. M. Durand, G. G. Skebe, and F. J. Lisy, “A novel integrable microvalve for refreshable braille display system,” J. of Microelectromechanical Systems, vol. 12, no.3, 2003, pp. 252-263,

[19] Y. Matsumoto, X. Arouette, T. Ninomiya, Y. Okayama and N. Miki, “Vibrational Braille code displaz with MEMS-based hydraulic displacment amplification mechanism,” Micro Electro Mechanical Systems (MEMS), IEEE 23rd International Conference, pp.19-22, January 2010

[20] M. Sadeghi, H. Kim, and K. Najafi, “Electrostatically driven micro-hydraulic actuator arrays,” Micro Electro Mechanical Systems (MEMS), IEEE 23rd International Conference, pp. 15-18, January 2010

[21] H. S. Lee, H. Y. Kwon, D. G. Kim, U. K. Kim, N. N. Linh, N. Canh Toan, H. Moon, J. C. Koo, J. Nam, and H. R. Choi, “SMD pluggable tactile display driven by doft actuator,” IEEE International Conference on Robotics and Automation, USA, pp. 2731-2736, May 2012

[22] H. S. Lee, H. Phung, D.-H. Lee, U. K. Kim, C. T. Nguyena, H. Moona, J. C. Koo, J. Namb, and H. R. Choi, “Design analysis and fabrication of arrayed tactile display based on dielectric elastomer actuator,” Sensors and Actuators A, 2014, pp. 191– 198

[23] R. Bansevicius, J.A. Virbalis, “Two-dimensional Braille readers based on electrorheological fluid valves controlled by electric field,” Mechatronics, 2007, pp. 570–577

[24] J. Streque, A. Talbi, P. Pernod, and V. Preobrazhensky, “New magnetic microactuator design based on PDMS elastomer and MEMS technologies for tactile display,” IEEE Transaction on Haptics, vol. 3, no. 2, pp. 88-97, 2010,

[25] M. Matysek, P. Lotz, T. Winterstein, H. and F. Schlaak, Dielectric elastomer actuators for tactile displays, 3rd Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, USA, pp. 290-295, March 2009

[26] R. Kornbluh, R. Pelrine, Q. Pei, S. Oh, and J. Joseph, “Ultrahigh streain repones of field-actuated elastomeric polymers,” In Smart Structures and Materials: Electroactive polymer Actuators and Devices, 2000, pp. 51-64

[27] S. Ukai, T. Imamura, M. Shikida, K. Sato, “Bubble driven arrayed actuator device for a tactile display,” Solid-State Sensors, Actuators and Microsystems Conference, pp. 2171-2174, June 2007

[28] Y. Bar-Cohen, “Electroactive polymer (EAP) actuators as artificial muscles: reality, ptorential, and challenges,” 2n edition, SPIE PRESS

[29] T. K. Kim, J. K. Kim, O. C. Jeong, “Measurement of Nonlinear Mechanical Properties of PDMS Elastomer,” Microelectronic Engineering, 2011, pp.1982–1985

[30] S. Rosset, M. Niklaus, P. Dubois, and H. R. Shea, “Mechanical characterization of a dielectric elastomer microactuator with ion-implanted electrodes,” Sensors and Actuators A, 2008, pp. 185-193

[31] S. P. Lacour, S. Wagner, Z. Huang, and Z. Suo, “Stretchable gold conductors on elastomeric substrates,” Applied Physocs Letters, vol. 82,no. 15, 2004, pp. 2404-2406

[32] G. Coley, “BeagleBone Black System Reference Manual”, Rev. C.1, USA, 2014.

Checkpoint/Restore in User-Space with Open MPI

Adrian Reber and Peter Väterlein
Hochschule Esslingen
Fakultät Informationstechnik
Flandernstraße 101
73732 Esslingen

***Abstract*—With the every increasing size of High Performance Computing (HPC) clusters new approaches to system management are required. A first step towards new system management approaches in an HPC environment is to employ process migration to dynamically move running processes to the most suited location. This paper describes the required steps towards process migration based on CRIU and Open MPI**

I. Introduction

Virtualization is one of the primary platforms providing services in a data center. With techniques such as virtual machine migration system management tasks like the installation of new software or hardware which in turn requires a downtime of the physical machine can easily be performed without interrupting the running applications. By migrating all running virtual machines off the physical machine it can be made available for system management tasks like hardware maintenance or software updates. Examples for easy-to-use off-the-shelf solutions which support virtual machine migration are the hypervisor implementations from VMware [1] and KVM [2].

Especially in an HPC environment with its every increasing number of nodes such system management approaches are not yet widely in use. This is related to the fact that in an HPC environment the CPU is usually the most important resource. Although the virtual machine performance penalty is minimal there is ongoing research to optimize the usage of the resources and a common approach in an HPC environment is to use para-virtualization or even container based virtualization [3] to reduce the overhead of the virtualization.

The approaches to reduce the virtualization overhead by using simpler virtualization techniques like para-virtualization and container based virtualization are a strong indicator that, no matter how small the overhead is, every CPU cycle is important and should not be wasted if possible.

Another problem with virtualization is access to network or I/O devices. For calculations which are running on not just a single node, but on multiple nodes, low latency communications are important to reduce the time the CPUs have to idle waiting on results from the involved nodes. In virtualized scenarios each access to the network or to an I/O device has also to pass through the hypervisor which adds latencies. Or in the case where the virtual machine is directly accessing the communications hardware via pass-through (or SR-IOV) it is not clear how virtual machine migration can work, especially if using out-of-band communication hardware like InfiniBand.

The small but existing virtualization overhead makes virtualization still not very common in an HPC environment and continuing the trend to reduce the virtualization overhead this work proposes to migrate single processes or process groups[1] and thus removing the need for any overhead which wastes CPU cycles.

To successfully use process migration in an HPC environment it has to support existing means of parallelization. With MPI being one of primary approaches to parallelize a computational task over multiple nodes and cores any kind of process migration has to support MPI parallelized applications. If the MPI environment can handle process migration, especially in combination with out-of-band communication like InfiniBand, it becomes easier to migrate processes as the knowledge of the underlying communication technology is no longer necessary to the instance triggering the migration.

After describing the motivation this work is based on in Section I. Section II. introduces different process migration approaches and why this work is based on checkpointing and restarting. Existing checkpointing and restarting implementations are discussed and CRIU has been selected as the currently best implementation. As this work focuses on an HPC environment existing MPI fault tolerance frameworks as a basis for process migration are introduced. Section III. discusses different migration approaches and possible optimizations and the actual steps to re-enable checkpointing and restarting in Open MPI in combination with CRIU are described in Section IV. The last Section (V.) summarizes this work and gives an outlook about what still needs to be done to use process migration for previously mentioned system management tasks.

II. Related Work

Process migration can either be seen as a special case of regular scheduling like it is performed by every preemptive multitasking operating system with the difference that the process can be scheduled to a different physical (or virtual) system instead of scheduling the process on a local CPU. On the other hand process migration can be seen as specialized form of checkpointing and restarting where the checkpointing is not used to write a process image on disk but instead is directly transferred to the memory of the destination system.

This work bases its process migration on checkpointing and restarting as there are already different existing implementations for checkpointing and restarting and it also does not require changes to a central part of the operating system like the process scheduler, like it would be required if basing process migration on the mechanisms of preemptive multitasking.

[1] a process with all its child processes

As this work is mainly targeting an HPC environment it focuses on Linux as Linux is dominant in HPC and used on over 90% of the worlds fastest systems[2]. In addition to its wide adoption in supercomputing the openness of Linux makes it a perfect basis for this work.

There are various implementations of checkpoint/restart for Linux based systems. There are user space variants like Distributed MultiThreaded Checkpointing (DMTCP) [4], which requires no changes to the operating system. It aims to be as transparent as possible but still needs a special environment for the process to be checkpointed. It requires certain libraries to be pre-loaded to intercept different system calls. This design reduces the usage of DMTCP as it is not possible to checkpoint a process which has not been started in this environment.

In contrast to pure user space implementations, there are also checkpoint/restart variants which require changes to the operating system. There are a few implementations for Linux which require a different number of changes to the Linux kernel.

Berkeley Lab Checkpoint Restart (BLCR) [5] provides an almost transparent Checkpoint/Restart implementation for Linux by extending Linux with a kernel module which contains most of the functionality. BLCR still requires, in certain scenarios, libraries to be pre-loaded. It is not intended to be a fully transparent checkpoint/restart implementation. To minimize the changes required to the Linux kernel, BLCR's code is located in a kernel module. The necessary functionality is located in this module and, depending on the Linux version, most systems can easily load the BLCR kernel module. Limiting the code changes to the BLCR module makes it easy to integrate it into a Linux system but, on the other hand, limits the functionality. Another problem with BLCR is that it is not part of the mainline Linux kernel, which requires additional effort every time the kernel needs to be updated.

An attempt to fix the limitations of implementations like BLCR was started in 2010 by implementing transparent application checkpoint/restart as part of the mainline Linux kernel [6]. The goal was to offer checkpointing and restarting capabilities by developing the necessary functionality from the beginning, hand in hand with the Linux community. The development ended with the Linux kernel version 2.6.37 and was later ported to Linux kernel version 3.2 [7]. At this point it became clear that it had little chance of being accepted by the Linux kernel community, because it consisted of over 100 patches, which changed almost every subsystem of the Linux kernel and was therefore considered too invasive and too difficult to maintain.

The latest approach to operating system level checkpointing and restarting in Linux tries to avoid the problems previous implementations faced. Instead of trying to introduce massive changes at many places in the Linux kernel, Checkpoint/Restore In Userspace[3] (CRIU) tries to do as much as possible in the userspace. In addition, to be as transparent as possible, it uses existing Linux kernel interfaces to avoid the necessity of processes having to run in a specially prepared environment (e.g., pre-loading libraries). As the interfaces provided by the Linux kernel did not offer all the information required to checkpoint and restore a process, the CRIU developers extended the existing interfaces and introduced new ones. These minimal changes to the Linux kernel were accepted as they (in most cases) can also be used by use cases other than checkpoint/restore.

Since it provides a transparent checkpoint/restart implementation CRIU has been adopted in several Linux distributions. Starting with Fedora 19, for example, it is possible to use checkpoint/restart without software or configuration outside of Fedora[4].

However, in an HPC environment, more is required than only the ability to checkpoint and restore a process to provide the ability to migrate processes. Process migration can only be employed if the used Message Passing Interface (MPI) implementation supports one of the existing checkpoint/restart solutions. LAM/MPI [8] has been the subject of different research projects providing checkpoint/restart features in combination with BLCR [9] [10].

LAM/MPIs successor Open MPI [11] has also been adapted to provide fault tolerance mechanisms [12] [13] through multiple checkpoint and restart implementations like BLCR and DMTCP. The existing infrastructure in Open MPI and its open and community based development model make Open MPI the optimal starting point to support process migration in combination with CRIU.

Single System Image (SSI) implementations like openMosix[5] have not been taken into account, as most of today's HPC applications are more likely to be running in a MPI environment than in an SSI environment.

III. Design

A process is a container or instance of an application or program which is currently being executed. According to [14, 89] a process consist of an entry in the process table "with one entry per process". Each entry in the process table includes all the information about the process and the resources which have been allocated to it. In the context of migrating a process from one system to another the following parts of a process have to be taken in account: Process management, Memory management, File management. Concentrating on process migration in an HPC environment with MPI parallelized processes removes the need to look at network connections as all communication is abstracted by the MPI layer.

As mentioned before, process migration can either be seen as a special case of regular scheduling as it is found in every preemptive multitasking operating system. The operating system schedules processes which are ready to run on the available CPUs, by copying the necessary data on and off the CPUs. Process migration could be implemented by scheduling the processes on another node and thus migrating the process by transferring the process table entry to that other node. The process' memory could then be transferred by on-demand paging. Or, on the other hand, process migration can be seen as a special case of checkpoint/restart. Instead of storing the information about the process which is collected during

[2]http://top500.org - TOP500 Release September 2014

[3]http://criu.org/

[4]https://fedoraproject.org/wiki/Features/Checkpoint_Restore

[5]http://openmosix.sourceforge.net/

checkpointing on disk, all the data are directly transferred to the memory of the destination node. As checkpointing and restarting already has all the functionality required to capture all necessary information about a process, it is well suited as a basis for process migration. In either case (scheduling from one node to another or checkpointing directly to the memory of the destination node) the process has to be suspended, all the information about the process (process table entry and the used memory) have to be transmitted to the destination node and finally the process has to be resumed. The biggest difference is the way the memory is transferred. Either, as in the case of scheduling, it would be a form of on-demand paging, or as in the case of checkpointing, the complete memory would be transferred before the process is restarted.

Comparing the amount of data to be transferred the memory management related information requires by far the most time to be transferred and therefore three different methods to transfer the memory from one node to another have been taken into account for this work. The first method is the simplest (see Figure 1).

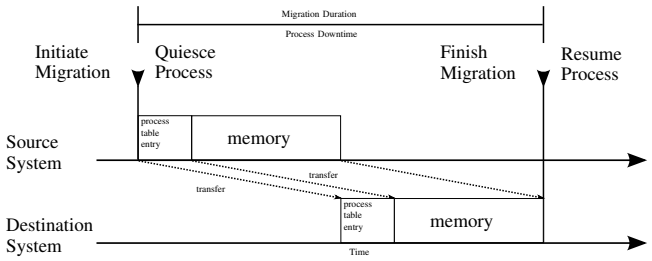


Fig. 1. Memory Transfer During Migration

The process is suspended during the whole time the process table entry and the memory is transfered. This leads to the longest downtime of the migrated process but it is also the simplest method, making memory transfer less complex. The second method of transferring memory is to pre-copy the memory of the process to be migrated, allowing the process to keep running during this time (see Figure 2).

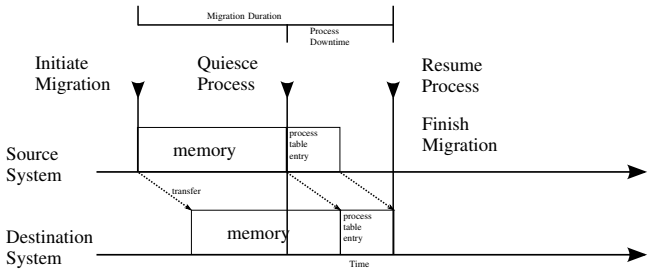


Fig. 2. Memory Transfer Before Migration

Pre-copying can be done multiple times and each time only the changes to the previous run have to be transferred. The final step is to suspend the process, transfer the process table entry and then resume the migrated process. The remaining method is to post-copy the memory which corresponds to on-demand paging (see Figure 3). During the process' downtime only essential information about the process are transferred (process table entry) to the destination node and the memory is copied when actually being accessed. Although this seems to be the method which is the most effective it is also the most complex method as it requires changes to the operating system's way of paging and scheduling.

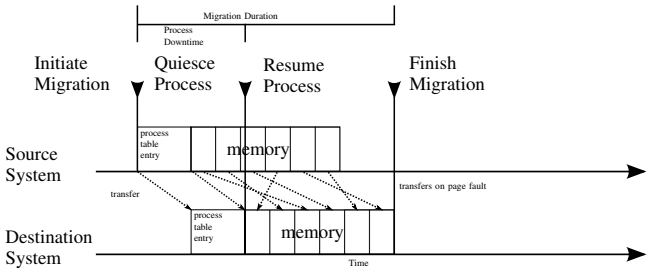


Fig. 3. Memory Transfer After Migration

In the scope of this work, the simplest way of transferring the memory will be used, with the option of using pre-copying at a later point in time.

Another simplification to reduce the impact of process migration on the operating system is to use indirect migration instead of direct migration. In the case of direct migration the memory of the to be migrated process would have been directly transferred from the source system to the destination system. This would require to include network transfer methods, authentication and encryption directly in the operating system's kernel. To avoid instabilities indirect migration using existing tools has been implemented. The drawback of multiple copies (kernel-space to user-space, user-space to kernel-space) of the process' data is reduced by the fact that the time required for the network transfer is several times higher.

Looking at process migration in an HPC environment it also requires support for process migration in a MPI implementation. Therefore the goal of this work is to support checkpointing and restarting in a MPI implementation as an intermediate step to process migration in an HPC environment. An important requirement for checkpointing and restarting is that it is as transparent as possible. With a transparent checkpoint/restart implementation, which is also accepted by the Linux community, it can be easily used on any systems and reduces the constraints placed on an HPC system. Being transparent is important in order to reduce the constraints on the application such as having to run in a special environment, which has to be set up prior checkpointing.

IV. Implementation

The checkpoint/restart implementation used in this work is CRIU. The reason for using CRIU is that it is the solution which provides the most transparent checkpoint/restart implementation and it is available in Linux distributions without additional effort. It is being actively developed and accepted by the Linux kernel community. All these factors currently make CRIU the best-suited checkpoint/restart implementation and process migration using CRIU has already been presented [7].

The MPI implementation used in this work is Open MPI. Open MPI supported checkpointing and restarting of processes independent of the checkpoint/restart implementation through

the Checkpoint and Restart Service (CRS) framework [13]. This work was started in 2007 [12] and looking at the history of the revision control system it seems to have ended in 2010.

Unfortunately, the remaining code and design of Open MPI have changed so much that the existing code was no longer functional. The changes to the code infrastructure in Open MPI were, in fact, so fundamental that it was no longer possible to compile Open MPI with the fault tolerance code paths enabled. The biggest change in Open MPI, which was not reflected in the fault tolerance code, was the discontinuation of the blocking communication. Many parts of the fault tolerance code used the blocking communication functions and they all had to be redesigned to use non-blocking communication.

The first step to re-enable the fault tolerance code in Open MPI was to allow the existing code to be compiled again, and to reflect the architectural changes of Open MPI (blocking vs. non-blocking communication, variable handling, ...).

In a next step a new module was added to the CRS, supporting checkpointing and restarting, using CRIU. With this new CRS module it is now possible to checkpoint and restart an Open MPI application using `orte-checkpoint`. The program `orte-checkpoint` is a stand-alone program which uses Open MPI internal interfaces to communicate with `mpirun` (which controls the programs running in the MPI runtime environment) to signal that the processes under `mpirun`'s control should be checkpointed (see Figure 4).

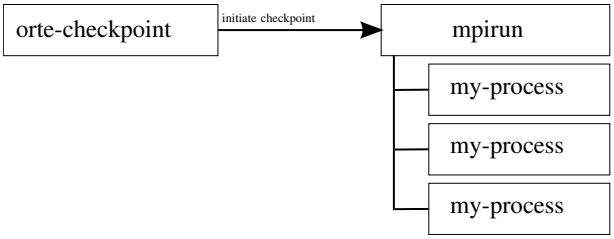


Fig. 4. Initiate a checkpoint in Open MPI

Depending on the configuration, the processes will continue to run after having been checkpointed or they will be terminated. The CRS framework tells CRIU at which place in the file system the image of the checkpointed processes should be stored and it also writes additional meta-data about the checkpointed processes, which will later be used to restart the application.

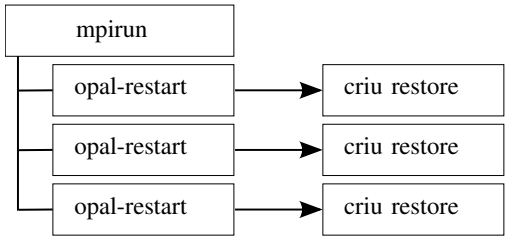


Fig. 5. Restart processes in Open MPI

Using `orte-restart` it is possible to restart a process which has previously been checkpointed using `orte-checkpoint`. `orte-restart` needs to know the location of the checkpoint meta-data and uses this information to restart the processes. It starts an instance of `mpirun` with the corresponding number of child processes. Instead of the original process (e.g., `my-process`) `mpirun` starts multiple `opal-restart` processes. These processes restart the actual processes using CRIU (see Figure 5). The expected result of the restart is a process tree resembling the one which was present during checkpointing (see Figure 4).

The main problem during restart is related to the interfaces provided by CRIU. The main interface provided by CRIU is the command-line interface. There are many different options but the main usage of CRIU is `criu dump -t <PID>` and `criu restore` to checkpoint or restart a process respectively. For inclusion into Open MPI a library-like interface would be preferred to the calling of an external command-line program like `criu`.

The possibility of linking against CRIU is the reason why it also provides a library which can be easily included into any software. The library `libcriu.so` provides all the necessary interfaces to checkpoint and restart processes. The library is only a thin Remote Procedure Call (RPC) wrapper which communicates with a CRIU process running as a daemon. Since the daemon is running with `root` privileges, any user can checkpoint and restart processes without any additional privileges required.

This approach, however, collides with the expectations Open MPI has on the utility restoring the checkpointed process. Processes running under Open MPI's control are child processes of `mpirun` (see Figure 4 right part) and after restoring the checkpointed processes, this is also the expected end-result.

Since CRIU's restore is also initiated by the daemonized `criu` through the RPC library, the restored processes will be a child processes of the daemonized `criu` service and not of `mpirun`. Re-parenting a process in Linux is not possible and therefore another solution on how to use CRIU is required. The easiest solution was to directly `exec() criu` instead of using the `libcriu.so` library. Thus the restored processes are child processes of `mpirun`. Restoring a process requires special privileges which can be achieved by executing `criu` with `root` privileges (`setuid`).

A similar problem exists with respect to the output of `mpirun`'s child processes. To collect all output of all attached child processes, Open MPI replaces the file descriptors for `stdout` and `stderr` with pipes. Thus all output can be redirected to Open MPI's Head Node Process (HNP). During checkpointing of the processes attached to `mpirun`, CRIU checkpoints the file descriptors for `stdout` and `stderr` just as the pipes as they are created by `mpirun`. Upon restart `mpirun` creates again the pipes to replace `stdout` and `stderr` but this time with different node numbers. For pipes like these, which are connected and created by something external to the restored process, the pipes have to be adjusted before CRIU finishes the restore. To solve this problem, a CRIU plugin has been created, which adjusts the pipes to the new environment.

V. Conclusion and Outlook

With CRIU being accepted in the official Linux kernel and thus providing checkpointing and restarting on most modern Linux systems out of the box, it is the best possible solution to provide transparent checkpointing and restarting. Being transparent as much as possible decreases the barrier for users and system administrators to actually use checkpointing and restarting which in turn will increase the acceptance for this technology. These factors (transparency and wide-spread acceptance) are making the combination of CRIU and Open MPI an important first step for process migration in an HPC environment. The next step to support process migration in an HPC environment is to adapt `orte-migrate` which shall be used to migrate a single MPI process from one system to another. Migrating parts of a parallel job in an HPC environment opens up the possibility to integrate process migration in management tools and resource schedulers to offer the flexibility of virtual machine migration in HPC environments without a hypervisor and the thereby connected overheads.

CRIU and Open MPI are well established in their respective communities. Thus the combination of those tools offers new perspectives in an HPC environment without requiring the installation of third-party software and/or limiting the operating system version and application environment to be used.

All code changes described in this work are part of the respective upstream projects and can easily be tried and used.

Acknowledgment

We would like to acknowledge the use of the computing resources provided in cooperation with bwGRiD [15] and bwHPC [16]. This work was partially funded by the MWK Baden-Württemberg.

References

[1] I. VMware, "VMware vSphere vMotion Architecture, Performance and Best Practices in VMware vSphere 5," http://www.vmware.com/files/pdf/vmotion-perf-vsphere5.pdf, Tech. Rep., 2011.

[2] KVM - Kernel-based Virtual Machine, "Migration - kvm," http://www.linux-kvm.org/page/Migration, 2012. [Online]. Available: http://www.linux-kvm.org/page/Migration

[3] M. G. Xavier, M. V. Neves, F. D. Rossi, T. C. Ferreto, T. Lange, and C. A. De Rose, "Performance evaluation of container-based virtualization for high performance computing environments," in *Parallel, Distributed and Network-Based Processing (PDP), 2013 21st Euromicro International Conference on*. IEEE, 2013, pp. 233–240.

[4] J. Ansel, K. Arya, and G. Cooperman, "DMTCP: Transparent checkpointing for cluster computations and the desktop," in *23rd IEEE International Parallel and Distributed Processing Symposium*, Rome, Italy, May 2009.

[5] J. Duell, "The design and implementation of berkeley labs linux checkpoint/restart," Tech. Rep., 2003.

[6] O. Ladaan and S. E. Hallyn, "Linux-cr: Transparent application checkpoint-restart in linux," in *The Linux Symposium 2010, Ottawa, July 2010*, 2010. [Online]. Available: http://systems.cs.columbia.edu/files/wpid-ols2010-linuxcr.pdf

[7] A. Reber and P. Väterlein, "Pos (isgc 2012) 031 live process migration for load balancing and/or fault tolerance," in *The International Symposium on Grids and Clouds (ISGC)*, vol. 2012, 2012.

[8] G. Burns, R. Daoud, and J. Vaigl, "LAM: An Open Cluster Environment for MPI," in *Proceedings of Supercomputing Symposium*, 1994, pp. 379–386. [Online]. Available: {http://www.lam-mpi.org/download/files/lam-papers.tar.gz}

[9] S. Sankaran, J. M. Squyres, B. Barrett, A. Lumsdaine, J. Duell, P. Hargrove, and E. Roman, "The LAM/MPI checkpoint/restart framework: System-initiated checkpointing," *International Journal of High Performance Computing Applications*, vol. 19, no. 4, pp. 479–493, Winter 2005.

[10] C. Wang, F. Mueller, C. Engelmann, and S. L. Scott, "A job pause service under LAM/MPI+BLCR for transparent fault tolerance," in *Proceedings of the* 21^{st} *IEEE International Parallel and Distributed Processing Symposium (IPDPS) 2007*. ACM Press, New York, NY, USA, Mar. 26-30, 2007, pp. 1–10. [Online]. Available: http://www.csm.ornl.gov/~engelman/publications/wang07job.pdf

[11] E. Gabriel, G. E. Fagg, G. Bosilca, T. Angskun, J. J. Dongarra, J. M. Squyres, V. Sahay, P. Kambadur, B. Barrett, A. Lumsdaine, R. H. Castain, D. J. Daniel, R. L. Graham, and T. S. Woodall, "Open MPI: Goals, concept, and design of a next generation MPI implementation," in *Proceedings, 11th European PVM/MPI Users' Group Meeting*, Budapest, Hungary, September 2004, pp. 97–104.

[12] J. Hursey, J. M. Squyres, T. I. Mattox, and A. Lumsdaine, "The design and implementation of checkpoint/restart process fault tolerance for Open MPI," in *Proceedings of the 21st IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE Computer Society, 03 2007.

[13] J. Hursey, J. M. Squyres, and A. Lumsdaine, "A checkpoint and restart service specification for open mpi," Indiana University, Bloomington, Indiana, USA, Tech. Rep. TR635, July 2006. [Online]. Available: http://www.cs.indiana.edu/cgi-bin/techreports/TRNNN.cgi?trnum=TR635

[14] A. S. Tanenbaum, *Modern operating systems*, 3rd ed. Upper Saddle River, NJ: Prentice-Hall, 2010.

[15] M. Dynowski, M. Janczyk, J. Schulz, D. von Suchodoletz, and S. Hermann, "Das bwGRiD - "High Performance Compute Cluster" als flexible, verteilte Wissenschaftsinfrastruktur." in *DFN-Forum Kommunikationstechnologien*, ser. LNI, vol. 203. GI, 2012, pp. 95–105, 1617-5468.

[16] H. Hartenstein, T. Walter, and P. Castellaz, "Aktuelle Umsetzungskonzepte der Universitäten des Landes Baden-Württemberg für Hochleistungsrechnen und datenintensive Dienste," *PIK-Praxis der Informationsverarbeitung und Kommunikation*, vol. 36, no. 2, pp. 99–108, 2013.

An Architecture for Cloud Accountability Audits

Thomas Rübsamen, Christoph Reich, Martin Knahl
Cloud Research Lab
Furtwangen University
Furtwangen, Germany
Email: {ruet, rch, knahl}@hs-furtwangen.de

Nathan Clarke
Centre for Security, Communications and Network Research
Plymouth University
Plymouth, United Kingdom
Email: N.Clarke@plymouth.ac.uk

***Abstract*—As more and more sensitive data is stored in the cloud, privacy and security become more important. Today's cloud services, their internal processes and details about how and by whom data is processed is opaque to the cloud user. To address the trust issues stemming from this loss of control, we propose a software agent-based system for auditing accountability policies, to increase transparency of cloud services.**

***Keywords*—Accountability, Audit, Cloud Computing, Digital Evidence**

I. Introduction

Cloud Computing is an increasingly popular paradigm for service delivery in today's Internet [1] and may lead to significant advantages such as reduced upfront investments [2], rapid provisioning and automatic scaling of resources [3]. However, the adoption of cloud computing is accompanied with several security and privacy problems. For instance, data breaches and data loss are amongst the major threats in cloud computing [4]. Therefore, two of the key issues are customer trust and compliance [2], [5]. Because of the loss of control, cloud customers have to trust cloud providers to handle their data appropriately and that sufficient data protection mechanisms are in place. Cloud Providers use terms of service (TOS), which are generally non-negotiable [5], and privacy agreements to describe how data will be handled in their services. However, beyond such documents, there is usually a lack of transparency regarding details about security processes and controls. Trust is also an issue when service providers use additional services provided by third-parties, because trust will not necessarily be transitive in such complex scenarios [2]. This lack of trust can be addressed by strengthening transparency and accountability [6], [7] on the cloud provider side.

The Audit Agent System proposed in this paper, strives to enable automated cloud accountability audits by addressing transparency and privacy protection issues associated with cloud computing. Accountability is regarded as a means to strengthen customer trust in cloud services. By auditing compliance with data policies, transparency and privacy of the cloud shall be improved. This includes the secure and privacy-aware collection of evidence supporting claims made in audit reports. In this paper, we describe an architecture for the Audit Agent System.

This paper is structured as follows: in Section II a brief overview about current research projects and industrial approaches is given. In Section III, we describe the proposed system architecture for the Audit Agent System. We close this paper with a conclusion in Section IV.

II. Related Work

There are several academic approaches to various aspects of cloud auditing. For instance, security auditing is a very important part of accountability auditing of a cloud provider, since it demonstrates that required security controls are put in place and are functioning correctly. There are some projects working on the architectural and interface level regarding the automation of security audits such as the Security Audit as a Service (SAaaS) project [8]. The Distributed Management Task Force (DMTF) is also working on cloud auditing with the Cloud Auditing Data Federation (CADF) working group. They are focusing mostly on developing standardized interfaces and data formats to enable cloud security auditing [9]. A similar project is the Cloud Security Alliance's (CSA) Cloud Trust Protocol (CTP), which defines an interface for enabling cloud users to "generate confidence that everything that is claimed to be happening in the cloud is indeed happening as described, ..., and nothing else" [10], which indicates an additional focus on providing additional transparency of cloud services. The latter two projects, however, do neither detail an actual architecture and how the interfaces shall be implemented nor do they describe explicitly focus on accountability.

A lot of current research is not focused on the overall automation of cloud accountability audits, but rather on aspects, that would be part of such an audit (i.e., may be implemented as part of the system described in this paper). Such approaches are for example concerned with the provenance of data in the cloud [11], proof of retrievability and provable data possession [12], virtual machine introspection [13] and replay as an advanced monitoring and forensic analysis technique.

When looking at cloud audits and the associated process of collecting evidence to assess policy compliance, it is important to look at industry practices regarding monitoring. Many such tools, such as the well-established Nagios [14] support agent-based data collection. New Relic [15], a Software as a Service (SaaS) software analytics solution enables the collection of data on various different scopes and devices. However, most of these tools are mainly concerned with performance monitoring and tracing, whereas our approach mainly considers the automation of security and accountability auditing. Security Information and Event Management (SIEM) systems are the main source of monitoring information in today's more complex IT infrastructures. They provide additional means of detecting security incidents by collecting information from various sources in the infrastructure. However, when it comes to auditing policies on the level data objects and regarding accountability requirements specific to individual customers,

there is still lacking functionality.

III. Architecture

In this Section, we describe the high-level architecture of the Audit Agent System and its components. We also describe the input to the Audit Agent System in the form of accountability policies and illustrate the data flow of evidence from its source to the processing components across the different architectural layers.

A. Audit Agent System Architecture Introduction

In Figure 1 the overall architecture of the Audit Agent System is depicted. In the following, we describe the tool's main actors, components and the general flow of information from the evidence-producing source to the audit report.

AAS Actors:
There is one actor using the Audit Agent System: the auditor. According to NIST, a cloud auditor is a "A party that can conduct independent assessment of cloud services, information system operations, performance and security of the cloud implementation." [16] Based on this, a cloud customer, cloud provider or any third-party can act as a cloud auditor. From this, requirements regarding depth and presentation of audit results can be derived. Also, since an auditor can be internal or external to an organization (i.e., a cloud service provider), data protection is an issue to consider, when potential confidential information is processed during an audit. These issues shall be addressed by the presentation and anonymization components described later in this section.

AAS Components:
The architecture of the Audit Agent System is based on using software agents to achieve flexibility, address requirements regarding the dynamics of cloud computing (e.g., rapid elasticity) and achieve the necessary extensibility required by the cloud, where evidence data may need to be gathered from highly diverse evidence sources. The proposed architecture comprises of four major functional components: *Audit Policy Module (APM)*, *Audit Agent Controller (AAC)*, *Evidence Processor and Presenter (EPP)* and *Evidence Store (ES)*. For a high-level overview of the system, refer to Figure 1.

Audit Policy Module: There are two types of input to the Audit Agent System:

1) Accountability policies, which define obligations that have to be fulfilled by the cloud provider, such as data access restrictions and usage policies, data retention requirements and general security requirements (e.g., use of encryption). The A4Cloud [17] research project develops a machine-readable policy language based on the Primelife Policy Language [18] called Accountability PPL [19], which will serve as input to the Audit Agent System.
2) Since the A-PPL does not address technical aspects, such as mapping policy requirements to specific tools to use for evidence collection and details of the processing of such evidence, additional manual input is required by the cloud auditor.

The Audit Policy Module (APM) uses both inputs to generate audit tasks. Audit tasks are managed by the Audit Agent Controller.

Audit Agent Controller: The Audit Agent Controller (AAC), can be regarded as the core component of the Audit Agent System. It is responsible for managing the life-cycle of evidence collection agents, controlling audit execution, storage of evidence records and managing data flow between the components. For instance, the Audit Agent Controller deploys, according to what's specified in the audit policy and audit task, evidence collection agents across the various architectural layers of a cloud infrastructure (i.e., in a virtual machine, on a virtualization host, in an application server). From there, data, such as logs, object storage information, block storage information and analysis application output is collected.

Evidence Processor and Presenter: The Evidence Processor and Presenter (EPP) component is responsible for evaluating policies based on the evidence gathered by the audit agents. This component is, similar to the Audit Agent Controller, logically formed by several agents; in this case Processing Agents are responsible for the evaluation of audit policies and Presentation Agents responsible for outputting results to the auditor. The audit results are produced by the audit process and prepared by Presentation Agents according to the auditor's preferred display settings (e.g., a report document or a web-based dashboard).

Evidence Store: The Evidence Store is the central repository for storing evidence records. Some of the more important characteristics of evidence records are, that they are associated with an accountability policy for which they were collected and contain supporting information such as important log entries collected by an agent, which points out a policy violation. For each cloud tenant, there is a separate Evidence Store. This addresses some of the confidentiality and privacy issues associated with a share data pool for potentially sensitive information. Only authorized persons in the role of an auditor may access the Evidence Store.

Multi-layer Evidence Collection:
Collecting evidence in a cloud infrastructure is a very complex process. The main problem lies in integrating a multitude of heterogeneous and distributed sources. As a basis for evidence source classification, we use a simple cloud architecture stack as depicted in Figure 1. There, we consider low-level evidence sources, such as data extracted from the network layer using NetFlow or SNMP, information collected on a virtualization host, information collected inside a customer's virtual machine and also information provided by the software layer (as in SaaS logging). Last, but not least, we consider the cloud management system (CMS), such as OpenStack or OpenNebula to be among the most important sources of evidence, since lots of information provided by CMS logging is directly relevant for auditing against accountability policies (e.g., virtual resource life-cycle and data transfer events for data provenance, and authentication and authorization logging for data security).

B. Policy Input and Audit Task Definition

In this Section, we describe the input to the Audit Agent System, which is derived from A-PPL policies. A-PPL policies capture accountability-related obligations in a policy language.

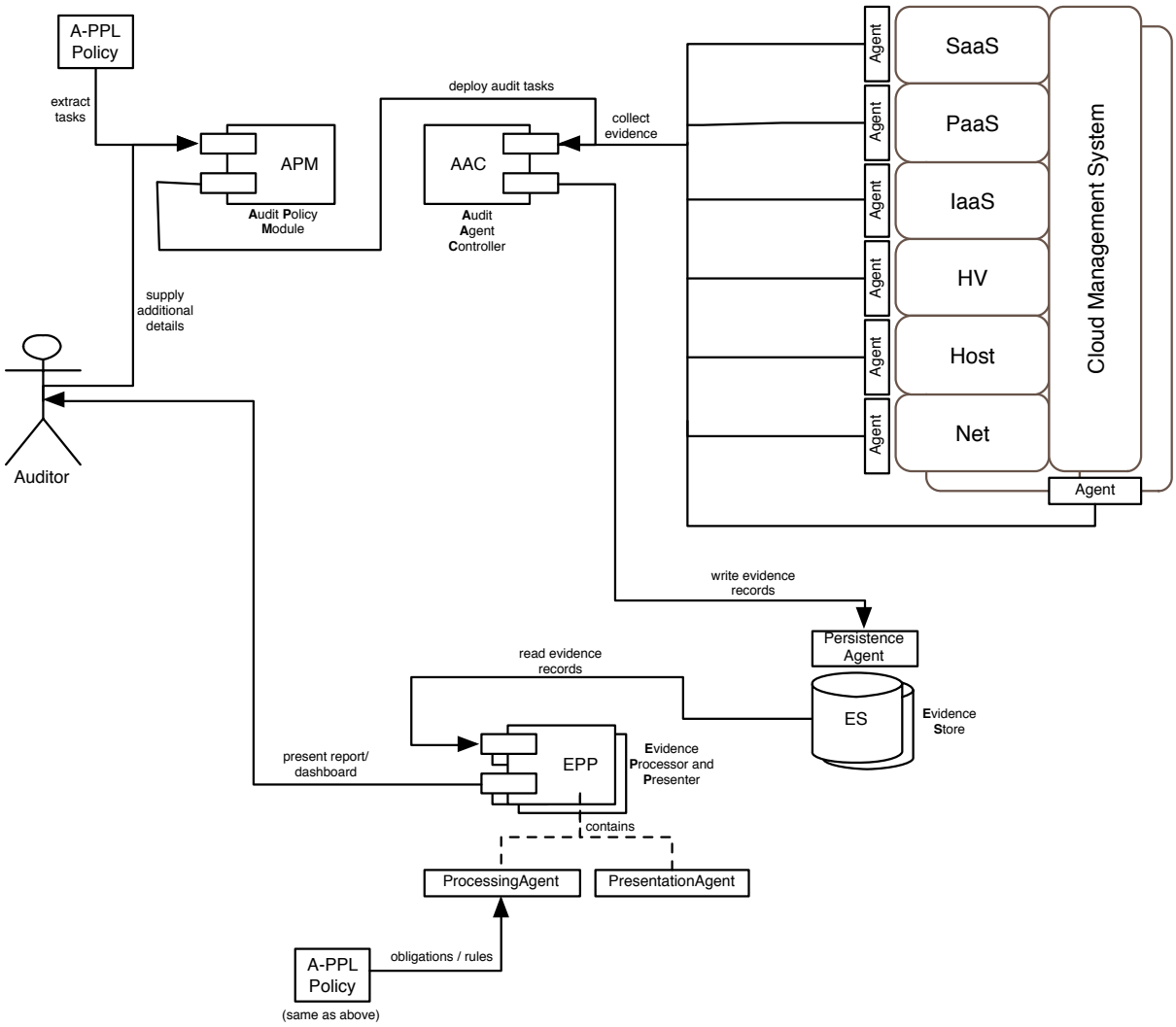


Fig. 1. Audit Agent System Architecture

A-PPL is not part of the Audit Agent System but rather serves as a means for describing what to audit and to decide which accountability requirements need to be fulfilled. A-PPL is developed as part of the A4Cloud research project. It is an extended version of the PrimeLife Policy Language (PPL), which itself is based on the well-established XACML access control management policy language. Therefore, XML as a defining technology, is a given.

Figure 2 depicts the policy-related input as well as the process of deriving audit policies from A-PPL policies. Based on the A-PPL policy input, Audit Tasks are extracted. An Audit Task is a combination of an evidence collection agent (describes where to collect information using which tool), its configuration (which information to collect from a possibly very large pool) and thresholds (limits and conditions that constitute a policy violation). Audit tasks are prepared somewhat similar to templates. For instance, a cloud management system agent is a program that is able to interface (e.g., via the logging and monitoring API) with the CMS and extract certain information. For this, it needs a basic configuration (e.g., how to connect to the CMS). The program and the basic configuration form a template. In the actual audit, the template is populated with all the required basic information (such as authentication credentials and IP address of the CMS), the actual information to collect (e.g., the agent is instructed to build a list of all life-cycle of a virtual machine in a specific time-frame) and possibly a failure condition (e.g., snapshot events are a policy violation).

An Audit Policy, similar to an A-PPL policy containing multiple rules and obligations, contains at least one Audit Task. Several Audit Tasks may need to be executed to be able to evaluate a policy. Performing the reasoning in case of multiple evidence items is part of the Evidence Processor and Presenter.

C. Audit Data Flows

As described in Section III-A, the actual flow of information in the Audit Agent System can be quite complex. In this Section, we describe three different layers that evidence data has to pass through from the collection up to the presentation

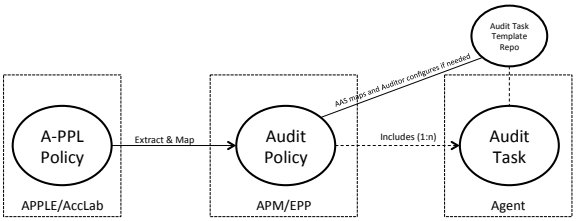


Fig. 2. Audit Agent System Policy Input

of audit results. An overview of this process is presented in Figure 3.

Raw Data: On this layer, information that can be used as evidence is generated. In Section III-A, we mentioned that evidence can be produced by diverse tools in a common cloud infrastructure. Typically, such evidence is generated in the form of logging information, cryptographic hashes (e.g., of files), configuration details or the output of analysis tools, such as from digital forensics tools. Logs, while being similar in general structure (e.g., typically one line per event, beginning with a time-stamp), they differ very much in the used syntax (e.g., time-stamp format, order of event elements, etc.). In our approach, this problem is addressed by interfacing with evidence sources on the Raw Data layer individually. More precisely, for every data source, there is a specialized agent, which is aware of the syntax, semantic and interfaces of the evidence source on one side and of the syntax, semantic and interfaces of the Audit Agent System on the other side. The method of interfacing with an evidence source can be diverse as well. For instance, the agent may (I) use an evidence-generating tool's API to collect information, (II) monitor log files or (III) parse the output of analysis tools.

Agent: Software agents collect evidence data from the Raw Data layer, where it is produced. An agent has two major components, a *Collector* and a *Minimizer*. The Collector interfaces with the evidence source and extracts evidence data. The Minimizer performs several pre-processing actions on the collected data. It removes unnecessary information to reduce the amount of data before transmission to the Audit Agent System core components. The anonymizer is an optional component that tries to remove sensitive information from the collected data in order to protect the confidentiality and privacy of affected persons. How data is to be anonymized and whether or not the anonymization/removal is actually feasible or negatively impacts the audit results has to be decided on a case-by-case basis by an auditor during preparation of the audit task. In any case, the principle of only collecting data that is absolutely required plays an important role at this point of evidence processing and should be observed during audit policy creation. The pseudonymizer works the same way as the anonymizer but allows reversal.

Evidence Processing & Presentation: After the evidence data has been collected and preprocessed by the agent, it is passed to the Evidence Processing & Presentation component. Here, the data passed by the agent is processed by the Evaluator, Aggregator and Presenter components. These components are software agents themselves, but together they logically form the EPP component. The Evaluator is used to

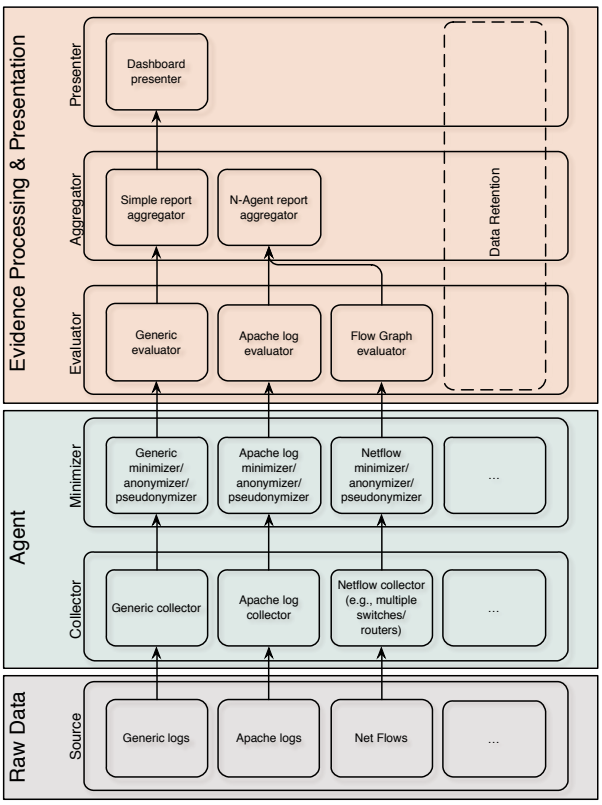


Fig. 3. Audit Agent System Data Flow

compare evidence collected by multiple agents against the policy. Since a single agent can only evaluate parts of the more complex audit policy, the Evaluator is required to put the individual results into context and generate an audit result for the whole audit policy. Evaluators are implemented as an additional agent type (similar to collectors, and aggregators). The evaluation function greatly depends on the input policy and can be as simple as keyword search in text files but also more complex when time lines from various log sources need to be constructed and analyzed. The Aggregator is used to combine the results of multiple audit policies into a single base for the Presenter. There are multiple Presenters, one for each method of presentation and also differing in level of detail depending on the technical knowledge of the auditor. It is very common to have an audit report as a document, which includes the audit result (compliance statement) and if necessary supporting evidence that has been collected. Such documents can be generated automatically to some degree. This form of presentation is most useful, when audit intervals are quite long (for instance in a monthly audit). There is also the presentation of the results in a web-based dashboard, as it is commonly done in monitoring solutions. This approach is more useful, if intervals are short or auditing is done continuously (i.e., as soon as a change event triggers a re-audit), because results can be displayed immediately.

IV. Conclusion

In this paper, we presented a software architecture for performing accountability audits on cloud ecosystems. We

based our approach on the use of software agents, to address problems arising from the wide range of data sources producing evidence and the dynamics of cloud infrastructures. The Audit Agent System is extensible, by allowing to easily develop new agents either on the collection, processing or presentation layer.

We also discussed the input and output interfaces of the Audit Agent System to demonstrate, how such a system can potentially be used by a cloud auditor to automate audit tasks and enable continuous auditing.

By providing cloud customers with such auditing functionality, transparency of cloud services as well as data processing in the cloud can be increased, which may have positive influence on the trust in such services. Additionally, the proposed system enables cloud providers to demonstrate, that they are acting according to the agreed upon policies (between them and their customers), which is a major part of demonstrating accountability.

REFERENCES

[1] IDC for the European Commission, "Quantitative estimates of the demand for cloud computing in europe and the likely barriers to take-up," http://cordis.europa.eu/fp7/ict/ssai/docs/study45-d2-interim-report.pdf, [retrieved: Sep 2014] 2012.

[2] S. Pearson, "Toward accountability in the cloud," *Internet Computing, IEEE*, vol. 15, no. 4, pp. 64–69, July 2011.

[3] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," National Institute of Standards and Technology, Information Technology Laboratory, Tech. Rep., 2011. [Online]. Available: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

[4] Cloud Security Alliance (CSA), "The notorious nine - cloud computing top threats in 2013," https://downloads.cloudsecurityalliance.org/initiatives/top_threats/The_Notorious_Nine_Cloud_Computing_Top_Threats_in_2013.pdf, [retrieved: Sep 2014] 2013.

[5] National Institute of Standards and Technology (NIST), "Guidelines on security and privacy in public cloud computing," http://csrc.nist.gov/publications/nistpubs/800-144/SP800-144.pdf, 2011.

[6] A. Haeberlen, "A case for the accountable cloud," *SIGOPS Oper. Syst. Rev.*, vol. 44, no. 2, pp. 52–57, Apr. 2010. [Online]. Available: http://doi.acm.org/10.1145/1773912.1773926

[7] D. J. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler, and G. J. Sussman, "Information accountability," *Commun. ACM*, vol. 51, no. 6, pp. 82–87, Jun. 2008. [Online]. Available: http://doi.acm.org/10.1145/1349026.1349043

[8] F. Doelitzscher, T. Ruebsamen, T. Karbe, C. Reich, and N. Clarke, "Sun behind clouds - on automatic cloud security audits and a cloud audit policy language," *International Journal On Advances in Networks and Services*, vol. 6, no. 1 & 2, 2013.

[9] Distributed Management Task Force, Inc. (DMTF), "Cloud auditing data federation (cadf) - data format and interface definitions specification," http://www.dmtf.org/sites/default/files/standards/documents/DSP0262_1.0.0.pdf, 2014.

[10] Cloud Security Alliance (CSA), "Cloud Trust Protocol," https://cloudsecurityalliance.org/research/ctp, [retrieved: Sep 2014].

[11] O. Q. Zhang, M. Kirchberg, R. K. L. Ko, and B. S. Lee, "How to track your data: The case for cloud computing provenance," HP Labs, Tech. Rep., 2012.

[12] S. Worku, Z. Ting, and Q. Zhi-Guang, "Survey on cloud data integrity proof techniques," in *Information Security (Asia JCIS), 2012 Seventh Asia Joint Conference on*, 2012, pp. 85–91.

[13] T. Garfinkel and M. Rosenblum, "A virtual machine introspection based architecture for intrusion detection," in *In Proc. Network and Distributed Systems Security Symposium*, 2003, pp. 191–206.

[14] Nagios Enterprises, LLC, "Nagios," http://www.nagios.org/, [retrieved: Sep 2014] 2014.

[15] New Relic, Inc, "New relic," http://newrelic.com/, [retrieved: Sep 2014] 2014.

[16] F. Liu, J. Tong, J. Mao, R. Bohn, J. Messina, L. Badger, and D. Leaf, "Nist cloud computing reference architecture," http://www.nist.gov/customcf/get_pdf.cfm?pub_id=909505, 2011.

[17] "A4Cloud FP7 Project," http://www.a4cloud.eu/, [retrieved: Sep 2014] 2014.

[18] C. A. Ardagna, L. Bussard, S. D. C. D. Vimercati, G. Neven, S. Paraboschi, E. Pedrini, S. Preiss, D. Raggett, P. Samarati, S. Trabelsi, and M. Verdicchio, "Primelife policy language," http://www.w3.org/2009/policy-ws/papers/Trabelisi.pdf, [retrieved: Sep 2014] 2009.

[19] M. Azraoui, K. Elkhiyaoui, M. Önen, K. Bernsmed, A. Santana De Oliveira, and J. Sendor, "A-PPL: An accountability policy language," in *DPM 2014, 9th International Workshop on Data Privacy Management, September 10, 2014, Wroclaw, Poland*, Wroclaw, POLAND, 09 2014. [Online]. Available: http://www.eurecom.fr/publication/4381

Future of Logging in the Crisis of Cloud Security

Sai Manoj Marepalli
University of Applied Sciences Offenburg
Badstrasse 24, 77652 Offenburg
smarepal@stud.hs-offenburg.de

Razia Sultana
University of Applied Sciences Offenburg
Badstrasse 24, 77652 Offenburg
razia.sultana@hs-offenburg.de

Andreas Christ
University of Applied Sciences Offenburg
Badstrasse 24, 77652 Offenburg
christ@hs-offenburg.de

***Abstract*— Logging information is more precious as it contains the execution of a system; it is produced by millions of events from simple application logins to random system errors. Most of the security related problems in the cloud ecosystem like intruder attacks, data loss, and denial of service, etc. could be avoided if Cloud Service Provider (CSP) or Cloud User (CU) analyses the logging information. In this paper we introduced few challenges, which are place of monitoring, security, and ownership of the logging information between CSP and CU.**
Also we proposed a logging architecture to analyze the behaviour of the cloud ecosystem, to avoid data breaches and other security related issues at the CSP space. So that we believe our proposed architecture can provide maximum trust between CU and CSP.

***Index Terms*— logging; Big Data; machine learning, cloud computing, data security.**

I. INTRODUCTION

We are in the midst of data revolution; adoption of digital technology is growing higher than before. Every individual in this world is directly or indirectly responsible for the growth of data. According to CSA research report everyday human beings generate 2.5 quintillion bytes of data [1], which is coming from every influence of human life, and it is too big and too complex. A short survey to show how big the data is, by Royal Pingdom Internet 2012 in numbers says every day, total 144 billion emails are exchanged worldwide, 5 billion times +1 button is used in Google+, 2.7 billion numbers of likes in Facebook, 300 million new photos are added in Facebook, 175 million tweets are posted in Twitter [2], this data which is generated everyday is highly unstructured. To extract the knowledge from this highly unstructured data, advance computing technologies like cloud and Big Data are really helpful [3].
The characteristics of Big Data described according to Gartner, which are divided into 3V's 1) Volume 2) velocity and 3) Variety, as shown in Fig 1.
Current cloud architectures can be abode of numerous Big Data technologies[4]; to compute the Big Data problems, technologies like cloud computing can help in more efficient way because of its inbuilt characteristics.
According to [5], cloud features are provision of on-demand

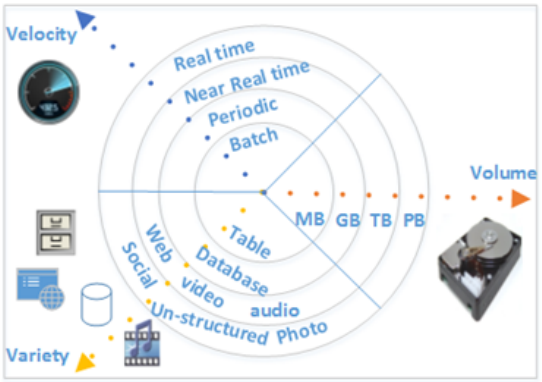


Fig. 1. Gartner 3V's Volume, Velocity, and Variety [4]

computing services with high reliability, scalability, and availability in distributed environments. For most of the IT organizations cloud is a paradigm shift, where computing power, data storage, and services are outsourced to Cloud Service Provider (CSP) and made them available as commodities. Services from the CSP are basically divided into three different forms 1) Software as a Service (SaaS), 2) Platform as a Service (PaaS), and 3) Infrastructure as a Service (IaaS) [3], [5].

A question to use cloud services from CSP, is still needed to be addressed by most of the IT organizations before planning to migrate their IT operations, from their in-house computing architectures to across the internet and into cloud computing. This question arises few opportunities and limitations for the IT organizations to migrate their infrastructure to the cloud
Opportunities, Applications/data are replicated across different geographical locations, and are made available at anytime anywhere through internet. Also running those applications in cloud is very less expensive
Limitations, Applications/data are stored in unknown hosts, which can triggers concerns about data privacy and security by IT organizations [5]
In this paper we tried to address the privacy and security concerns of the IT organizations, by monitoring interactions to their services/data when they are in rest, or in live within the cloud. To give an idea about privacy and security issues from a the perspective of IT organization, one of the most common privacy concern can be lack of control [6]. Because most of

the services in the cloud are currently controlled by the cloud facilitator/CSP this leads IT organizations to fear/concern about their sensitive/private information. Likewise, one of the many other security concerns is fear of vulnerabilities of cloud services [6]. What if a malware-based attack such as worms, viruses, and DoS exploit vulnerabilities of a cloud service? It gives a chance for intruders to gain unauthorized access and get hold on to their critical information [7].
The goals of the proposed architecture is to prevent, detect, and respond to the vulnerabilities of the cloud space, which enables CSP and CU to be proactive instated of being reactive towards data/application thefts.
Log analysis can help security response teams at CSP or at CU to

- Prevent hijacking the control from intruders
- Minimize damage of the theft and help for forensic analysis
- Catch exploiters before they succeed

The rest of the paper is organized as follows. Background and Challenges are described in Section II. Section III Related Works. Section IV describes the Logging Infrastructure at Cloud with MAR Principle. The proposed architecture is described in Section V and conclusion and future work is part of Section VI.

II. Background and Challenges

With the infinite computing power, minimum cost and maintenance cloud looks very advantageous for IT organizations to migrate their IT operations into the cloud infrastructure. But unfortunately most of the current cloud architectures often lack in providing trust to the IT organizations in terms of transparency [8], we try to discuss it more in this section.
Current cloud solutions are like a black box, security measures are always uncertain to the cloud Users (CU) (or) IT organizations. If compared with traditional computing architectures they are always predictable and understandable to the IT organizations. In cloud because of reduced visibility of the underlying infrastructure, it became very difficult for IT organizations to predict or monitor the vulnerabilities within their cloud space.
To predict or monitor the CUs cloud space in cloud infrastructure, there are few challenges they still remain as open and still need to be debated further.
Challenge 1: who will take the responsibility of log monitoring? CSP or CU

- **CSP**
  Questions
  1) **If yes**, will it be cloud user specific logs or whole cloud infrastructure logs?
  2) **If yes**, does CU allow CSP to view his logs? Because logs are considered as a private information by CU
  3) **If no**, responsibility goes to Cloud User (CU)
- **CU**:
  can CU afford another infrastructure to monitor logs?
  Questions
  1) **If yes**, will the infrastructure be within his organization or at another trusted CSP?
  2) **If yes**, will CSP allow CU to view log data? Because it also contain the infrastructure information of CSP, which is his private data
  3) **If no**, uncertainty of monitoring logs still exits

Challenge 2: Collection of Log sources for forensic evidence
Collecting these evidences is more complicated in cloud because of its nature of multi-tenant computing models [9], where each user share same processing and networking resources at CSP. An exploiter can enter into the cloud through virtual machine (VM) and then can exploit the vulnerable applications in the cloud and when he/she goes out from the VM, there will be no evidence to prove the culprit.
Challenge 3: Control over log files, who will take control CSP or CU?
Currently CSP has more control over log files compared to CU. As stated by [9] at present CSP is not providing any log data from the network components. Also it is very difficult to get the log data from SaaS and PaaS where CSP has maximum control over those services. For CU it is very difficult to monitor his cloud space, and for forensic investigators it will be difficult to investigate cyber attacks in cloud infrastructure.
Challenge 4: Protection of Log data
Ensuring log data is not tampered is also very important, when it comes for monitoring security attacks and for knowing evidences for forensic investigations within the cloud ecosystem.
It is very early to decide about the best possible place to record the log files. This can be at CSP or at CU. The question still remains open, yet to be addressed. In this paper we are proposing a Logging infrastructure, which can be carried out at both places either at CSP or at CU. If both of them mutually agree upon sharing log data.

III. Related Works

As cloud computing evolved, concerns related to security and privacy are overseen by cloud users due to lack of transparency between CU and CSP. There are several researchers, who tried to address similar issues on security and privacy in different ways. According to [10] proposed a Public Key Infrastructure operating in concert with Single Sign On (SSO) and Lightweight Directory Access Protocol (LDAP), their idea is to ensure the authentication, integrity and confidentiality of involved data and communications. [7] proposed a new approach for securing the customers virtualized workloads in a cloud, in their idea they are closely monitoring variety of guest operating systems, and quarantined promptly in case of compromise.
The researchers [10], [7] are addressing most important challenging issues in the cloud, but still lot more have to be done for the important issue Trust in the cloud. CU is always skeptical about his/her own cloud space, which we see as the most important issue, but a very little research has done to address this.

[11] their approach is to ensure that any access to cloud users data will trigger authentication and automated logging to the JARs, this approach addresses the accountability in the cloud. Another research [12] SecLaaS they tried to address ensuring confidentiality to cloud users by storing virtual machines logs and provide access to forensic investigators.
The researchers [11][12] tried to partially address the issue of trust by using log data between CU and CSP, but the challenges mentioned in Section II still remain unclear. The question of who owns the Log data, CSP or CU is still remaining as same.
In the proposed architecture we tried to give an overview on processing of log files, either at cloud or at CU infrastructures with big data analysis and using machine intelligence algorithms, still we believe it needs lot of research inorder to answer the specific questions about the algorithms to be used, and log data owning etc.

IV. Logging Infrastructure at Cloud with MAR Principle

Having logging infrastructure at CSP or CU, can help to be aware of different issues in cloud ecosystem. Cloud as a whole generates millions of events from a simple system logins to complex system errors, if they are divided into specific CU based logs, then the complexity of logging can be reduced a bit, still research has to be done to prove it. A simple log file is typically a collection of events, which determines the actions of reading, writing, deleting and modification of data. And log file can also able to determine the process who owns it, when it was initiated, where the action occurred, and why the process ran, and what are the rights [13], [14]. In this paper for efficient monitoring of log data, we followed MAR principle which contains three basic steps to be carried out.

A. MAR Principle

MAR principle is the basic principle which is integrated into the proposed architecture for Monitoring, Analyzing, and Reporting the Cloud Infrastructure. These three phases are proposed in our previous paper for more details [15], and in this paper we elaborate more insights on monitoring and analyzing.

Workflow of MAR-Principle is:

- Monitoring: allows recording the log events from network log-ins to random system errors
- Analyzing: analyzes log files and mines required information out of it, and correlates them with the rules defined by the organization (policies)
- Reporting: if any misbehavior found/occurred it reports to the user proactively.

B. Importance of using MAR-Principle

By monitoring logs a

1) System Administrator in an organization can troubleshoot or fix the system level bugs
2) Developer can find the bugs in application development
3) Forensic investigators can use them for investigation

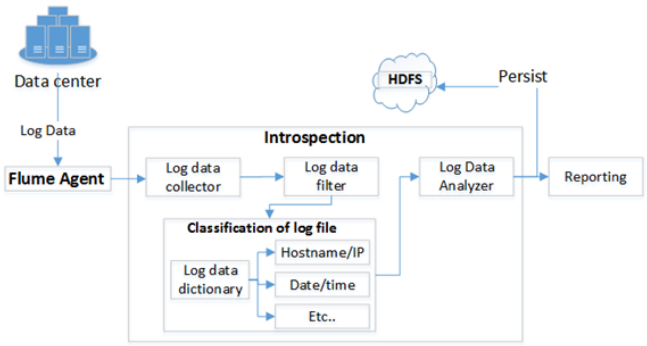


Fig. 2. System architecture for logging cloud data

By Analyzing logs a system can

1) Heal itself in the time of failure
2) automatically revokes the privileges, if an unauthorized user access trying to access

By Reporting

1) Health of the system is communicated
2) CU is aware of actions performed at his/her cloud space

V. Proposed Architecture

As cloud is distributed in nature, collecting logs into a centralized storage from different sources is a challenging task. Tools like Apache Flume are able to overcome this challenge for details see [16], in the proposed architecture Apache Flume is added to collect logs and persist them into a centralized storage.
The proposed architecture is divided into three basic parts as shown in Fig 2

1) Flume Agent
2) Introspection
   - Log Data Collector (LDC)
   - Log Data Filter (LDF)
   - Log Data Analyzer (LDA)
3) Reporting

A. Flume Agent

It is responsible for collecting the log data and putting them into a centralized storage, from different sources within the data center. The log data can be divided in to User Specific (US) or Cloud Specific (CS) mainly it depends on who take responsibility for monitoring log sources.

B. Introspection

Inside the Introspection phase, three processes are handled 1) Log Data Collector (LDC), 2) Log Data Filter (LDF), and 3) Log Data Analyzer (LDA).
LDC is a temporary buffer space, which acts as a tunnel to carry the log data from generated sources to LDF.
LDF in the LDF process, redundant as well as unimportant information from the log sources are discarded. To reduce the size of the log data.

Further classification mechanism is introduced to classify the log data; this uses Log Data Dictionary (LDD). LDD classify the filtered log data according to the list of dictionary values as show in Fig 2, and makes sure that all the single row items in every column are unique. For example, if there is an IP address repeated multiple times on the same data in different timings then that particular IP address is classified by the IP value and different events performed on the same date in different times.

Once the data is classified according to the LDD, then the Modified Log Data (MLD) are sent for analysis to LDA.

LDA as mentioned in Section II trust is the major setback for IT organizations to migrate their services into cloud infrastructures [17], yes it is true once the organizations sensitive/private data/applications moves out from their own servers to unknown CSP servers, fear persists. We try to argue bit more on this issue to address it, our argument in this paper is trust can be solved by monitoring the logs, because logs obtain enough information to know, who, what, when, how the person or someone accessed the assets of CU.

This argument can be fulfilled by incorporating two main algorithms 1) correlation analysis 2) behavior analysis, only idea of these two algorithms are described below because it needs still more investigation to finalize the right algorithms.

1) In correlation analysis, organization specific policies and access control rules are compared with the filtered log data, this comparison happens in real-time while log data is streaming from the cloud. This can enables the cloud eco system/CU to promptly respond to the unhandled exceptions, or unpredicted circumstances.
   For example, suppose an IT organization using a Cloud service for managing their Sales Management Service (SMS), where all the employees in the organization are having different access privileges to use the SMS in the Cloud. If one of the employees in the IT organizations gained unauthorized/illegal access to the SMS service, by the help of correlation algorithms which correlates the polices and access control rules with the filtered log data can stop the user to gain more control over the SMS.
   Correlation algorithm needs further more investigations on
   - How to compare policies with the filtered log data
   - Time constraint, how fast can this algorithm will find unauthorized users, and deny their access
2) Behavior Analysis, for understanding the behavior of the cloud space of each CU, more investigations have to done. There are very few researchers who tried to address this issue, but there are many reasons to defend the idea of behavior analysis of individual CU. Few of them can be
   - Difficulties in understanding the user behavior among different cloud services, because everything in cloud is referred as a service XaaS [5]
   - Multiple tenants are sharing common physical infrastructure, to understand their behavior is a challenging task
   - Analyzing CU behavior when more than one cloud is involved (Hybrid cloud), in the case of federated cloud environments [18]
   - Self healing the cloud space after any malicious attack
   - Pulling or pushing of log data among different servers for analyzing

   Challenges are there but it is not an impossible task to complete, due to enormous capabilities in processing huge unstructured data [19]. Big Data technologies and with the help of machine intelligence algorithms can able to provide a solution.
   Through analyzing behavior of the cloud space, a system tries to give intelligence to the log data. In order to predict vulnerabilities of the cloud services, and stop future attacks which are not imagined while developing the program.
   For example, the Heartbleed bug [20] occurred in 1st April 2014, was happened because of developer writing vulnerable code, and it was noticed after approximately one year. This vulnerability can find even faster if sophisticated monitoring technologies are developed [14].

C. Reporting

The real value of proposed log management architecture can be seen, when the process such as filtering, and analyzing functions simultaneously and leads critical events to respond an immediate alert.

Reporting must respond to strong likelihood of malicious activity, excessive system activity, unexpected system activity, and when machine critical application performance fails.

Reporting also must be aware to respond efficiently in different user behaviors as shown:

- Normal behavior (as stable state)
- Confidential Behavior (special permissions acquired by admin are in this state)
- Unauthorized behavior (Anonymous person accessing the system without proper privileges)

Reporting tool can also be carried out with the aggregated statistics of the events, based on the Alerts, security issues, denial of service attacks, intruder attacks etc.

VI. Conclusion and Future Work

Securing the cloud infrastructures from malicious activities, vulnerabilities, and intruder attacks, etc is a challenging task with existing security mechanisms, because of the growing size of the data. In this paper we tried to address the above mentioned challenge by proposing a new logging architecture by processing log files. Log files contains precious information, it contains the execution of the program if log files are mined and analyzed with proper care many of the issues related to security and privacy can be solved in the cloud infrastructure. In our approach by analyzing log data, starting by filtering redundant data, classifying them,

correlating with the policies, analyzing behavior for finding vulnerabilities and then reporting the state to the CSP or to the CU helps in increase the trust between CU and CSP. Still we believe that lot of research has to be done, to answer specific questions about transparency, analysis algorithms, and log data ownership.

REFERENCES

[1] "Big Data Analytics for Security Intelligence," Cloud Security Alliance, 2013.

[2] "Royal Pingdom," 16 January 2013 . [Online]. Available: http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/. [Accessed 20 August 2014].

[3] M. Peter and G. Timothy, "The NIST Definition of Cloud," National Institute of Standards and Technology, Gaithersburg, 2011.

[4] Stamford, "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data," Gartner, 27 June 2011. [Online]. Available: http://www.gartner.com/newsroom/id/1731916. [Accessed 18 September 2014].

[5] P. R. Bhaskar, J. Admela, K. Dimitrios and G. Yves, "Architectural Requirements for Cloud Computing Systems: An Enterprise Cloud Approach," Springer Netherlands, vol. 9, no. 1, pp. 3-26, 2011.

[6] S. S and K. V, "A survey on security issues in service delivery models of cloud computing," Journal of Network and Computer Applications, vol. 34, no. 1, p. 111, 2011.

[7] C. Mihai, S. Reiner, L. S. Douglas, S. Daniele and S. Daniele, "Cloud security is not (just) virtualization security: a short paper," in 16th ACM Conference on Computer and Communications Security, Chicago, 2009 .

[8] P. Siani and C. Andrew, Accountability as a Way Forward for Privacy Protection in the Cloud, Berlin Heidelberg: Springer , 2009.

[9] D. Birk and C. Wegener, "Technical Issues of Forensic Investigations in Cloud Computing Environments," in Systematic Approaches to Digital Forensic Engineering (SADFE), 2011 IEEE Sixth International Workshop on, Oakland, CA, 2011.

[10] Z. Dimitrios and L. Dimitrios, "Addressing cloud computing security issues," Future Generation Computer Systems, vol. 28 , no. 3, pp. 583-592, 2012.

[11] S. Sundareswaran, A. Squicciarini, D. Lin and S. Huang, "Promoting Distributed Accountability in the Cloud," in Cloud Computing (CLOUD), ., Washington, DC, 2011 .

[12] Z. Shams, K. D. Amit and H. Ragib, "SecLaaS: secure logging-as-a-service for cloud forensics," in Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security, 2013.

[13] B. Jacob, K. Dave, F. C. J. Everett and F. Jeremy, Security Log Management: Identifying Patterns in the Chaos, Syngress, 2006.

[14] K. Karen and S. Murugiah, "Guide to Computer Security Log Management," National Institute of Standards and Technology, Gaithersburg, 2006.

[15] S. M. Marepalli, R. Sultana and A. Christ, "Introduction to MAR Principle: A Log-based Approach towards Enhanced Security in Cloud," in Software-Technologien und -Prozesse, Furtwangen, Walter de Gruyter GmbH, 2014 , p. 101114.

[16] "Apache Flume," 16 July 2014. [Online]. Available: http://flume.apache.org/. [Accessed 10 August 2014].

[17] H. Kai and L. Deyi, "Trusted Cloud Computing with Secure Resources and Data Coloring," in Internet Computing, IEEE , 2010.

[18] H. Xueli and D. Xiaojiang, "Efficiently secure data privacy on hybrid cloud," in Communications (ICC), Budapest, 2013.

[19] S. Lohr, "The Age of Big Data," 11 February 2012 . [Online]. Available: http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted. [Accessed 15 September 2014].

[20] K. Mohit, "HeartBleed Bug Explained," The Hacker News, 14 April 2014. [Online]. Available: http://thehackernews.com/2014/04/heartbleed-bug-explained-10-most.html. [Accessed 15 September 2014].

Autonomic Service Level Agreement Management as a Service

Stefan Frey, Claudia Lüthje, Christoph Reich
Furtwangen University
Cloud Research Lab
Furtwangen, Germany
{stefan.frey, claudia.luethje, christoph.reich}@hs-furtwangen.de

***Abstract*—After an initial hype, cloud computing is establishing itself as adequate means of providing resources on demand. By now cloud computing provides a practical alternative to, locally hosted resources for companies. This paper presents the basic concept and architecture for independent, autonomous management of Service Level Agreements (SLAs) for cloud environments, which was developed as part of the research project Autonomic SLA Management as a Service (ASLAMaaS) at Furtwangen University. The purpose of the presented architecture is to enable cloud users to modify the SLAs of his "outsourced" cloud services at any time to adapt to the changing requirements of the business processes. The proposed ASLAMaaS architecture therefore is based on the MAPE-K concept and provides the cloud customers with an easy to use SLA control interface to adjust the quality of service at any time. Through this effective use of cloud services and cloud resources the overall costs can be reduced. Due to this adaptability, the use of cloud resources can be made more efficient and business risks can be reduced.**

I. Introduction

According to a market analysis by the Gartner Group [8], the IT budgets of German companies has been reduced by 2.7% in 2011. The study also predicted that many companies will increasingly rely on outsourcing their IT to cloud computing to reduce costs. The idea behind cloud computing is to deliver computing resources on-demand over a network on an easy pay-per-use business model[1]. Due to the low upfront costs, rapid provisioning, elasticity and scalability, the adoption of cloud services is steadily increasing [2]. In order to make cloud services effectively usable [3] and reliable for enterprises [4], service level agreements (SLA) are needed, which state the precise level of performance, as well as the manner and the scope of the service provided.

This practice, which is widespread in the area of IT services, is currently of limited use for cloud computing, due to the fact that existing cloud environments offer only rudimentary support and handling of SLAs, if any. Moreover, the classic SLA management approach is a rather static method, whereas due to the dynamic character of the cloud, the QoS attributes respectively service levels must be monitored and managed continuously [5]. In addition, performance indicators [7] and measurement methods for service level objectives in cloud computing have been studied inadequately [6].

The proposed research focuses on cloud specific SLA management as well as the associated functional and non-functional QoS parameters and measurement methods. In addition, an architecture for providing autonomous management of cloud services together with an easy to use SLA control interface will be presented.

A. Quality of Service in Cloud Computing

In general SLA contracts describe the exact service quality a user can expect, how fast a provider must response in case of problems and what redress the provider has to give when the SLA contract gets violated and the user suffers a loss of business. SLAs are the cornerstones of every IT service provider to deliver services to its customers. According to a survey by Vanson Bourne, commissioned by Compuware [9], German companies suffered heavy losses due to poor performance in cloud applications. More than half a million euros, is the average annual loss due to lack of SLAs according to the study.

For the users of cloud services, especially small and medium sized businesses, it would be very desirable to find a cloud provider who can guarantee the quality of the provided services by offering and enforcing SLAs. Cloud infrastructures offer the potential to negotiate individual SLAs and adapt services on-demand, but this is currently not utilized. Large cloud providers such as Amazon are currently not willing to offer customer-specific SLAs and offer only rudimentary "one size fits all" general agreements. In the case of Amazon, for example, this means they guarantee all customers a general availability of 99.95% for their cloud services but for their Elastic Block Store (EBS) services no service quality guarantee is given.

In addition, problems can arise from the international locations of such large providers, for example when a breach of contract occurs the jurisdiction may lie outside of the EU. Therefore analysts of the Experton Group [10] recommend german business customers to choose cloud providers with german contracts and service level agreements (SLAs), and local jurisdiction. Additionally current cloud services are hard to monitor for the customer, because none or only sparse information is given by the provider. However, for businesses, it is essential to monitor their services and check on the compliance with their SLAs.

During a lifecycle of a service a company is often confronted with changing demands, which is often reflected in a change of the agreed service quality, emergency procedures, costs, legal compliance, and so on. The classical approach of negotiating SLAs is a very static based process. Unfortunately, this can not keep up with the dynamic character of cloud

computing. For this new methods have to be created the cope with the fast-paced dynamics, but so far there is no solution on this field. All these problems can make cloud computing unattractive for small and medium sized enterprises (SMEs).

II. Related Work

Many of the cloud-specific service level objectives (eg resource reservation and allocation times, scaling, load balancing, etc.) are not yet sufficiently modeled and must be further investigated. Many of the technical parameters in a cloud are variable and dynamically changeable, which make optimizations, such as the search for the optimal allocation of resources, to a multi-dimensional search problem. Goudarzi & Pedram [15], for example, tried to optimally allocate resources under consideration of technical parameters (Memory, CPU, etc.) with an intelligent search. The monitoring and reporting for SLAs presents also a wide field for research, because cloud infrastructures are so dynamic. For example it takes information from the cloud user to tell the difference between a non-available resource and a deliberately disconnected resource. Also it requires a lot of logging and keeping backlogs since all resources and their conditions have to be documented. Currently, the provider is only able to provide simple technical data such as load, memory usage, and so on, which is given by the cloud management system or the resources themselves. The White Paper ”SLA-Driven Dynamic Resource Management for Multi-tier Web Application in a Cloud” [16] clearly indicates that the utilization of a virtual environment can be visualized, however this does not mean that the actual weak points of the system are knowable. This makes it very difficult for provider to predict on their infrastructures.

There are currently no suitable interfaces for virtualization software like VMWare, Xen or KVM, which are able to take on SLA parameters directly. Standards such as the Open Virtualization Format (OVF) [17], which describes an open interface for packaging and distributing virtual appliance and software or DeltaCloud [18], which provides a cloud technology independent REST API for Clouds, are still only base specifications. Teckelmann et. al. [11] a study on standards and interoperability criteria of the service model Infrastructure as a Service (IaaS) compares some of the currently existing standards.

Lately there are some approaches originated trying to integrate SLAs within cloud environments. The previously known approaches deal only either with the negotiation of SLAs between the customer and the provider, such as SLA Support System for CC (SLACC) [19], which automatically select the appropriate SLA, based on KPIs and SLOs, or to better guarantee SLA-specific sub-area SLOs. Casalicchio and Silvestri [20] analyzed the problem from the perspective of an Application Service Provider that uses a cloud infrastructure to provide scalable services according to QoS, but this approach is limited to VM load distribution. Especially for a cloud databases, Pengcheng Xiong et.al. [21] used a machine learning approach to intelligently manage resources among the clients. An approach to SLA negotiations between different Cloud (interoperable) is discussed in [22] and [23].

III. Autonomic SLA Managment as a Service

To address the shortcomings of current state of the art cloud computing SLA landscape and enable customers to dynamically create, adjust and monitor SLAs for their cloud services, the research projcet Autonomic SLA Management as a Service (ASLAMaaS) investigates in techniques and methodologies to integrate SLA management into cloud computing. The goal of ASLAMaaS is, by using a Software as a Service (SaaS) architecture, to monitor the agreed service performance, facilitate SLA compliance and enable dynamic customer-specific adaptation of SLAs for cloud services. For this purpose cloud management system interfaces will be used to allow monitoring and adaptation of services in a manner as if the services would be made available to the local IT environment. For an all-encompassing SLA management process the cloud services must be integrated. Crucially for this is the usage of existing standards such as the Open Virtualization Format (OVF), Delta Cloud, Unified Service Description Language (USDL), Web Service (WS) -Agreement, etc. [11]. If a cloud provider offers such interfaces, a way simpler monitoring and management of cloud services is possible.

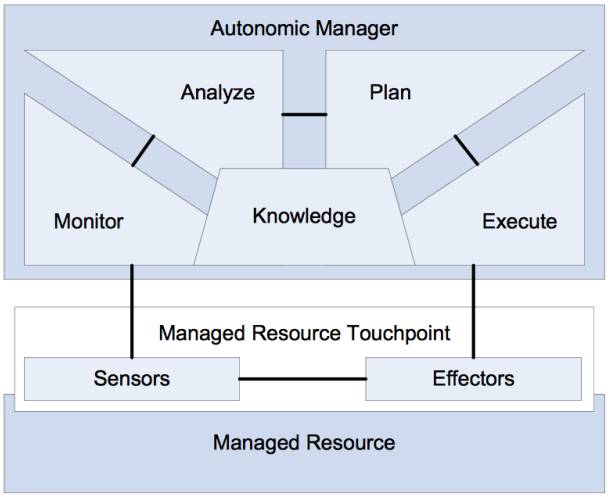


Fig. 1: Autonomic Computing Control Loop [12]

A. Autonomic SLA Managment as a Service Architecture

As the basis of the ASLAMaaS architecture the IBM developed autonomic manager concept MAPE-K [12] (Monitor-Analysis-Plan-Execute - Knowledge) is used. Automomic computing systems are capable of continuous self-monitoring and adjustment. For the monitoring, a sensor is connected to the managed resource touchpoint, which in our case would be an interface on an observed cloud service. Here the monitoring unit collects the data from the sensors and hands it over to the analysis unit, which then with the help of a knowledge base creates plans. These plans are then carried out by the implementation unit (Execute), which influences the managed resource through the effectors. For example by upscaling a Web server cluster to decrease the observed request response time. The use of the MAPE-K approach within the ASLAMaaS architecture can be seen in Figure 2.

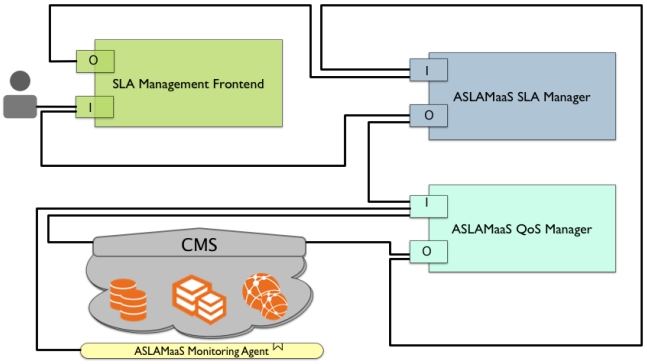


Fig. 2: ASLAMaaS Architecture

In the presented architecture cloud users connects to the SLA Manager via browser or machine interface, where they have the ability to create and adapt SLAs for their cloud services according to their needs. To facilitate the process of creating new SLAs there is a repository of pre-made SLA templates available within the SLA editor. Thus users can easier and faster create their own specific SLAs by selecting a fitting template, which are based on best-practice and experience data for generalized types of services such as web services, databases systems, ERP systems and so on. An overview of the graphical user interface can be seen in Figure 3.

The user chooses key performance idicators (KPIs) from different categories, which he can use for his specific SLA. KPIs describe in each case a specific QoS parameter and the associated metric to monitor. The available KPIs are provider dependent and must be accepted beforehand. The utilized KPIs include, both general QoS parameters such as availability, response time or bandwidth, which are commonly used for almost all services today, but also cloud-specific KPIs, such as the deployment-times for PaaS, virtual image management or scaling schemes. A more detailed presentation of the relevant KPIs and metrics can be found within ”Low level metrics to high level SLAs - LoM2HiS Framework” [13] or ”SLA-Richtliniendokument für Cloud Computing” [14].

Based on the provider accepted KPIs, by using historical infrastructure data and expert know-how, margins are calculated as basis for the offered SLA templates. These margins mark the boundaries in which the user can choose the Service Level Objective (SLO). The expected usage data together with the performance of the infrastructure is matched and business concerns and the strategic focus is taken into account create a pricing model for the offered services and the corresponding SLAs. Figure 4 shows an overview of the SLA Management Frontend and its corresponding components. If, for example an infrastructure constantly delivers a response-time of below 170ms and only a few fluctuations of users are expected the provider may set the margins for this KPI to 200ms or above. The cloud user then can choose which exact response-time for his service he wants to be guaranteed. It’s common practice to make prices not directly on each possible service level, but to create pricing categories, such as low, medium, high or silver, gold platinum for example. This enables a simpler management of the managed service levels and makes the gradation of

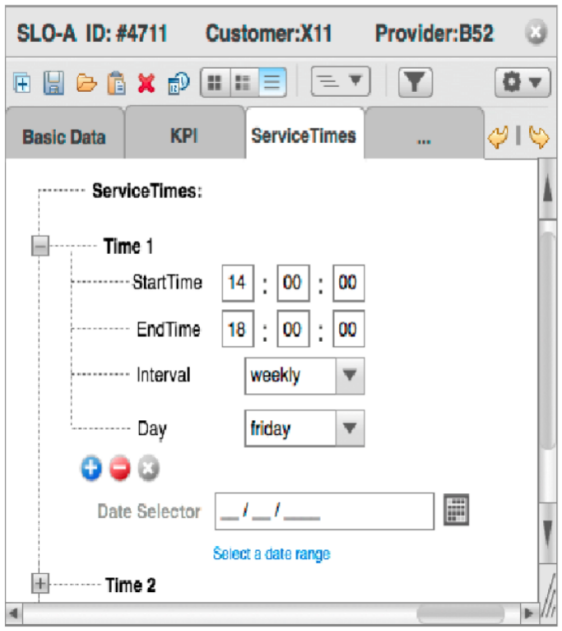


Fig. 3: ASLAMaaS SLA Editor GUI

customers easier for the provider.

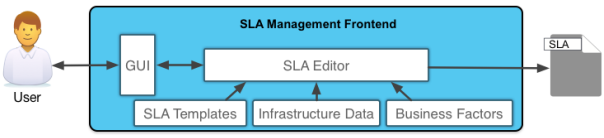


Fig. 4: ASLAMaaS SLA Management Frontend

The pre-defined SLA templates and the finished filled out SLAs used must be represented in a machine processable form. For this an machine readable agreement description language is used. The Adaptable Service Level Objective Agreement (A-SLO-A) model [24], which is based on the SLA* model, enables the use of dynamic and constant agreement alteration and therefore is used as the basis of the SLAs in ASLAMaaS.

After creating a SLA for a specific cloud services within the SLA Editor the ASLAMaaS SLA Manager stores it in his repository and starts controlling it. Such a SLA may consist of several independent or conditional KPIs. To ensure service quality and operate within the agreed on limitations every KPI has to be controlled and monitored separately. This means for each of the KPIs guaranteed in a SLA an instance of the ASLAMaaS QoS Manager, as seen in Figure 2, takes control of the corresponding parameters and starts monitoring them. Every instance of the QoS Manager uses the autonomic MAPE-K principle. An overview of the ASLAMaaS QoS Manager and its components is shown in Figure 5.

Starting by the Monitor the KPI relevant technical parameters are collected. This is done by using the stored metrics for each KPI deposited within the Knowledgebase. Such metrics can be measured either directly within the Cloud Management System (CMS), like for example the availability

of the resource, or are collected directly by connecting to the resource or using a monitoring agent.

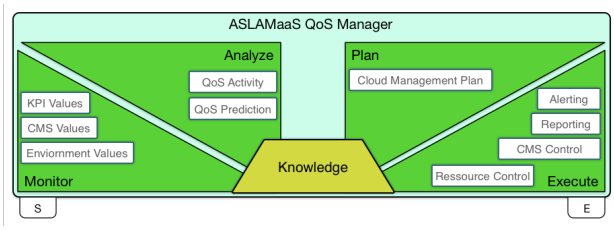


Fig. 5: ASLAMaaS QoS Manager

The Analysis part of the QoS Manager investigates whether or not the recorded service quality behavior is within the SLA specifications or not. Within the Analyze part prediction algorithms are used to estimate future problems. Based on these calculations, respectively the current state of the monitored parameters, an action Plan is conceived. Such plans may consist of actions like scaling up or down, allocation or deallocation of resources, altering the infrastructure, or alerting the corresponding service provider if an adaption is not automatically possible. The QoS manager can thereby intervene directly or indirectly with the CMS. By executing these plans both the cloud infrastructure and the reporting are operated.

Reporting is an important core component of the SLA management, since it used as base for the billing of the claimed services, but also is used as evidence in case of SLA violations. Providers have to exhibit to their customers that the in a SLA agreed service levels have been consistently met, or were within the defined deviations. Therefore the reporting is integrated in the Execute part of the QoS Manager as well as in the SLA Manager of the ASLAMaaS architecture.

An overview of the ASLAMaaS SLa Manager and its components is shown in Figure 6. The SLA Manager collects data about all the QoS parameters and KPIs which belong to a SLA. Additionally data about the status and the relations of all the SLA, which are currently being managed by the system is collected. Within the Analyze part this data is then processed in order to perceive the current state and predict the future behavior of the managed SLAs. Here the mutual influencing SLAs and the strategic business direction comes to bear, and it may also be decided at the violation of SLAs which contracts are abandoned in favor of others.

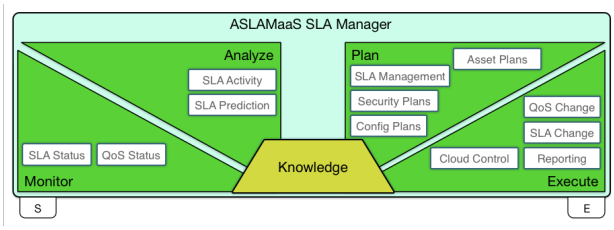


Fig. 6: ASLAMaaS SLA Manager

The Plan part of the MAPE-K concept in this module creates the action plans for the infrastructure, the CMS, the SLA Manager itself and additionally delivers adjustments directly back into the SLA Editor, where for example the margins could be readjusted to cope with the changes. Thus the SLA templates and the corresponding KPI margins are always compliant with the actual performance of the cloud system. The action plans result in adaptation of the cloud infrastructure, like for example if the stated response-time inside a certain SLA tends to be broken the SLA manager can instruct the virtual network agent to re-route the traffic if the problem is network dependent. If this problem is related to overloaded CPU or virtual instances the SLA Manager could start additional new instances or allocate more CPU cores.

The various methods to manage the QoS parameters have to be modeled individually. A sample for the regulation of the KPI response-time can be found at Frey et al. [25]. There scalable cloud services have been started and stopped based on a fuzzy control set. This and other similar control mechanisms enable the Execute part to adapt the infrastructure and the services so that SLA violations can be avoided and the QoS is guaranteed. Within both MAPE-K loops the Knowledge consist mainly of the historic data about the services, the SLAs and the cloud environment, which was measured continuously and expert knowledge in the form of best practices and empirical values as well as strategic business plans.

IV. Conclusion & Future Work

Within ASLAMaaS based on historical monitoring data and provider expertise SLA templates are determined for each cloud service type. Cloud customers can therefore specify SLA contracts freely and without any provider negotiation based on the deposited margins for each KPI. ASLAMaaS serves both the cloud customer and the cloud provider for SLA management and monitoring of the defined service qualities and is well integrated the client's and provider's IT management. Due to the autonomous character of the presented architecture, it is possible that cloud services and environment are adapted constant and continuously. In case of impending SLA violations ASLAMaaS can automatically counteract with the help of predefined policies and action plans, for example by providing additional resources or readjust the allocation. Furthermore, new, cloud-specific SLAs such as service-scaling, service reservation, peak-usage scaling, etc. can be enabled and monitored. These special cloud-specific mechanisms should be examined further to comply better with the goal, to provide and guarantee the qualities of services. Additionally, the search for new optimization techniques and processes and prediction methods for utilization and behavior of the cloud infrastructure, offer further topics for future work.

Acknowledgment

This research is supported by the German Federal Ministry of Education and Research (BMBF) through the research grant number 03FH046PX2.

References

[1] National Institute of Standards and Technology, NIST Definition of Cloud Computing, http://csrc.nist.gov/groups/SNS/cloud-computing/, [retrieved: Jan. 2014]

[2] G. Cattaneo, M. Kolding, D. Bradshaw, G. Folco, IDC for the European Commission, Quantitative Estimates of the Demand for Cloud Computing in Europe and the Likely Barriers To Take-up, 2012, http://cordis.europa.eu/fp7/ict/ssai/docs/study45-d2-interim-report.pdf, [retrieved: Sept. 2013]

[3] Distributed Management Task Force, Architecture for managing clouds. http://dmtf.org, [retrieved: June, 2014]

[4] ISO/IEC SC 38 Study Group, JTC SC 38 Study Group Report on Cloud Computing, International Organization for Standardization, Tech. Rep., 2011. http://isotc.iso.org, [retrirved: June, 2014]

[5] A. Keller and H. Ludwig, The WSLA Framework: Specifying and Monitoring Service Level Agreements for Web Services, Journal of Network and Systems Management, vol. 11, no. 1, pp. 5781, Mar. 2003

[6] J. Happe, W. Theilmann, A. Edmonds, and K. Kearney, "Service Level Agreements for Cloud Computing", Springer-Verlag, 2011, A Reference Architecture for Multi-Level SLA Management, ISBN 978-1-4614-1613-5

[7] S. Ferretti, V. Ghini, F. Panzieri, M. Pellegrini, and E. Turrini, QoS-Aware Clouds, in Proceedings of the 2010 IEEE 3rd International Conference on Cloud Computing, ser. CLOUD 10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 321328.

[8] [1] Gardner Group, "CIO-Prioritäten und Budgets", 2011, http://www.cio.de/strategien/analysen/2262709/[retrieved: Sept. 2013]

[9] Bourne, V., "Companies struggling with cloud performance", 2011, http://servicemanagement.cbronline.com/news/cloud-performance-issues-costing-firms-600000-a-year-survey, 2011, [retrieved: Nov. 2013]

[10] Experton Group, "Experton-Analyse: Wer f¨llt die Cloud mit Business Services",2012, http://www.computerwoche.de/management/cloud-computing/1934027/ ([retrieved: Sept. 2012]

[11] Teckelmann, R.; Reich, C.; Sulistio, A, "Mapping of Cloud Standards to the Taxonomy of Interoperability in IaaS," CloudCom, 2011 IEEE Third International Conference on Cloud Computing Technology and Science, pp.522,526, 2011

[12] Manoel, E., Nielsen, M.J., Salahshour, A., S.S., Sudarshanan, "Problem Determination Using Self-Managing Autonomic Technology", IBM RedBooks, June 2005

[13] Emeakaroha, V.C.; Brandic, I; Maurer, M.; Dustdar, S., "Low level Metrics to High level SLAs - LoM2HiS framework: Bridging the gap between monitored metrics and SLA parameters in cloud environments," High Performance Computing and Simulation (HPCS), 2010 International Conference on , vol., no., pp.48,54, June 28 2010-July 2 2010

[14] Frey S.; Lüthje C.; Reich C.; Maier M.; Zutavern T.; Streif M.; "SLA-Richtliniendokument für Cloud Dienstleistungen", 2013, http://www.wolke.hs-furtwangen.de/assets/files/ASLAMaaS-SLA-Richtliniendokument.pdf, [retrieved: Jun. 2014]

[15] Goudarzi, H.; Pedram, M., "Multi-dimensional SLA-Based Resource Allocation for Multi-tier Cloud Computing Systems," CLOUD 2011, IEEE International Conference on Cloud Computing, pp.324,331, 4-9 July 2011, doi: 10.1109/CLOUD.2011.106

[16] Iqbal, W.; Dailey, M.N.; Carrera, D., "SLA-Driven Dynamic Resource Management for Multi-tier Web Applications in a Cloud," Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference on , vol., no., pp.832,837, 17-20 May 2010, doi: 10.1109/CCGRID.2010.59

[17] DMTF, "Open Virtualization Format (OVF)" http://www.dmtf.org/standards/vman, (November 2007), [retrieved: March 2014]

[18] Apache Software Foundation, "Deltacloud", http://incubator.apache.org/deltacloud/, (September 2009), [retrieved: March 2014]

[19] Chrisment, I.; Couch, A.;Badonnel, R.; Waldburger, M.; Sperb Machado, G.; Stiller, B., "An SLA Support System for Cloud Computing", Managing the Dynamics of Networks and Services, 2011, ISBN 978-3-642-21484-4, pp- 53-56

[20] Casalicchio, E., Silvestri, L., "Architectures for Autonomic Service Management in Cloud-based Systems", ISCC 2011, IEEE Symposium on Computers and Communications (28 2011-july 1 2011) pp. 161 - 166

[21] Xiong, P., Chi, Y., Zhu, S., Moon, H.J., Pu, C., Hacigumus, H., "Intelligent Management of Virtualized Resources for Database Systems in Cloud Environment" ICDE 2011, IEEE 27th International Conference on Data Engineering (April 2011) pp. 87 - 98

[22] Mehdi, N.A., Mamat, A., Ibrahim, H., Subramaniam, S.K., "On the fly Negotiation for Urgent Service Level Agreement on Intercloud Environment", Journal of Computer Science. Volume 7. (2011) pp. 1596 -1604

[23] Zhiliang Zhu; Jing Bi; Haitao Yuan; Ying Chen, "SLA Based Dynamic Virtualized Resources Provisioning for Shared Cloud Data Centers" CLOUD 2011, IEEE International Conference on Cloud Computing, pp.630 -637, 4-9 July 2011, doi: 10.1109/CLOUD.2011.91

[24] Frey, S., Lüthje, C., Teckelmann, R.; Reich, C., "Adaptable Service Level Objective Agreement (A-SLO-A) for Cloud Services" CLOSER 2013, pp. 457-462, SciTePress. ISBN: 978-989-8565-52-5

[25] Frey, S., Lthje, C., Huwwa, V., Reich, C., "Fuzzy Controled QoS for Scalable Cloud Computing Services" CLOUD COMPUTING 2013, The Fourth International Conference on Cloud Computing, GRIDs, and Virtualization, pp. 150-155

An Ambient Assisted Living Platform as a Service Architecture for Context Aware Applications and Services

Hendrik Kuijs, Christoph Reich
Faculty of Computer Science
Furtwangen University of Applied Science
Furtwangen, Germany
Email: {Hendrik.Kuijs, Christoph.Reich}@hs-furtwangen.de

***Abstract*—The main goal of technology in the field of Ambient Assisted Living (AAL) is to support and assist people in their daily life. Especially for elderly people this opens up the possibility, that they are able to stay in their own homes and their known environments longer. In addition to that another focus is to introduce technological approaches to support the social inclusion for elderly people with reduced mobility in rural regions.**

This paper presents an overview of research projects for central management platforms in the field of AAL. Taking these into account we propose a platform architecture that builds the basis for various applications to assist elderly people and support the social inclusion. In our approach the user is seen as the central concept and services or applications are able to adapt to a person's needs. These changes are triggered by personal information, that is stored in a person centered ontology. It is able to store medical information, interests, habits, as well as personal information or contact information. This data and data of the user's environment is used to make intelligent decisions to provide adapted services.

The platform is realized as a Platform as a Service (PaaS) in the cloud. The setup is a Private Cloud, that shares central services and interfaces in the Public Cloud for better flexibility, scalability and maintainability.

***Index Terms*—PaaS, AAL, Cloud, OSGi, software agents, context aware**

I. INTRODUCTION

Because of the increasing average life expectancy and a decrease of birth-rate, the proportion of the young working population in Germany [1] and world-wide [2] is shrinking continuously. On the other hand, families are getting smaller and extended families that are able to care for their elderly relatives are more and more disappearing. Politics try to compete with this trend by introducing programs for new nursing or day-care facilities, but there are too few trained care assistants for the elderly or too little financial resources [3]. The only chance is to keep the needed time-span for professional care facilities at a minimum. This trend is supported by the target group as surveys show that elderly people want to stay at home as long as possible [4].

The field of Ambient Assisted Living (AAL) tries to address these demands by utilizing technology to connect everyday objects and the social environment to build up special services to support elderly people to stay independent in their known environment. These projects and services can be divided in four different areas of application [5]:

- Health and health care
  This area is focused on health prevention and functional rehabilitation at home. The applications range from remembering assistance systems for medication or exercise programs to emergency systems, that are triggered by sensor data or vitality and movement data of the user.
- Household and supply
  This includes the growing market of smart home products, that are able to communicate with other products or external services to deliver a richer service to the user [6]. Another trend is to re-think user interfaces for a better user experience, by using easily comprehensible displays or implementing help-dialogs to guide the user through complex tasks.
- Safety and privacy
  Applications in this area range from devices that are secured against accidental operation and presence detectors to alerting-functionality or automated emergency calls.
- Communication and social environment
  Technology is used to support social integration by providing easy to use interfaces to get connected to family members, neighbors or other social networks. This initial communication and social inclusion can lead to more mobility and a better access to cultural or leisure activities.

However the developed applications often lack the interoperability and try to solve just one area of concern. This leads to a highly fragmented market with system incompatibilities that are intransparent for the customer [7]. Therefore many projects try to combine existing solutions by delivering platform architectures for AAL that integrate existing services into one manageable system.

In Section II we present different projects that integrate existing smart home devices in single platforms, provide tools for developing new applications, combine the solutions of different projects into one and projects that try to migrate partial

data services or the whole platform to the cloud. In Section III we clarify which research outcomes are considered for the *Person Centered Environment for Information, Communication and Learning* (PCEICL) platform, which is presented with its main ideas in Section IV. Section V gives a conclusion and a further outlook of upcoming research topics.

II. State of the Art in AAL Architectures

Memon et al. [8] define the role of an architecture for AAL as follows:

> "An AAL solution is an integrated system-of-systems composed of systems, subsystems and components, providing a part of the overall AAL system and its services. The architecture defines the distribution and relationship among the AAL systems, subsystems and components."

Many of the currently developed solutions rely on integration in a networked environment connected to a middleware, that manages the different sensors, actuators and devices. The *European Ambient Assisted Living Innovation Alliance AALIANCE* [9] recommends using a modular and open approach for new services and software in the context of AAL. For remote management for software upgrades and installing new software to an existing platform, the *Open Service Gateway initiative (OSGi)*[10] is seen as one of the possible solutions for middleware. The central feature of OSGi is the possibility to include, update, start and stop applications or services as bundles during runtime without the need for a restart of the whole environment. Therefore it is adopted in different AAL architectures and many developers are delivering their applications or device-drivers as OSGi bundles.

A. Architectural approaches

The first step towards AAL architectures are smart homes with interconnected sensors, actuators, computers, and other devices in the environment. Due to the complexity of the first systems the main controller often was referred to as a black box and interconnecting it with other or new systems was not possible or feasible [11][12].

Different projects try to open up or standardize the functionality of the middleware in order to create a common basis for future AAL projects and to provide a centralized management platform and distribution platform.

The *Gator Tech Smart House* [12] tries to achieve this with a *Programmable Pervasive Space*, that can be extended with new emerging technologies. Devices in the physical layer are converted to software services by the platform, that can then be programmed or merged with other services to create complex applications.

The *SOPRANO (Service Oriented PRogrammable smArt enviroNments for Older Europeans)* project [13] developed an open middleware for AAL solutions with another object of research: The *SOPRANO Ambient Middleware (SAM)* enriches user commands or sensor data semantically and determines an adequate system response, that is then performed in the living environment by the connected actuators. The middleware was introduced together with guidelines to develop new services or integrate actuators and sensors [14]. These guidelines apply for different stake-holders that are keys to success for systems in the field of AAL: These are developers of actuators and sensors, providers of value-added services, solution developers and care providers, relatives or the users themselves.

ProSyst delivers a framework for eHealth scenarios based on OSGi to manage sensors and devices over different protocols [15] . The management software (*mBS Smart Home*) is installed at the user's home and able to be adopted to already installed sensors or actuators. It delivers applications for predefining home automation scenarios, notification of specific events and configuring interfaces for collecting data of installed sensors. On top of this it is possible to deploy domain specific applications for eHealth. ProSyst offers a backend (*mPower Remote Manager*) for remote management of services and applications: These can be installed, updated, started, stopped or deleted remotely through a Software Management service in the backend. The backend is installed off-site and connected via LAN or internet over a secured connection. Developers are able to use a *Software Development Kit* for new applications in the Smart Home environment. The *mBS Smart Home* and *mPower Remote Manager* are considered to be closed source.

The *AMIGO (Ambient Intelligence for the networked home environment)* project [16] developed an architecture that is based on a middleware, that operates across different application domains and across different homes and environments. Therefore they developed the *Amigo Community Sharing Services (CHESS)* that uses web-services to communicate or share time together with relatives via web-services. Most of the applications are web-based and can be accessed by any device with a web-browser. The main focus of the middleware lies on automatic device and service discovery.

MPOWER (Middleware platform for eMPOWERing cognitive disabled and elderly) created a middleware with a strong focus on rapid development of applications by implementing standards-based web services in the home domain [17]. The system supports the interoperability between profession and institution specific systems (e.g. Hospital Information System). It supports security and safe social and medical information management. and addresses the need of mobile users (e.g. professional caregivers) which often change context and tools.

The *OASIS (Open architecture for Accessible Services Integration and Standardization)* project [18] focuses on an ambient intelligent (AmI) framework with software-agents based upon a *OASIS hyper-ontology* as common language. The ontology is able to combine multiple ontologies in the same application domain or different domains. The *OASIS System* consists of the *AmI Framework*, based on software-agents, and a *Interaction Platform* with a user interface to combine new sensor data to new services and the ability to self-adapt to different devices.

B. Combining architectural approaches

Another attempt is to merge promising or already successful partial solutions with different technical requirements into one big and flexible framework.

The project *universAAL* [19] tries to build up one "Consolidated European AAL platform" [20] and integrates different software modules of other European research projects such as *Amigo* [21], *MPOWER* [22], *OASIS* [23], *SOPRANO* [13] and *PERSONA* [24] into one single AAL solution. This first release of universAAL consists of the *AAL Studio*, an integrated development environment with different Eclipse [25] plugins, the *Runtime Support Platform (RSP)*, *universAAL Control Centre (uCC)*, for managing the platform, and *universAAL store (uStore)*, to provide one single repository for new software and services. Its goal is to make it viable for developers to create new AAL services. Interested developers are therefore provided with extensible knowledge-bases, online courses, wikis and personal training sessions. The uStore is seen as central marketplace to distribute the developed applications based on universAAL.

C. Cloud-backed AAL environments

The presented AAL platforms mostly require an installation of the whole system in a users environment. The needed working power can mean high costs for the initial setup of such a system. Another possibility that is currently discussed is the use of cloud systems for special services that can be accessed via web-interfaces by different systems or institutions.

Kim et al. [26] present a platform approach to share health data in the cloud in a secure way. The system is built around *Microsoft HealthVault* [27] and *DACAR* [28]. The patient-centric solution provides strong security and privacy characteristics and is entirely governed by the patient. It allows sharing of health data between hospitals, trained care-personnel or relatives to indicate changes in the health conditions amongst different support groups of the user.

Ekonomou et al. [29] introduced a cloud-service for maintaining an installation of an AAL solution in a home environment. They developed an extensible OSGi-based architecture for highly heterogeneous smart home systems. This architecture is focused on the integration of new devices by using a cloud-based service for discovering drivers in a manual, semi-automatic and automatic way. The user interface for the auto-discovery is displayed on a smart phone for ease-of-use.

D. Cloud-based AAL environments

The project *CoCaMAAL (Cloud-oriented Context-aware Middleware in Ambient Assisted Living)* [30] tries to move the AAL platform in the cloud. Their focus is on the implementation of a service-oriented architecture (SOA) for unified context generation. Data of installed sensors and devices in the smart living environment is collected by a *Data Collector* on-site and transferred into the cloud. This data is combined by a *Context Aggregator* and interpreted based on classifications obtained by *Context Providers*. A *Context-aware Middleware* matches this context with services provided by a *Service Provider Cloud* and sends appropriate actions back to the *Data Collector* to activate actuators or devices. Besides the architectural description no further information or code is provided by the authors.

III. Reusing existing Platforms for PCEICL

The two most promising platforms for research on new approaches for ambient intelligent services in the field of AAL at the moment are *universAAL* and *CoCaMAAL*. *UniversAAL* delivers a platform that can be configured for developing applications, testing applications and testbed installations. The complexity of the platform itself and the still ongoing development makes it difficult to test and implement new mechanics in the architecture. Nevertheless there are aspects that have to be considered when building the system architecture, like the agent-based ambient intelligent framework, the uStore for providing new applications or OSGi as a common basis for developed plugins. Futhermore universAAL is based on local installations. *PCEICL* tries to build up an AAL PaaS (Platform as a Service) in the cloud. The project *CoCaMAAL* is presenting an architecture that is cloud-based as well, but above that no further information of the development state is provided. As the research is still in progress it will be possible, that aspects of this approach have to be taken into consideration as well.

IV. The PCEICL Platform

The project *Person Centered Environment for Information, Communication and Learning (PCEICL)* introduces adaptation of functionality and presentation of information based on the medical state and the physical environment of the user.

The PCEICL ontology [31] is implemented to store and retrieve information about the user (e.g. medical information, interests or habits). Information about the environment is provided by attached sensors and external web-services. Based on this input information the platform is able to make intelligent decisions by integrating a software agent platform, in this case JADE (Java Agent Development Framework) [32]. These decisions are submitted to applications that use the information to adapt the functionality.

An example is a user with mobility impairment, that lives in a rural region. The platform provides a calendar of events for his region through a public *Events Service*. He wants to attend an event and the system tries to find a nearby neighbor, who will fetch him up for this event. The system automatically schedules a reminder prior to the event to get ready for the pick-up. Due to an accident the user is not able to attend the event. The system registers the change in health-state and cancels the appointment. This scenario uses external services, like a calendar of events provided by the municipality, a route planner to estimate the pick-up time prior to the event and an external search platform for the search for a lift as well as health-information of the user provided by the ontology.

Figure 1 shows the proposed architecture of PCEICL. It is based on the OSGi platform with several OSGi bundles for common services, like *Sensor Bundles*, *Smart Home Control*

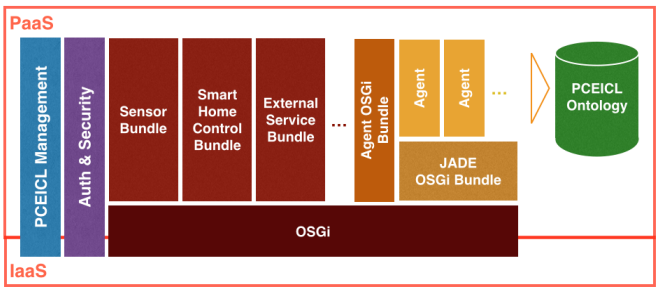


Fig. 1. Overview of PCEICL Architecture

Bundles, an *Address Book Bundle*, *External Services Bundle* or a *Web-Interface Bundle*. The *JADE-OSGi Bundle* [33] is hosting the software-agents system but can also register and communicate with software-agents, that are OSGi bundles themselves. The *PCEICL ontology* is only accessed by agents, that are evaluating and aggregating data and pass the relevant context to the service bundles. The platform features an *Authentication and Security Module*, that is also securing the OSGi framework. For example, a *Message Access Control Module* controls the messages flow between the various OSGi bundles in detail.

Special OSGi bundles provide services for installing new bundles and functionality or for updating existing ones during runtime through a central *Bundle Repository* based on OBR [34].

The context aware platform is realized as a Platform as a Service (PaaS) in a Private Cloud, that shares central services and interfaces in the Public Cloud for better flexibility, scalability and maintainability. Each environment is separated from other environments for privacy reasons. For preconfiguration and customization it provides a PCEICL management system to add new services on demand or adjust the configuration of the platform to the user's needs or different environmental settings.

The PCEICL system is able to access central services in the Public Cloud like a Cloud Management System, that has the ability to assign resources based on the workload of the PaaS instance, or public services, as the Events Service described in the example above.

V. Conclusion

In this paper we presented an overview of different projects for AAL architectures and the different research topics. Although the goal of AAL is understood the same way across all the projects, emerging technologies allow for new research approaches.

Several ideas of the described projects were considered while designing the PCEICL platform. With the PCEICL platform we focus on adaptability of applications based on software-agents, a user-centric ontology and environment information. The delivery of these services in the Private Cloud as a PaaS system with central services in the Public Cloud is another main research topic. The presented architecture gives a short introduction of the proposed PCEICL platform.

In this early stage there still remain uncovered aspects, like a definition of which services are running in the Private Cloud or the Public Cloud, a specification for interfaces between cloud services and their interfaces.

When data is transferred to services in the cloud, security and privacy constraints have to be taken into consideration as well.

Acknowledgment

The project ZAFH-AAL ("Zentrum für Angewandte Forschung an Hochschulen für Ambient Assisted Living") is funded by the Ministry of Science, Research and the Arts of Baden-Württemberg, Germany. The funding program for the universities of applied science is called: Zukunftsoffensive IV "Innovation und Exzellenz" (ZO IV). The PCEICL project is a sub-project of the project ZAFH-AAL [35].

References

[1] Statistisches Bundesamt, "Prognose der Bevölkerungsentwicklung in Deutschland nach Altersgruppen im Zeitraum der Jahre von 2007 bis 2050," http://http://de.statista.com/statistik/daten/studie/248090/umfrage/entwicklung-der-bevoelkerungsstruktur-deutschlands-nach-altersgruppen/, [retrieved: 2014.09.16].

[2] United Nations, "World Population Ageing: 1950-2050," UN: Department of Economics and Social Affairs - Population Division, http://www.un.org/esa/population/publications/woldageing19502050/, Report, 2001.

[3] tagesschau.de, "Gröhe will 30 Prozent mehr Altenpfleger," http://www.tagesschau.de/inland/pflegegroehe100.html, January 2014 [retrieved: 2014.09.16].

[4] J. Grauel and A. Spellerberg, "Akzeptanz neuer Wohntechniken für ein selbstständiges Leben im Alter," in *Zeitschrift für Sozialreform*, June 2007, vol. Heft 2 Jg. 53, pp. 191–215.

[5] P. Georgieff, *Ambient Assisted Living - Marktpotenziale IT-unterstützter Pflege für ein selbstbestimmtes Altern*. MFG Stiftung Baden-Württemberg, October 2008.

[6] Miele & Cie. KG, "Miele@home - Intelligente Vernetzung für Zuhause," http://www.miele.de/haushalt/hausgeraetevernetzung-1912.htm, [retrieved: 2014.08.03].

[7] R. Savage, Y. Yon, M. Campo, A. Wilson, R. Kahlon, and A. Sixsmith, "Market Potential for Ambient Assisted Living Technology: The Case of Canada," in *Ambient Assistive Health and Wellness Management in the Heart of the City*, ser. Lecture Notes in Computer Science, M. Mokhtari, I. Khalil, J. Bauchet, D. Zhang, and C. Nugent, Eds. Springer Berlin Heidelberg, 2009, vol. 5597, pp. 57–65.

[8] M. Memon, S. R. Wagner, C. F. Pedersen, F. H. A. Beevi, and F. O. Hansen, "Ambient Assisted Living Healthcare Frameworks, Platforms, Standards, and Quality Attributes," in *Sensors*. MDPI, Basel, Switzerland, March 2014, no. 14, pp. 4312–4341.

[9] AALIANCE, *Ambient Assisted Living Roadmap - AALIANCE Project - Deliverable 2.7*, March 2010, ch. Enabling Technologies, pp. 95–96.

[10] "The OSGi Architecture," http://www.osgi.org/Technology/WhatIsOSGi, [retrieved: 2014.09.17].

[11] S. Korff, "Pets in your home – how smart is that ?" *Symposium on Usable Privacy and Security*, 2013.

[12] S. Helal, W. Mann, H. El-Zabadani, J. King, Y. Kaddoura, and E. Jansen, "The gator tech smart house: A programmable pervasive space," *Computer*, pp. 64–74, March 2005.

[13] D. Balfanz, M. Klein, A. Schmidt, and M. Santi, "Partizipative Entwicklung einer Middleware für AAL-Lösungen: Anforderungen und Konzept am Beispiel SOPRANO," in *GMS Medizinische Informatik, Biometrie und Epidemiologie*, vol. 4(3), http://www.egms.de/static/de/journals/mibe/2008-4/mibe000078.shtml, October 2008.

[14] A. Schmidt, P. Wolf, M. Klein, and D. Balfanz, "SOPRANO Ambient Middleware: Eine offene, flexible und marktorientierte semantische Diensteplattform für Ambient Assisted Living," 2009.

[15] M. Petzold, K. Kersten, and V. Arnaudov, “OSGi-based E-Health / Assisted Living,” ProSyst, http://http://www.prosyst.com/fileadmin/ProSyst_Uploads/pdf_dateien/ProSyst_M2M_Healthcare_Whitepaper.pdf, Whitepaper, September 2013.

[16] M. D. Janse, “AMIGO - Ambient Intelligence for the networked home environment,” Final activity report, 2008.

[17] S. Walderhaug, E. Stav, and M. Mikalsen, “The MPOWER Tool Chain - Enabling Rapid Development of Standards-based and Interoperable Homecare Applications,” Tech. Rep., 2008.

[18] M. Panou and E. Bekiaris, “Oasis – open architecture for accessible services integration and standardization,” CERTH/HIT, Project Presentation, April 2008.

[19] R. Sadat, P. Koster, M. Mosmondor, D. Salvi, M. Girolami, V. Arnaudov, and P. Sala, “Part III: The universAAL Reference Architecture for AAL,” in *Universal Open Architecture and Platform for Ambient Assisted Living*, R. Sadat, Ed. SINTEF, November 2013.

[20] “universAAL - Project Description,” http://universaal.org/index.php/es/about/about-project-description, [retrieved: 2014.09.18].

[21] M. D. Janse, “AMIGO - Ambient Intelligence for the Networked Home Environment,” http://www.hitech-projects.com/euprojects/amigo/amigo.htm, [retrieved: 2014.09.18].

[22] M. Mikalsen, “MPOWER,” http://www.sintef.no/Projectweb/MPOWER/, September 2009 [retrieved: 2014.09.18].

[23] “OASIS - Open architecture for Accessible Services Integration and Standardisation,” http://www.oasis-project.eu/, [retrieved: 2014.09.18].

[24] Fraunhofer AAL, “PERSONA - PERceptive Spaces prOmoting iNdependent Aging within dynamic ad-hoc Device Ensembles,” http://www.aal.fraunhofer.de/projects/persona.html, [retrieved: 2014.09.18].

[25] Eclipse Foundation, “What is Eclipse?” http://help.eclipse.org/luna/index.jsp?topic=%2Forg.eclipse.platform.doc.isv%2Fguide%2Fint_eclipse.htm, [retrieved: 2014.09.18].

[26] J. E. Kim, G. Boulos, J. Yackovich, T. Barth, C. Beckel, and D. Mosse, “Seamless Integration of Heterogeneous Devices and Access Control in Smart Homes,” in *Eighth International Conference on Intelligent Environments*, 2012.

[27] Microsoft, “Microsoft healthvault,” https://www.healthvault.com, 2014 [retrieved: 2014.09.18].

[28] L. Fan, W. Buchanan, C. Thummler, O. Lo, A. Khedim, O. Uthmani, A. Lawson, and D. Bell, “DACAR Platform for eHealth Services Cloud,” in *Cloud Computing (CLOUD), 2011 IEEE International Conference on*, July 2011, pp. 219–226.

[29] E. Ekonomou, L. Fan, W. Buchanan, and C. Thüemmler, “An Integrated Cloud-based Healthcare Infrastructure,” in *Third IEEE International Conference on Cloud Computing Technology and Science*. IEEE Computer Society, 2011, pp. 532–536.

[30] A. Forkana, I. Khalil, and Z. Tari, “CoCaMAAL: A cloud-oriented context-aware middleware in ambient assisted living,” in *Future Generation Computer Systems*, G. Fortino and M. Pathan, Eds., vol. 35, 2014, pp. 114–127.

[31] C. Fredrich, H. Kuijs, and C. Reich, “An ontology for user profile modeling in the field of ambient assisted living,” in *SERVICE COMPUTATION 2014, The Sixth International Conferences on Advanced Service Computing*, A. Koschel and A. Zimmermann, Eds., vol. 5. IARIA, 2014, pp. 24–31.

[32] F. Bellifemine, G. Caire, G. Rimassa, A. Poggi, T. Trucco, E. Cortese, F. Quarta, G. Vitaglione, N. Lhuillier, and J. Picault, “Java agent development framework,” http://jade.tilab.com/, [retrieved: 2014.07.12].

[33] E. Quarantotto and G. Caire, “JADE OSGi GUIDE,” http://jade.tilab.com/doc/tutorials/JadeOsgiGuide.pdf, April 2010.

[34] W. J. Gédéon, *OSGi and Apache Felix 3.0*, 1st ed. Packt Publishing, November 2010, no. 978-1-84951-138-4, ch. Using the OSGi Bundle Repository.

[35] “ZAFH-AAL - Zentrum für angewandte Forschung an Hochschulen für Ambient Assisted Living,” http://www.zafh-aal.de, [retrieved: 2014.07.18].

Towards Smart Watch Position Estimation employing RSSI based Probability Maps

Stefan Knauth, Alfonso A. Badillo Ortega, Habiburrahman Dastageeri, Tommy Griese, Yentran Tran,
Faculty for Geomatics, Computer Science and Mathematics
HFT Stuttgart –
Stuttgart University of Applied Sciences
Schellingstr. 24, 70174 Stuttgart, Germany

***Abstract*—We investigate a RSSI based indoor positioning setup and algorithm which is based on probability maps, for estimating positions of RF equipped smart watches. For each measurement and receiver we calculate a probability function which indicates the probability for the mobile node to be at a certain position. For a given RSSI measurement, we do not represent the map in an analytical way (i. e. as formula) but by values stored in a 2-dimensional grid of numbers. The indexes of the array are related to spatial positions, one could speak as a "gray scale bitmap" representation of the probability function. The maps obtained for different reference points (receivers) are merged to obtain an overall probability map indicating the likelihood for the mobile node to be at a certain position. The map may be used for seeding in particle filter approaches. In this paper, we directly estimate the mobile nodes position by using the position of the highest likelihood value. Experiments were performed in a typical laboratory/office building environment. 868 MHz packets were transmitted at a rate of about 30 per second. In a first measurement setup an average positioning error of less than 1.5 m has been observed. Obtained results indicate that the accuracy of the method is comparable to fingerprinting, but besides a one-time calibration to resemble the internal path loss and antenna pattern of the reference node model, no calibration has to be performed.**

Keywords—RSSI localization; probability map; fingerprinting;

I. INTRODUCTION AND RELATED WORK

RSSI (Radio Signal Strength Indicator) based methods play an important role in the field of indoor positioning localization. WiFi access points are deployed in most buildings today and WiFi transceivers are available for example in smartphones and other personal devices. This has made WiFi RSSI based indoor positioning schemes the first choice for simple positioning setups, of course with restricted accuracy in the range of some meters. RSSI is also a much-used element in combined methods, for example as a base for dead reckoning methods employing for example inertial sensors (LMU) and allowing simultaneous localization and mapping (SLAM). RSSI is becoming even more attractive through the introduction of Bluetooth 4.0 (BLE, Bluetooth Smart Energy, "iBeacon"), since BLE will allow the cheap deployment of large amounts of battery-powered long lasting reference nodes. Current smartphones are already equipped with this technology.

There are three commonly used approaches to determine the position of a mobile node by means of reference nodes:

- Fingerprinting can reach accuracies in the range of 1..2 meters. In order to deploy fingerprinting, elaborate reference measurements have to be negotiated and updated on changes in the environment.
- Multilateration is more straightforward in the setup, besides determination of a suitable propagation coefficient no calibration has to be performed. Unfortunately, Multilateration is not convincing and in typical environments accuracies are well below fingerprint approaches.
- Proximity is typically not used to determine an accurate position but more to detect presence at certain points of interest. However it can also be used for positioning where accuracy is not so important. Commonly used for example in behavioral monitoring applications.

A detailed survey on representative existing systems can for example be found in the EvAAL series [1-3].

The probabilistic approach we present in the next section merges to some extend the concepts of proximity and multilateration to obtain accuracies comparable to those of fingerprinting algorithms but lacks the need to record fingerprintig maps. It is also related to methods used for example in radio tomography positioning schemes [4].

II. BACKGROUND

The relation between a transmitted radio signal and a received signal is described by the attenuation equation

$$P_R \propto P_T \frac{G_T G_R}{4\pi d^p} \qquad (1)$$

with the received power P_R, the transmitted power P_T, the transmitter and receiver antenna gains G_T and G_R, the distance d and the propagation exponent p [5]. Under free space conditions the propagation exponent p has the value of 2 describing undisturbed isotropic propagation. In building environments p has to be determined experimentally, with typical values between 4 and 6. RSSI algorithms actually operate on the logarithmic path loss L:

$$L = \log(P_T) - \log(P_R) + c_1 \qquad (2)$$

where L is given in the unit dBm, and c_1 is a constant describing antenna gains etc. Note that $\log(P_R)$ is the obtained RSSI value, as RSSI is also logarithmic and measured in dBm. Looking at equation (1) we can write

$$L = c_2 - p \log(d) \qquad (3)$$

where c_2 is a constant representing the transmission power and other constants from (1). In typical experiments, the path loss L is not directly observed. Instead the RSSI value (corresponding to P_R) is obtained in dBm. Putting (2) and (3) together we find

$$\log(P_R) = \mathrm{RSSI} = C - p \log(d) \qquad (4)$$

where C is again a constant representing c_1 and c_2, and is the RSSI value obtained at a distance of 1 meter from the transmitter. C is easily determined experimentally. Eq. (4) can be disssolved to give the distance as a function of the RSSI value:

$$d = 10^{(C - \mathrm{RSSI})/p} \qquad (5)$$

The distances d obtained between the mobile node and several reference nodes can be used as input to a multilateration based algorithm. Unfortunately the RSSI values and the obtained distance values are typically quite error prone, especially in indoor environments. The most prominent reasons for this deviations from this simple propagation model are (for example [6-8]):

- Attenuation of radiation by walls, furniture, or bodies.
- Random orientation of mobile antenna.
- Multipath propagation

In practice, for a given distance, the obtained RSSI value may be reasonably lower than predicted i. e. by (4). The main reasons are antenna misalignment, attenuation and multipath. Of course, the value may also be higher than predicted, for example due to constructive multipath interference or waveguide effects in corridors. Our algorithm respects this behavior by not converting RSSI values into a range, but assuming a probability distribution for the distance, based on the given RSSI value. In the next chapter, we outline how these ideas form the base for our algorithm.

III. PROBABILITY BASED ALGORITHM

Like in other approaches, path losses L_i between a mobile node and a set of n fixed nodes N_i at known positions R_i are obtained by RSSI measurement. For simplicity we start with only one fixed node N_1 and one path loss L_1. We work in 2 dimensions, i. e. $R_i = (x_i, y_i)$ and so on.

We define a probability density function $P(R,L_1)$, where R is any position in the reference frame and L_1 is the measured path loss between the mobile node and reference node 1. Fig. 1 (a) sketches a simple variant of P: here the probability is 1 for points R within a certain range d around the fixed node position R_1, and 0.3 for other positions. The figure represents this function as a grey scale image. We refer to this discrete representation of the probability density function as *probability map*. The actual range d (later called d_{RSSI}) in this simple case is given by the observed RSSI value according to (5).

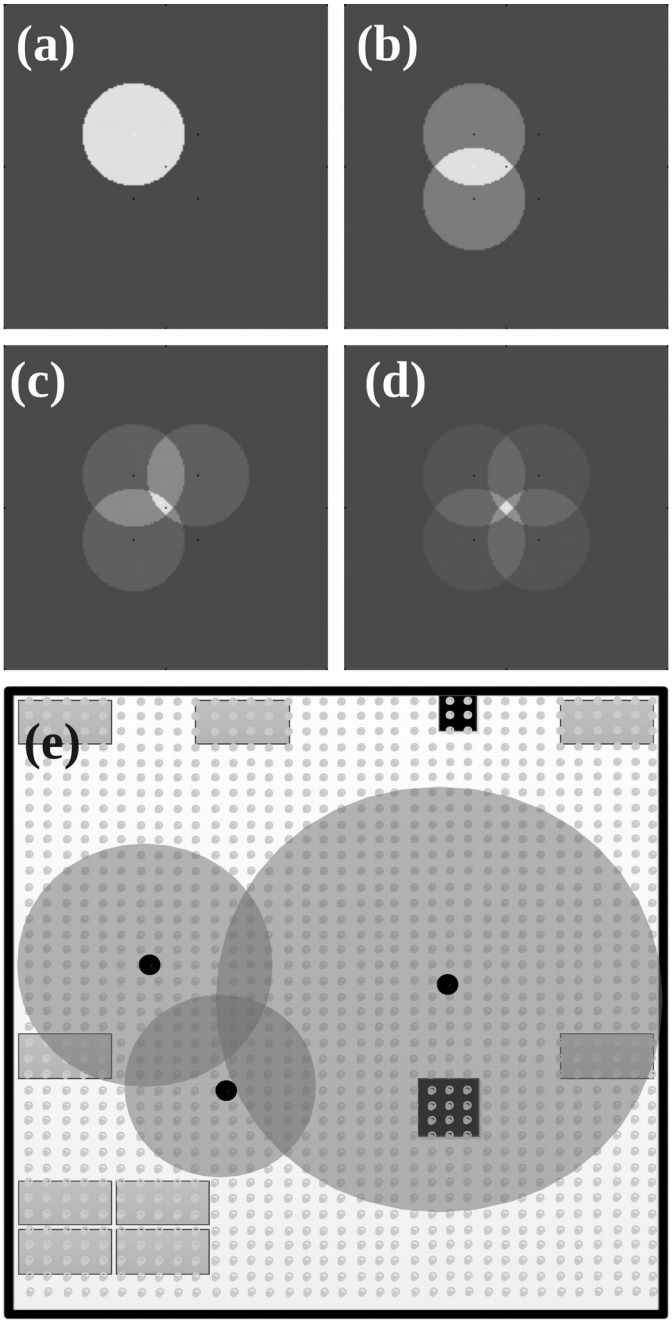


Fig. 1. (a) Probability map (representation of a probability density function) of a single receiver, for a RSSI value corresponding to the diameter of the circle. (b)..(d): merged maps for several receivers. (e) Merged map showing also the underlying calculation raster, and some orientation marks in the room. Overall size of (e) is about 6m x 6m. Circle areas with small radius correspond to high RSSI values / low path losses.

Already this easy probability density function respects some important properties of the indoor path loss, namely that the distance between fixed node and mobile node can be small, even if only low RSSI values are reported (high attenuation), but for a high RSSI value, it is unlikely that there is a large distance between transmitter and receiver.

For each receiver node, a respective map is created. Figures 1 (b-d) represent subsequent introduction of further fixed

nodes. The probability maps of the nodes are merged by pointwise multiplying thus creating an overall merged map. The merged map indicates the probability of presence of the mobile device for a given coordinate.

This map could for example be used to seed new particles in a particle filter. In the presented work we estimate the position of the mobile node by selecting only the areas of highest probability and calculating the center of gravity of these pixels. Figure 1 (e) displays again the merging of 3 probability functions, and indicates the discrete points which form the two-dimensional array of the probability maps.

In practice, more complex probability density functions are used. For example, the function may model the directional behavior of the fixed node antenna. A more smooth function will typically result in one absolute maximum value in the merged probability map. Results with such a function are discussed in the following chapter.

IV. Experiment

A. Setup

A general operation schema is shown in Fig. 2. Fig. 3 describes the actual setup in the Lab: Fixed receiver nodes were installed in a 6m x 6m sized room. Six CC1110DK MINI nodes were positioned on tables in the lab. Eight ground positions (Filled circles in Fig. 3) were defined and marked on the floor. For these positions, measurements were recorded. The mobile node consisted of a student wearing an EZ430-Chronos smart watch. The student visited the 8 marked positions. A second person recorded accurate timing information i.e. at what time which position was visited. The smart watch emits 868 MHz RF Packets of about 3 ms duration each, at a rate of 30 Hz. The 6 receivers collect their observed RSSI values and transmit them to a PC equipped with an 868 MHz USB dongle receiver (Fig. 2). A TDMA scheme is used to avoid collisions.

B. Calculation of probability map

For the real experiments we used a triangular function which is biased with a residual probability of 0,3 and has its maximum probability at the range determined by (5). This function $P(d_{RSSI}, d)$ delivers the probability for the mobile node to be at a distance d from a node, which recorded the signal strength RSSI. From this one-dimensional function a probability map is created by calculating for each point R of the map the value $P(d_{RSSI}, d)$, where d is the distance between R and the corresponding reference node location R_i. In Fig. 4 an example of such a mao is shown.

C. Results and discussion

TABLE I. Obtained Positioning Results

	real		calculated		error
point	x	y	x	y	
1	0,00	5,50	0,00	5,38	0,12
2	2,00	5,50	2,44	5,94	0,62
3	4,00	5,50	4,70	5,00	0,86
4	6,00	5,50	5,35	6,00	0,82
5	3,70	4,00	2,80	3,70	0,95
6	2,30	2,45	2,11	0,60	1,86
7	3,70	2,00	3,50	0,25	1,76
8	5,00	2,45	4,51	4,83	2,43

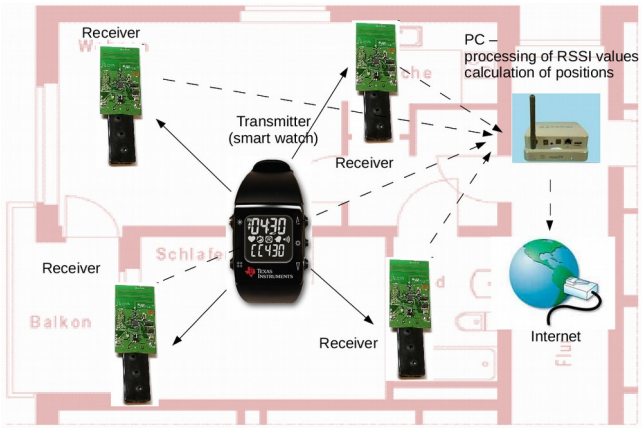


Fig. 2. Schematic setup used for the test measurements: An EZ430 Smartwach emits 868 MHz radio packets, which are received by CC1110DK MINI nodes (solid arrows). These fixed nodes transmit their obtained RSSI values to a central PC (dashed arrows), which logs the values and runs the algorithm. Note that in the experiment, 6 receiver nodes were used as shown in Fig. 3.

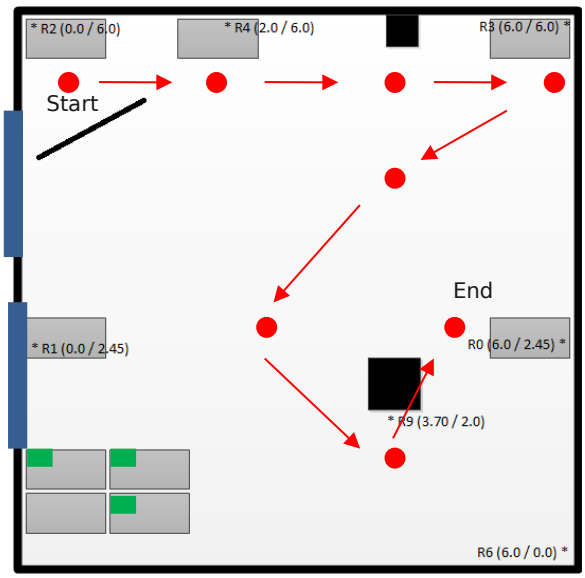


Fig. 3. Situation in the lab. Grey rectangles indicate tables, receiver positions are marked with asterisks. The test path is indicated with arrows, measurement positions are indicated by filled circles.

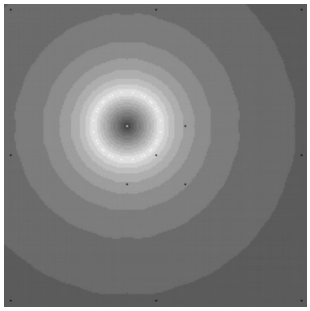


Fig. 4. Typical Probability map as used in the experiments, in a grey scale representation. Dark parts indicate low probability, bright parts indicate high probability. The size of the map is 10m x 10m, the d_{RSSI} value is 2m. The function is based on a triangular density function and geometric correction. For lower RSSI values, the radius of highest probability (ring with bright grey values) would be larger, and vice-versa.

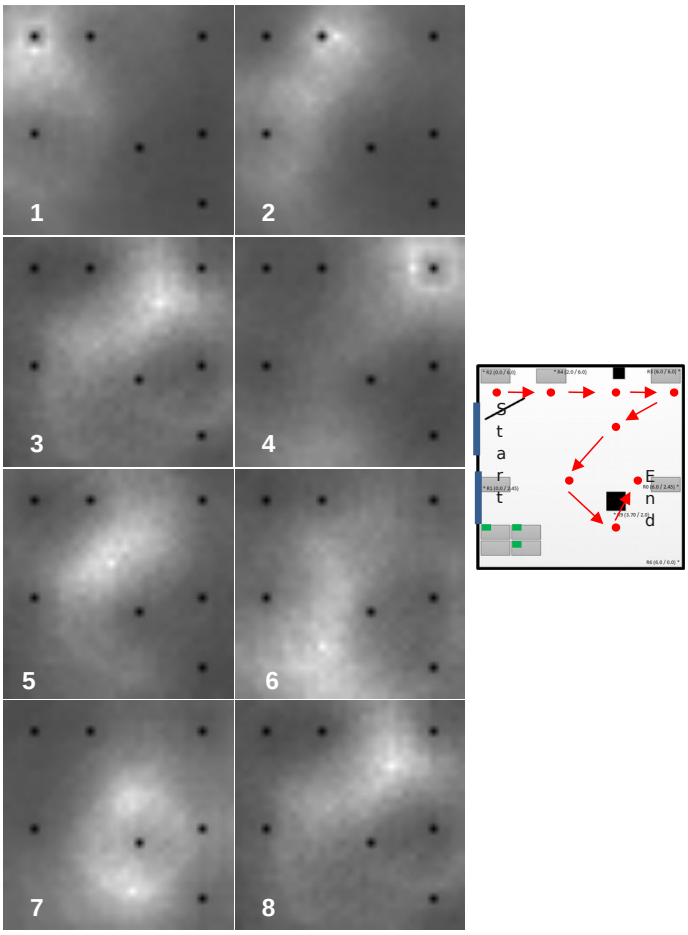


Fig. 5. Obtained merged maps for 8 measurement points. The map numbering corresponds to the sequence of the points in the path indicated in the left lower inset. The positions of the fixed nodes are indicated as dark points in the maps.

Table 1 lists the positioning results of the algorithm. Obtained errors lie in the range of some cm up to about 2.5 meters, with an average error of about 1.2 meters.

As it can be seen in Table 1, high errors were obtained for positions 6, 7 and 8 (see also Fig, 5). This might be attributed to the fact that these points lie more or less outside of the area spanned by the reference nodes which generally leads to less good algorithmic conditions for range related positioning methods.

V. Conclusion and Outlook

We consider this early work as successful proof of concept for the described method employing a RSSI dependent probability density function and merging the resulting probability maps. We already obtain accuracies comparable to those of fingerprinting methods. The method has the advantage that no fingerprint maps have to be generated, allowing for less engineering effort during deployment.

While we have taken the “brightest point”, i.e. the most likely position, for the reported position estimate, the generated maps allow also for more sophisticated position estimators, especially when using them as seeding information in particle filters.

Currently a new measurement campaign is negotiated on a larger area, and more accurate probability density functions are investigated on the recorded data. We see a reasonable chance to develop the method to an extent where it can produce reliable results under different deployment scenarios which are at least comparable to those of fingerprinting methods.

Acknowledgment

This work was funded in the frame of the German federal Ministry of Education and Research programme “FHprofUnt2013” under contract 03FH035PB3 (Project SPIRIT)

References

[1] Chessa, S., Knauth, S. (eds.) (2012); Evaluating AAL Systems Through Competitive Benchmarking; EvAAL 2011; CCIS vol. 309, Springer Berlin Heidelberg ISBN: 978-3-642-33532-7

[2] Chessa, S., Knauth, S. (eds.) (2013) Evaluating AAL Systems Through Competitive Benchmarking; EvAAL 2012; CCIS vol 362, Springer Berlin Heidelberg, ISBN: 978-3-642-37418-0

[3] J. A. Botía ,J. A. Álvarez-García, K. Fujinami, P. Barsocchi, T. Riedel (Eds.) (2013) Evaluating AAL Systems Through Competitive Benchmarking; EvAAL 2013, CCIS vol 386, Springer Berlin Heidelberg, ISBN: 978-3-642-41042-0

[4] Wilson, J., Patwari, N. (2010); Radio Tomographic Imaging With Wireless Networks; IEEE Transactions on Mobile Computing 9(5), 621–632

[5] Mautz, R. (2012): Indoor Positioning Technologies, SVH, ISBN 978-3-8381-3537-3, no. 3754, 136p.

[6] Bultitude, R.J. (1987): Measurement, characterization, and modeling of indoor 800/900 MHz radio channels for digital communications. IEEE Communications 25(6), 512

[7] Hashemi, H. (1994): A Study of Temporal and Spatial Variations of the Indoor Radio Propagation Channel. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 127–134

[8] Rappaport, T. (2001): Wireless Communications: Principles and Practice, 2nd edn. Prentice Hall PTR, Upper Saddle River

SmartMetering

Frederik Laasch, Philipp Klein and Dirk Benyoucef
Furtwangen University, Robert-Gerwig-Platz 1, 78120 Furtwangen, Germany
Email: {laaschfr, p.klein, dirk.benyoucef}@hs-furtwangen.de

***Abstract*—SmartMetering is a project of the Signal Processing Research group of Furtwangen University. The project is concerned with the research in Non-Intrusive-Load-Monitoring (NILM) systems. The key idea in this paper is that different appliances have different distinct properties. In order to increase the accuracy of a disaggregation, those properties are exploited by using different disaggregation algorithms.**
The paper summarizes the development of the NILM system. Information is provided about the event detection mechanism, the classification algorithm, the energy tracking and about the testbench used for data acquisition.

I. Introduction

Today, one of the most pressing topics is energy consumption and its reduction. Monitoring the energy consumption of each appliance is the first step to consumption reduction in a facility. The disaggregation of the measured energy consumption and the matching of the parts with its appliances is the second step.
A tool for monitoring the consumed energy is the smart meter. A smart meter provides digital information on the total energy consumption of a facility. Non-Intrusive Load Monitoring (NILM) allows the disaggregation of the consumed energy and a matching of the disaggregated energy to the appliances. The term non-intrusive refers to the fact that next to the smart meter no additional hardware is needed. The disaggregation and matching of the parts to the appliances is done by using algorithms.

The structure of a NILM system is shown in Fig. 1. Based on the measurements the event detection determines the point in a time dependent signal where a transient occurs. Between two transients there is a steady-state. In a NILM system the time of occurrence of a transient and the duration of a steady-state are of interest. Being aware of the timing it is possible to perform a classification. The classification determines the probable parts of the aggregated signal. Knowing the energy consumption and its parts, it is possible to perform the energy tracking: To determine the energy consumption of each appliance.

The SmartMetering project of Furtwangen University has three stages: preparatory work, analysis and algorithm development as defined in [3].
As part of the preparatory work a systematic collection and examination of appliances is carried out. A database containing profiles of different appliances is built. During the second stage the profiles of the appliances are analysed. It is shown that the appliances show different types and levels

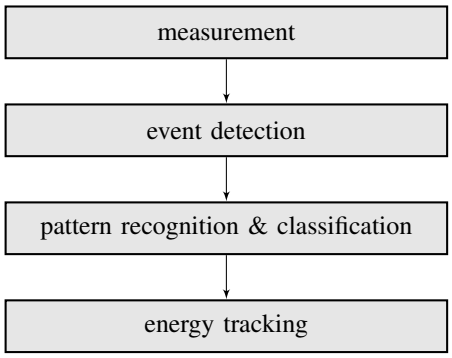


Fig. 1. Linearly structured disaggregation

of characteristics. For the classifier development the idea is to combine different classifiers instead of using one classifier for all appliances. Each classifier addresses an appliance specific characteristic. Two different classifiers are used. The first classifier is based on time-series power data fingerprints. The second classifier is based on Artificial Neural Networks (ANN). The ANNs are used to detect appliances with a periodic switching behaviour.
In the third stage first tests are carried out using ANNs [4]. The ANN is embedded in a linearly structured NILM system. The performance achieved with the neural network is compared to the performance of the algorithm Hart [5] developed.
Currently the first stage is completed, the second stage is almost completed and work on the third stage has begun.

II. State of art

NILM can be divided into two parts. The first part deals with the classification of appliances using their steady-states. Only the parts not containing switching events are analyzed. The first work on this topic was published by George Hart. In the 1980s he recorded real and reactive power in intervals of one second [5]. In order to classify loads a comparison of the power differences occurring when appliances were turned on or off was used. The detected events were classified using a predefined database as reference. Harts work was carried on by several other researchers, e.g. Baranski [6], Murata [7], [8] and Nakano [9].
In contrast to the method described before the transient state analysis tries to classify appliances by investigating switching events. Their characteristics can be found in the short time periods right after the turning on of an appliance. Works in

this area were carried out by Shaw [10], Cox et al. [11], Lee [12] and Laughman [13].

III. Measurements & analysis

Measurements were done during the first stage of the project. The measurements were motivated by the fact that a comparison of the proposed algorithms is difficult and requires a lot of manual tuning effort. Since it is hard to get reliable ground truth data associated with an algorithm. Reliable ground truth data must contain a list of all true switching events in order to compare the results achieved with an algorithm [14].
For the generation of a dataset an integrated control and measuring system is used in order to keep the effort for generating the ground truth data small. The system [14] generates switching events for connected appliances. For each appliance, the generated events are synchronously recorded together with the consumed energy. In parallel other research groups started to work on the same problem generating various datasets: [15], [16], [17], [18], [19], [20].

The idea of the system is to measure the energy consumption of appliances while recording their switching cycles. The first time, the system was described in [14]. It consists of a Measurement Box (MB), a Switching/Detection Box (SDB) and a Data Acquisition Card (DAC) and a PC. The block diagram of the test bench is shown in Fig. 2. The MB is

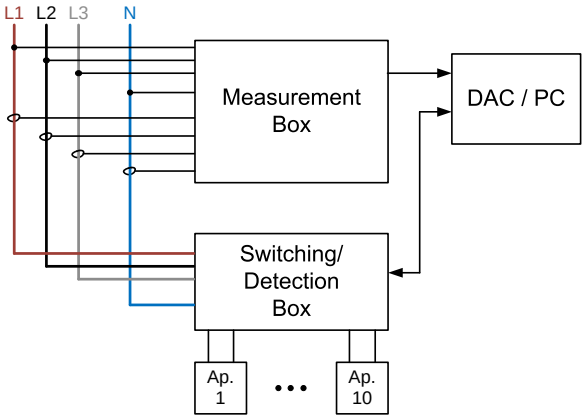


Fig. 2. Three phase test bench system

used to measure voltage and current, the SDB controls the appliances and the DAC is responsible for data conversion and storage. The test bench can be used in two different ways. First it can generate switching events in order to actuate appliances. Second, if the switching of the appliances is performed by humans or automatically by the appliances than the test bench can detect the switching events with the internal current sensors of the SDB.

The first set of measurements was collected under laboratory conditions. It consists of the following appliances: water heaters, freezers, hand held mixers, hair-dryers, incandescent lamps, a heater oven, energy saving lamps, micro ovens, TFT monitors and televisions. In the course of the project further measurements were performed and more appliances were included.

In the following the power signals of selected appliances right after the turn on event are shown. The figures show the turn on event of different monitors, micro-ovens and of refrigerators. The power signals are analyzed for characteristics which can be used for the disaggregation.
Monitors of different manufactures, as shown in Fig. 3, show only little similarity. The differences are mainly due to the large number of different switching power supplies available. In Fig. 4 the turn on behavior of micro ovens is shown. The

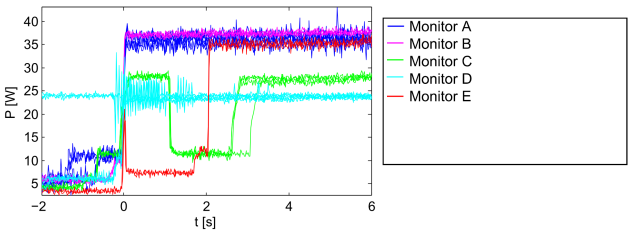


Fig. 3. Turn on behavior of monitors

different micro-ovens show similar patterns although they were manufactured by different companies. Contrary to the micro-ovens the power signal of the refrigerators in Fig. 5 is different in dimensions and amplitudes from model to model. Fig. 6 shows the power consumption of a household. As can be seen in the figure, between 8 pm and 6 am there are several switching events. Those switching events are caused by a refrigerator. The power signal of the refrigerator is periodic over long periods of time.

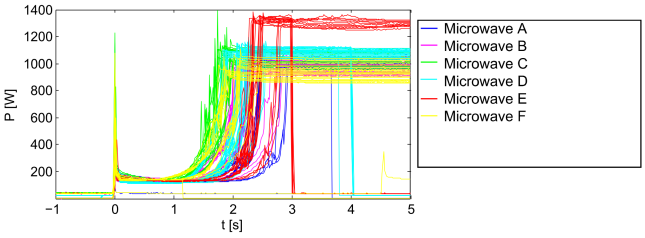


Fig. 4. Turn on behavior of micro-ovens

The analysis of the collected data aims at the improvement of the performance of the disaggregation. In order to improve the performance of the disaggregation the different characteristics of different appliances are taken into account choosing the classification algorithm.
Two different characteristics are distinguished: There are appliances whose time dependent power signal proves to be periodic, i.e. the refrigerator. The other kind of appliances have

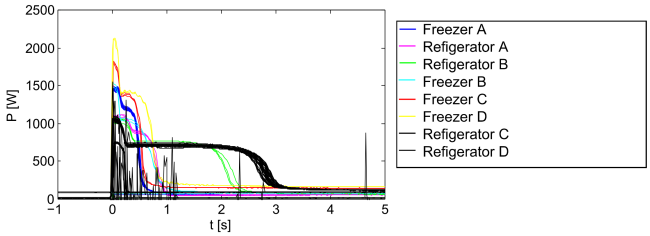


Fig. 5. Turn on behavior of a refrigerator

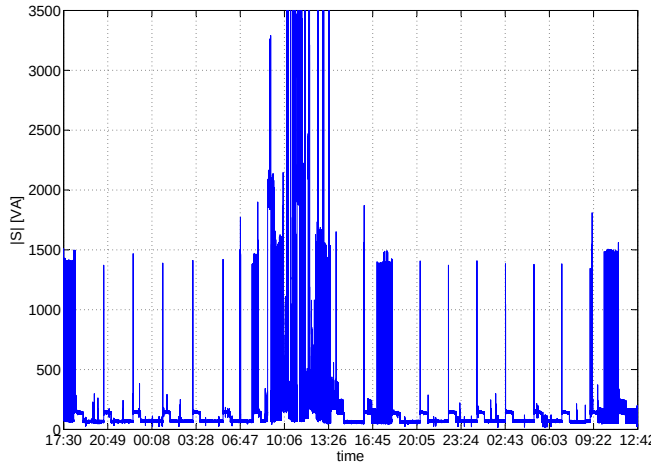


Fig. 6. Switching behavior of a refrigerator

power signals whose pattern is similar despite the fact that the appliances are manufactured by different companies. In this paper those appliances are the micro-ovens.

IV. Event Detection

Event detection algorithms are used to find the start point of the operation of an appliance, i.e. switching ON events. Referring to one dimensional signals the event detection is called *Step Detection*. Examples for step detection algorithms are [21], [22], [23], [24], [25]. The following paragraphs provide an overview over the event detection method used in a paper written by Thomas Bier [14].

There are two kinds of events:

- Switching ON Events
- Switching OFF Events

Fig. 7 shows a measurement of the instantaneous power. It's a simple pattern of different appliances. The pattern is composed of the power consumption of a refrigerator, a waterheater, a toaster, a hairdryer and lamps. Time dependent changes in power, as shown in Fig. 7, are used to detect the events:

- $\Delta P = P(t_2) - P(t_2 - \Delta T)$
- $\Delta T = t_2 - t_1$

If ΔP is greater than a predefined threshold an ON event is detected, as defined in (2). The second parameter is the time ΔT. It is taken into account since it is assumed that a switching

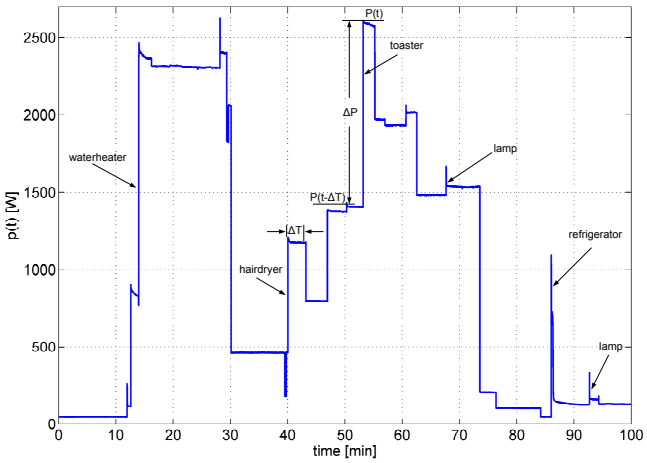


Fig. 7. Example for the instantaneous power of different appliances

ON event is always followed by a steady-state.

$$SwONEvent \hat{=} \Delta P \geq Threshold \quad (1)$$

$$SwONEvent \hat{=} P(t) - P(t - \Delta T) \geq Threshold \quad (2)$$

The switching OFF events are calculated analogously to the switching ON events.

$$SwOFFEvent \hat{=} \Delta P \leq Threshold \quad (3)$$

$$SwOFFEvent \hat{=} P(t) - P(t - \Delta T) \leq Threshold \quad (4)$$

The major drawback of the approach is the fixed threshold. An appliance might have a power consumption lower than the threshold. The appliance with the power consumption lower than the threshold cannot be detected. Therefore event detection algorithms with a variable threshold are currently investigated.

V. Pattern Recognition & Classification

Based on the detected events the classification algorithm finds the best combination of known using Hartsappliances for a given power signal. Three different approaches are examined: Classification using Harts approach [5] is compared to the ANN developed by Thomas Bier [4]. Finally the fingerprinting method developed by Philipp Klein [26] is evaluated. Each classification method is used to address different characteristics of the appliances.
The ANN is used for the classification of refrigerators. It is trained using the overnight periodic switching events. As pointed out in Chapter III, the switching events of different refrigerators vary in dimensions and amplitude and therefore are not suited for classification by fingerprints. Instead the fingerprints are used to detected micro-ovens, since the power signals of micro-ovens show a similar pattern.

The results achieved with the ANN are compared with Hart's method [5]. The ANN is trained using the apparent power of the appliance and a LMS algorithm for the computation of the weights of the neurons of the ANN. The transferfunction of the neuron is a step function [4].

In Table I, the performance of Hart's classifier is compared to the classifier based on the ANN. Altogether 1235 switching ON events of a refrigerator are taken into account. The classification algorithm based on Hart's approach has classified 1080 (87.5 %) of the events correctly. The algorithm based on the ANN approach has classified 1162 (94.1 %) of the events correctly. Considering the number of correct classifications,

Measurements of appliances	Approach based on	
	Hart	ANN
1235	Correctly classified	
	1080 (87.5%)	1162 (94.1%)
	Classification Errors	
	262	13

TABLE I
RESULTS OF THE CLASSIFICATION

the approach based on an ANN outperforms Hart's approach by 6.6 %. Both approaches have also classification errors. Hart's approach has 262 classification errors, caused by the presence of appliances that have a similar power consumption as the refrigerator. The ANN approach has 13 missclassifications. Now, considering the missclassifications, the ANN is also outperforming Hart's approach.

The fingerprints are generated using the characteristics of time series of the micro-ovens. In order to define a fingerprint time dependent signals are analyzed. The signals of different appliances are searched for similarities. By exploiting the similarities it is possible to define classes of appliances with similar time dependent behavior [26].
Figure 4 shows the power signals of different microwave ovens during the first few seconds right after the turn-on event. The common pattern off all the devices are used to generate the finger print. A fingerprint is described by dividing the signal into segments. The segments are mathematical described by functions with independent parameters which are fitted to each of the segments.
For example the fingerprint of a class describing micro-oven is defined by three parts and each part is described by a different function:

$$
\begin{aligned}
P_1(t) &= P_{01} && && t_0 \le t < t_1 \\
P_2(t) &= P_{02} && P_{02} < P_{01} && t_1 \le t < t_2 \\
P_3(t) &= P_{03} + E_3 \cdot e^{-\frac{t-t_2}{\tau_3}} && \tau_3 < 0 && t_2 \le t < t_3
\end{aligned}
$$

$P_1(t)$, $P_2(t)$ and $P_3(t)$ are power signals each belonging to a different segment. The very short turn-on impulse is approximated by a constant power P_{01}. The second part can be modeled as a constant power P_{02}. In the third part τ_3 is the time constant of the power increase. The exponential term is scaled by E_3. Both parameters show only small fluctuations for different appliances.
The parametrization of the models is done using training data and an iterative nonlinear fitting method based on the Levenberg-Marquardt algorithm.

The application of the algorithm is shown on the basis of a simulation. The signals of three appliances were superposed as listed in Table II. The power signal and the fingerprints are shown in Fig. 8. The turn-on events 1 and 2

TABLE II
PATTERN OF THE SIMULATION

No.	Turn-on time	Appliance
1	6.25 s	LG intellowave (microwave oven)
2	25.00 s	Medion MD12801 (microwave oven)
3	56.25 s	Bosch FD8705 (water heater)
4	56.50 s	LG intellowave (microwave oven)

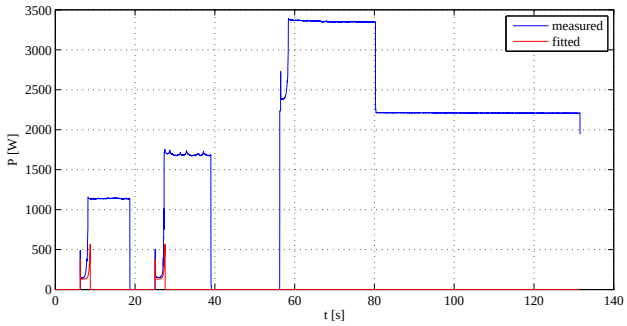


Fig. 8. Results of the appliance classification using characteristic signals

can be classified by the same fingerprint model although they are caused by different types of micro-ovens. If the fingerprint (No. 4) is superposed by an unknown power level (the power of the water heater, No. 3) it cannot be correctly identified.
The shown method is suitable for detection of some distinct appliances. A detection of all possible load profiles is unrealistic as shown before, since some devices of the same kind but manufactured by different companies show too little similarity. Nevertheless, this method can be used as a complement to other methods. Further analysis have to be made in order to estimate the potential of this method.

VI. ENERGY TRACKING

Energy tracking is a topic of ongoing research. Currently the problem is that it is possible to detect the switching on events but it is a problem to detect the switching off events. Both events are needed in order to calculate the on-time of an appliance and therefore the power consumption of the appliance.
Furthermore the currently used linear structure, as shown in figure 1, of the NILM system has a disadvantage. An erroneous measurement or event detection reduces the chance for a

successful disaggregation and matching of the parts of the total energy consumption to the appliances. So it becomes necessary to provide some reliability information together with the results of the functional blocks of figure 1.

VII. Conclusion & Future Work

In this paper the SmartMetering project of Furtwangen University was introduced. It is a system that uses non-intrusive load monitoring for energy consumption disaggregation. In the course of the project a measurement system was developed. The system is used to acquire data for the performance analysis of NILM systems. Special attention was payed to the availability of information about the switching events. Only by knowing the timing of the switching events it is possible to evaluate the performance of the event detection algorithms.

An event detection mechanism and two different approaches for the classification of events were introduced. First artificial neural networks were used for the classification of autonomous appliances with only two switching states. The results achieved were compared to Hart's [5] approach. For appliances with more than two switching states fingerprints were used. The use of different classifiers for the disaggregation of the energy consumption is believed to improve the classification accuracy. Therefore further possibilities to combine different classifiers and to thereby enhance performance are researched.

References

[1] A. Zoha, A. Gluhak, M. Imran, and S. Rajasegarar, "Non-intrusive load monitoring approaches for disaggregated energy sensing: A sur vey," *Sensors*, vol. 12, no. 12, pp. 16 838–16 866, Dec. 2012. [Online]. Available: http://www.mdpi.com/1424-8220/12/12/16838/

[2] Y. F. Wong, Y. Ahmet Sekercioglu, T. Drummond, and V. S. Wong, "Recent approaches to non-intrusive load monitoring techniques in residential settings," in *Computational Intelligence Applications In Smart Grid (CIASG), 2013 IEEE Symposium on*. IEEE, 2013, pp. 73–79.

[3] D. Benyoucef, P. Klein, and T. Bier, "Smart meter with non-intrusive load monitoring for use in smart homes," in *Energy Conference and Exhibition (EnergyCon), 2010 IEEE International*, 2010, pp. 96–101.

[4] T. Bier, D. Abdeslam, J. Merckle, and D. Benyoucef, "Smart meter systems detection & classification using artificial neural networks," in *IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society*, 2012, pp. 3324–3329.

[5] G. Hart, "Non-intrusive Appliance Load Monitoring," in *JRC Technical Report*. Proceedings of the IEEE: IEEE, December 1992, pp. 1870–1891.

[6] M. Baranski, "Energie-Monitoring im privaten Haushalt," in *Ph.D. Thesis*, University of Paderborn, 2006.

[7] H. Murata and T. Onoda, "Applying support vector machines and boosting to a non-intrusive monitoring system for household electric appliances with inverters," 2000.

[8] ——, "Estimation of power consumption for household electric appliances." 9th International Conference on Neural Information Processing, 2002.

[9] Y. Nakano, "Non-Intrusive Load Monitoring System – Part 5: Performance Test at Real Households," in *Technical Report*. System Engineering Research Laboratory, 2005.

[10] S. Shaw, "System identification and modeling for nonintrusive load diagnostics," in *Ph.D. Thesis*. Massachusetts Institute of Technology, 2000.

[11] R. Cox, S. Leeb, S. Shaw, and L. Norford, "Transient event detection for nonintrusive load monitoring and demand side management using voltage distortion." in *Applied Power Electronics Conference and Exposition*. APEC 06. Twenty-First Annual: IEEE, 2006.

[12] K. Lee, L. Norford, and S. Leeb, "Development of a functioning centrally located electrical-load monitor," May 2003, technical Report, Massachusetts Institute of Technology.

[13] C. Laughman, K. Lee, R. Cox, S. Shaw, S. Leeb, L. Norford, and P. Armstrong, "Power signature analysis," in *Power and Energy Magazine*. IEEE, Mar-Apr 2003, pp. 56–63.

[14] T. Bier, D. Benyoucef, D. Abdeslam, J. Merckle, and P. Klein, "Smart meter systems measurements for the verification of the detection & classification algorithms," in *IECON 2013 - 39th Annual Conference on IEEE Industrial Electronics Society*.

[15] J. Z. Kolter and M. J. Johnson, "REDD: A public data set for energy disaggregation research." San Diego, CA, USA: Massachusetts Institute of Technology, 2011.

[16] K. Anderson, A. Ocneanu, D. Benitez, A. Rowe, and M. Berges, "BLUED: A fully labeled public dataset for event-based non-intrusive load monitoring research," in *Proceedings of the 2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD)*, Beijing, China, 2012.

[17] A. Reinhardt, P. Baumann, D. Burgstahler, M. Hollick, H. Chonov, M. Werner, and R. Steinmetz, "On the accuracy of appliance identification based on distributed load metering data," in *Proceedings of the 2nd IFIP Conference on Sustainable Internet and ICT for Sustainability (SustainIT)*, 2012, p. 1–9.

[18] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, and J. Albrecht, "Smart*: An open data set and tools for enabling research in sustainable homes," in *Proceedings of the 2012 Workshop on Data Mining Applications in Sustainability (SustKDD 2012)*, Beijing, China, Aug. 2012.

[19] S. Makonin, F. Popowich, L. Bartram, B. Gill, and I. Bajic, "AMPds: A public dataset for load disaggregation and eco-feedback research," in *2013 IEEE Electrical Power Energy Conference (EPEC)*, Aug. 2013, pp. 1–6.

[20] J. Kelly and W. Knottenbelt, "'UK-DALE': A dataset recording UK domestic appliance-level electricity demand and whole-house demand," *arXiv:1404.0284 [cs]*, Apr. 2014, arXiv: 1404.0284. [Online]. Available: http://arxiv.org/abs/1404.0284

[21] Qi Li, Jinsong Zheng, A. Tsai, and Qiru Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 3, pp. 146–157, 2002.

[22] J. Canny, "A computational approach to edge detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 8, no. 6, pp. 679–698, 1986.

[23] H. Moon, R. Chellappa, and A. Rosenfeld, "Optimal edge-based shape detection," *Image Processing, IEEE Transactions on*, vol. 11, no. 11, pp. 1209–1227, 2002.

[24] A. Paplinski, "Directional filtering in edge detection," *Image Processing, IEEE Transactions on*, vol. 7, no. 4, pp. 611–615, 1998.

[25] Ruey-Shy Guh and Yi-Chih Hsieh, "A neural network based model for abnormal pattern recognition of control charts," *Computers & Industrial Engineering*, vol. 36, no. 1, pp. 97–108, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360835299000042

[26] Philipp Klein and Dirk Benyoucef, "SmartMetering: Classification of appliances without device-specific data base," Ilmenau, Germany, 2012, pp. 6–9.

A service robot platform for individuals with disabilities

Wolfgang Ertel[1], Steffen Pfiffner[1], Benjamin Reiner[1], Benjamin Stähle[1], Markus Schneider[1]
Jörg Schmal[2], Barbara Weber-Fiori[1] and Maik H.-J. Winter[1]

***Abstract*— Given the background of demographic change, the German national project AsRoBe addresses the question whether people with physical disabilities can benefit from the support by mobile service robots. Based on the analysis of user requirements obtained from a literature study and an inquiry among people with disabilities an innovative service robot was designed. This robot is presented here for the first time.**

I. INTRODUCTION

The project AsRoBe is an interdisciplinary project involving three partners. The institute for artificial intelligence provides an innovative service robot and the user interfaces for the disabled people. The faculty for social sciences does the user requirement analysis and the design and evaluation of tests with disabled people in their homes. These people live in the facilities of the third partner, a large social services provider in south Germany. This project aims at answering the following questions:

- What type of help do physically disabled people need?
- Can medium-term available service robots provide such help at acceptable cost?
- How must such a robot be constructed?
- How should human and robot interact?
- Will the disabled people accept such machines as a substitute or an extension of their human caretaker?
- Are other technical aids better suited?

This paper presents results from a literature study and from an inquiry among people with physical disabilities. The primary focus of this study was to get a survey of the requirements. The results shown in Section II and III lead to the conclusion that none of the currently commercially available service robots is able to fulfill all the requirements. It turned out that already the mechanical features of the existing robots are insufficient. In Section IV we will present a service robot that solves all these problems.

II. META-ANALYSIS OF REQUIREMENTS

For elderly people and individuals with disabilities, even simple everyday activities are problematic [8]. In [16], we performed a systematic literature review about requirements and demands on service robots in domestic environments from the perspective of different user groups (mainly elderly people). The results are evaluated with Evidence-based Practice (EbP)[3]. We shortly summarize the results in this section. The original version of Figure 1 and Table I can be found in [16].

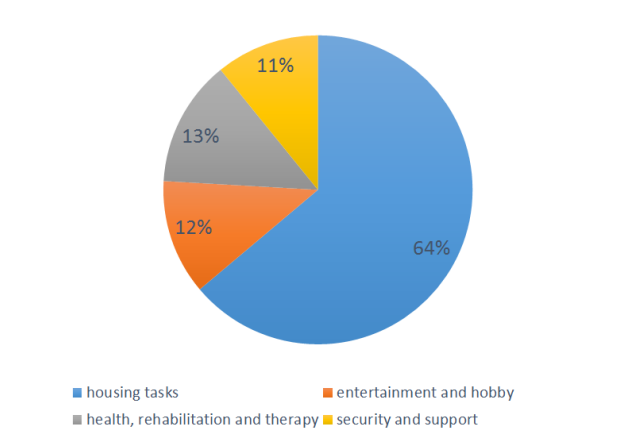


Fig. 1: Demands grouped by segment

The meta-analysis shows the requirements and demands on service robot systems. However (with the exception of [5]) the participants are not limited to humans with existing physical limitations. Sometimes the requirements are gathered from industrial and business experts and not from actual users of such systems. Table I shows the demands identified from literature grouped in segments *housing tasks*, *entertainment and hobby*, *health, rehabilitation and therapy*, *security and support*. Figure 1 summarizes the demands per segment. 64% of the demanded functions are located within the group of housing tasks. The other three segments are nearly equally distributed with 11% to 13% each.

The segment of housing tasks contains work like *pick and place*, *cleaning*, *ironing* and others which become more and more difficult with age or disabilities. Most studies show that pick and place tasks are especially important. This includes to serve drinks and food, but also to lift heavy objects. Tasks which takes very long such as window cleaning should be (according to test persons) done by a service robot. Also the handling of light switching functionality is often requested. According to [15] more personal or even confidential tasks, such as cooking, are reluctantly given to service robots whereas tasks like vacuum cleaning can willingly handed over to machines. Robots can assist in communication such as e-mail and telephony, but reading a book to somebody is explicitly not demanded according to [11], [16]. Health monitoring seems to be a reasonable functionality to be expected from a service robot. However people are concerned

[1]Ravensburg-Weingarten University of Applied Sciences, Weingarten, Germany, Email: {reiner, steahle, pfiffner, markus.schneider, ertel, weber-fiori, maik.winter}@hs-weingarten.de
[2]Email: joergschmal@gmail.com

Segment	Task	[1]	[2]	[4]	[5]	[6]	[9]	[11]	[12]	[15]
Housing tasks	turn on and off				x					
	tidy up		x		x			x	x	x
	cleaning floors		x				x	x		
	ironing							x	x	x
	cleaning ceilings and walls							x		
	cleaning windows						x	x	x	x
	pick up objects		x		x		x		x	x
	pick objects from high positions		x				x			
	dish washing		x					x	x	x
	pick and bring / transport		x		x			x	x	
	prepare meals		x		x		x		x	
	clean a surface							x		
	water plants							x		x
	repairing							x		
	vacuum		x			x	x		x	x
	dusting									x
	open doors		x							x
	laundry		x				x	x	x	x
Entertainment and hobby	banking		x							
	informations acquisition		x							
	communication		x	x	x	x				
	reading		x							
	sewing									x
	orientation				x					
	entertainment / hobby		x			x				
Health, rehabilitation and therapy	remainder functionality			x	x		x	x	x	
	health monitoring			x	x					
	mobility		x		x					
	self-care assistance	x	x							
Security and support	fire alarm / water damage prevention						x			
	accompaniment				x					
	facility surveillance		x	x		x				x
	locking doors				x					
	emergency calls functionality			x	x		x			

TABLE I: All demanded tasks of a service robot identified in literature

about safety issues. Housing security is another segment where a service robot can be useful and help people without interfering too much with their autonomy. This could for example be an alarm system prevent to fire or water damage. Also an emergency call to get medical help in case of an accident could be performed by a service robot. Additionally a robot could be used to secure a home, for example doors could be locked and monitored by a robot. However the surveyed people are concerned about their privacy and in their opinion the best approach would be a system which is only active in an emergency [4], [5], [16].

III. QUALITATIVE ANALYSIS OF REQUIREMENTS AND DEMANDS

In order to specify the needs and demands of individuals with disabilities more concretely, we performed a requirement analysis for potential future users of our service robot. Therefore face-to-face interviews were conducted as proposed in [7]. This supplements the previously shown meta-analysis and contributes additional information for potential possibilities of support by a service robot in a domestic environment for individuals with different physical disabilities (this inquiry is still in process).

A. Method

For the exploratory inquiry we perform guideline based interviews (face-to-face). The guideline is structured in two parts and contains open and closed questions with multiple dimensions to clarify the research questions [10]. Part I contains questions about the type, progress and perspective of the disease, demanded support (the classification is modeled after [13]) types of support (formally and informally), the application of (technical) means and more.

Part II discusses questions about demands, requirements and the attitude towards the planned service robot. Prefixed to the questions in part II the application of a service robot is illustrated in an exemplary movie (vignette). The concrete determination of demands concerning service robots is oriented on the empirical data inquired and is limited to the (technical) objective within our project. The present results are now characterized in a descriptive manner.

B. Sampling description

- Number of test persons: $n = 8$ (state June 2014).
- Types of diseases / disabilities: Tetra-spasticism, poliomyelitis and post-polio syndrome, cerebrovascular accident, multiple sclerosis, osteogenesis imperfecta and osteoarthrosis, spinal canal stenosis. All persons have restricted mobility (wheelchairs) and have, additional to

the diagnoses mentions above, functional or sensorical restrictions (seeing, hearing, grasping, and others).
- Age: $43 - 82$ years.
- Sex: five female and three male test persons.
- Domestic situation: All live alone at home and get support through various home care services.

C. First results

First qualitative results are mostly in accordance with the conclusion shown in the meta-analysis. Shown are tasks considered as essential as well as additionally identified supporting functionalities of a service robot in a domestic environment: *Pick and bring services* offered by a service robot are mostly assessed as useful functionality, independently of how difficult the mobility within the apartment is for the test person. Also assessed as useful is to *pick up objects from the ground* if bending down is not possible or dangerous, whereby in this case mechanical gripping pliers are already used. *Power on and off* devices and switches is assessed frequently as relieving for example if a person lies already in bed. To *open and close doors* is partially considered as interesting feature depending on the domestic situation. *Reminder functionality* is considered extremely useful by older as well as younger persons. Safety mechanisms such as *stove monitoring* are preferred primarily by older people. Test persons which already had negative experiences with emergency systems could imagine to have such an *emergency system* integrated into a service robot for improved service (for example to automatically call help after a fall). In order to gather more autonomy by getting more independent from social service, people are partly willing to accept direct *physical assistance* if it would be possible for the robot to support transfers from/to bed or the wheelchair. The connection of a service robot to an existing *fire alarm system* to call nearby help would contribute to safety, especially for bedridden people or wheelchair users. The willingness or necessity of *health monitoring* through a robot depends on the individual health situation. We will publish a complete and more detailed analysis by the end of 2014.

The goal of this work is now to build a mobile service robot which is capable of fulfilling many of the demanded tasks while being cost efficient and extensible for future requirements.

IV. THE ROBOT DESIGN

As we can see from the meta-analysis the demands and requirements on a service robot are versatile. There are even more additional requirements on the robot hardware. For example such a system should be able to operate long hours with maybe little time in between tasks to re-charge the battery. It must be stable, i.e., it must not fall over while driving even if it has some payload on board. Since indoor environments are sometimes narrow it is important that the turning radius is small. It needs a manipulator that is able to grasp objects in a wide range and (for safety reasons) must be certified to work with humans.

Fig. 2: Current state of the robot

Figure 2 shows our robot platform. As a mobile base platform we are using a MPO 700 from Neobotix[1]. The four omni-drive modules enable it to move in any direction and to rotate while driving. It has a weight of 120kg with a payload of 300kg. Its dimensions are $(L \times W \times H)$ $741\text{mm} \times 509\text{mm} \times 348\text{mm}$. The large battery (28Ah) allows about 5 hours of autonomous driving. The low center of mass combined with the high weight form a stable foundation for the robot.

We use two Sick 2D laser scanners for mapping, localization and collision avoidance.

Mounted on top is a linear guiding for the Jaco robot arm by Kinova[2]. The Jaco robot arm was originally intended as an assistive robot arm mounted to a wheelchair operated in close proximity to humans. The target audience are people with dexterity impairment in the upper body such as muscular dystrophy, amyotrophic lateral sclerosis, spinal cord injury, multiple sclerosis and neurological disorders. It has a total weight of only 5.7kg and a maximum payload of 1.0kg to 1.5kg, depending on the arm extension. With its 90cm reach we can grasp objects which would be out of range for other robot arms. The linear guiding allows us to move the arm up to grasp objects from a high shelf or move down for objects on the floor.

For object detection and face recognition we use the 3D sensor Kinect from Microsoft. In order to be able to use

[1] http://www.neobotix-robots.com
[2] http://kinovarobotics.com

Fig. 3: Manipulator which can hold a touch pad for human robot interaction.

the 3D information to enrich the localization and mapping information from the laser scanner, we mounted the Kinect on a pan-tilt unit. This allows us to see and map obstacles which are located in higher positions which are not reachable via laser.

Our service robot runs on the *Robot Operating System (ROS)* [14]. This allows us to choose from a variety of software modules such as *Simultaneous Localization and Mapping*, *Planning*, *Grasping*, *Object Detection* while developing new modules that can be used by other robotic groups.

A very important question is how robot and human interact. Our primary communication channel is voice recognition and speech synthesis which allows a very natural way of controlling a robot. We already have a primitive system which can process simple, pre-defined commands like *"Go to kitchen"* or *"Grasp cup"* which are mapped to actions. We are planning to offer several other possibilities for humans with a limited sense of hearing or which have difficulties to control a robot per voice such as a 3D mouse, a keyboard and a touch pad mounted on a second manipulator shown in Figure 3. This manipulator will be placed behind the other robot arm.

V. AsRoBe Milestones

Important research questions in the context of the AsRoBe Project are:

- **Demands on assitive service robots:** The question which support is needed by humans with different physical disabilities was manly answered by the meta analysis and summarized in this work.
- **Human robot interface:** How can and want elderly people or persons with disabilities control a service robot, command it and communicate with it in general? We will realize and evaluate different human robot interfaces in field trials.
- **Technical feasibility:** Is it possible to transfer intelligent behaviour, especially learning capabilities of the robot from laboratory to every day life in the living facilities of humans with disabilities or elderly people? Is it, in general, possible to meet the expectations on service robots with the technologies available today or in a medium-term time frame? Can robots be intelligent and adaptive enough to full fill all these demands?
- **Acceptance:** To what extent will the assistive service robot be accepted by the users? Where are the possibilities and limits of the robot applications from a user point of view? In what sense are caveats existent and what are their sources?
- **Robot programming by lay people:** The apartments of humans with disabilities or elderly people are much less structured than industrial environments and highly individual. Furthermore, the tasks required to fulfill by the robot may be very different and may not be known in advance. Hence it must be easy to reprogram the robot. Is it possible for users or nursing staff without technical knowledge to program or train the robot for new tasks? We will realize and evaluate this with *learning by demonstration* techniques.
- **Robots and Smart-Home:** Is it practical to connect a mobile service robot to smart-home technologies such as wireless connections to various domestic appliances? For example such a robot could control the heater, window shades or electric stoves. Is it feasible to attach RFID tags to various objects in order to facilitate a reliable object recognition? To what extent can such a robot be still considered autonomous?
- **Behaviour modelling:** In order to require less active user control and equip the robot with capabilities to plan useful tasks ahead the robot must be able to build a model of the user's behaviour and preferences. Can such a behaviour modelling be realized with statistical methods and machine learning techniques?

VI. Summary

In this work we presented a meta-analysis for requirements and demands on mobile service robot platforms with focus on people with physical disabilities. Based on this knowledge we built a service robot that is especially designed to work with humans in order to help them in their daily life. The next step in this project is to deploy this robot in the home of several test persons with physical disabilities and evaluate its performance. The gained information will be used to further improve this service robot design and its interaction with humans.

VII. Acknowledgment

The authors gratefully acknowledge the financial support from the Baden-Württemberg-Stiftung.

REFERENCES

[1] Heidrun Becker. *Robotik in Betreuung und Gesundheitsversorgung*, volume 58. vdf Hochschulverlag AG, 2013.

[2] Sandra Bedaf, Gert Jan Gelderblom, Dag Sverre Syrdal, Hagen Lehmann, Hervé Michel, David Hewson, Farshid Amirabdollahian, Kerstin Dautenhahn, and Luc de Witte. Which activities threaten independent living of elderly when becoming problematic: inspiration for meaningful service robot functionality. *Disability and Rehabilitation: Assistive Technology*, (0):1–8, 2013.

[3] J. Behrens and G. Langer. *Evidence-based Nursing and Caring. Methoden und Ethik der Pflegepraxis und Versorgungsforschung. 3.A. Bern.* 2010.

[4] Patrick Boissy, Hélène Corriveau, François Michaud, Daniel Labonté, and Marie-Pier Royer. A qualitative study of in-home robotic telepresence for home care of community-living elderly subjects. *Journal of Telemedicine and Telecare*, 13(2):79–84, 2007.

[5] Elizabeth Broadbent, Rie Tamagawa, Anna Patience, Brett Knock, Ngaire Kerse, Karen Day, and Bruce A MacDonald. Attitudes towards health-care robots in a retirement village. *Australasian journal on ageing*, 31(2):115–120, 2012.

[6] Kerstin Dautenhahn, Sarah Woods, Christina Kaouri, Michael L Walters, Kheng Lee Koay, and Iain Werry. What is a robot companion-friend, assistant or butler? In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 1192–1197. IEEE, 2005.

[7] Uwe Flick. *An introduction to qualitative research. 4th ed.* Sage, Los Angeles, 2009.

[8] W. Friesdorf, D. Mayer, and A. Heine. *Sentha - seniorengerechte Technik im häuslichen Alltag: ein Forschungsbericht mit integriertem Roman*. Springer, 2007.

[9] Panu Harmo, Tapio Taipalus, Jere Knuuttila, José Vallet, and Aarne Halme. Needs and solutions-home automation and service robots for the elderly and disabled. In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 3201–3206. IEEE, 2005.

[10] Cornelia Helfferich. Die Qualität qualitativer Daten. *Manual für die Durchführung qualitativer Interviews*, 2:152–153, 2011.

[11] Zayera Khan. Attitudes towards intelligent service robots. *NADA KTH, Stockholm*, 17, 1998.

[12] Sibylle Meyer. *Mein Freund der Roboter: Servicerobotik für ältere Menschen-eine Antwort auf den demografischen Wandel?; Studie im Auftrag von VDE-Verband der Elektrotechnik Elektronik Informationstechnik eV...* VDE-Verlag, 2011.

[13] World Health Organization. *ICF. The International Classification of Functioning disabilities and health : short version*. Geneva: World Health Organization, 2001.

[14] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. ROS: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5, 2009.

[15] Céline Ray, Francesco Mondada, and Roland Siegwart. What do people expect from robots? In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 3816–3821. IEEE, 2008.

[16] Joerg Schmal. Assistenz- und Serviceroboter für Menschen mit körperlichen Funktionseinschränkungen: Eine Analyse der geforderten Tätigkeiten im häuslichen Umfeld, 2014.

Big Data improving Ambient Assisted Living Solutions

Carina Rosencrantz, Christoph Reich
Faculty of Computer Science
Furtwangen University of Applied Science
Furtwangen, Germany
Email: {Carina.Rosencrantz, Christoph.Reich}@hs-furtwangen.de

***Abstract*—Big Data methods offer great opportunities to research in many different domains. In the research domains of social media or medicine they are very famous. With Big Data it is possible to revolutionize the technical systems and make them more adaptable, context-aware and customizable. Especially in the Ambient Assisted Living (AAL) domain such systems are needed to assist the older adults in their daily tasks and enable them a largely independent life. This paper discusses some existing work in this research area and presents ideas of reaching a better assistance by using Big Data methods in AAL. It also shows problems and challenges, like data privacy or data variety, rising up with this evolution.**

***Keywords*—Big Data, Ambient Assisted Living, Privacy, Volume, Velocity, Variety, Veracity**

I. Introduction

With the growing age of the population, intelligent systems are needed, which assist elderly people in staying longer in their preferred living environment. Most of these Ambient Assisted Living (AAL) systems offer a saver life by observing the assisted person and their environment like in [1], [2] and [3]. They use different sensors (wearables or environment sensors) to detect emergency cases like a fall of the person, danger of fire caused by home appliances or a critical health condition. But there are also sensors and actuators for home automation helping the person to master their daily life. Other systems provide the, in most cases alone living, elderly person the opportunity to stay socially integrated by helping to maintain the communication with other people and to gather information relevant to the person's needs [4]. All these assistance systems produce a lot of data and a large variety of data. This data could be used to improve existing solutions (e.g. to reduce false emergency detections or to increase the user experience of software products) and to develop new AAL products (e.g. a system that could monitor the feel-good-factor of a person). But for realizing this, there have to be technologies to gather, save, filter and analyze the data, which can be found in Big Data methods.

Big Data can't be handled with traditional data management principles and has according to Laney [5] three characteristics, the 3 V's: Volume (how much data), Velocity (how fast that data is processed) and Variety (the different types of data). Sometimes the 3 V's are complemented by Veracity (data integrity).

The Big Data approach is widespread in the domain of social media (keyword: Big Social Media Data), of medicine and of research of the buying behavior. The first two domains are also interesting for AAL solutions. More information about the person and their environment is needed to adapt the services to the individual needs of the person or to improve the assistance systems in general.

In Section II some related work is described. In Section III the idea of combining Big Data methods with Ambient Assisted Living solutions is presented and various questions and problems accuring with this idea are outlined. In Section V a short conclusion is given.

II. Related Work

In the field of Ambient Assisted Living the keyword Big Data is still not widespread. But there is research done in different aspects or related technologies of Big Data. There are solutions for home automation for the elderly people analyzing data of different sensors to react through actuators in a suitable way. In the field of medicine the data of wearables can be evaluated and sometimes the processed data can be sent to caring relatives or directly to healthcare professionals like the approach of the AAL project REMOTE [6] or like the ERMHAN platform, which supports care networks in providing them the prepared health data of the patient to reach a better home-based assistance [7].

In the paper of Dohr et al. [8], they present a study showing the benefits of the Internet of Things in the Ambient Assisted Living domain. In particular, they examined the usage of Keep In Touch (KIT), which uses smart objects and technologies (Near Field Communication, Radio Frequency Identification and smart phones) to facilitate the telemonitoring of health data, in combination with Closed Loop Healthcare Services, which enable communication channels between the care-givers and the elderly people to analyze the data. But the Internet of Things is just one part of the possibilities given by Big Data and AAL. Better assistance could be gained by analyzing more data like, for example, social, emotional or behavioral data to become a better understanding of the needs of the elderly people.

In the work of Jiang et al. [9] a Big Data pilot system for health monitoring for elderly people is presented. They developed a wearable sensor, which includes an accelerometer to measure the activities of the wearer, an ambient temperature sensor, a skin temperature sensor and a sensor for heartbeat and for SPO2 in blood measurement. This wearable sends the data to the connected smart phone, on which an intelligent

information forwarder based on a Hidden Markov Model is deployed. This forwarder is capable to monitor the behaviors of the wearer continuously, to detect anomalies and inform a caregiver about it and to forward only the interesting information to the healthcare Big Data system for further analysis. So the sensors are enriched with context-awareness and the communication loads and data storage can be reduced. A better understanding of the health status and the behavior of a user could be reached if there would be more data to combine with the measured data of the wearable sensors. But with a growing volume of data, another more powerful unit has to be used than a smart phone.

In [10] Dobrican and Zampunieris presented a work-in-progress describing a model of a distributed network of proactive, self-adaptive and context-aware systems. The idea is to have local proactive systems for each user which are all connected to one network. So it is possible, for example, to connect people with the same interests or connect many patients with their physician. A simple use case could be somebody with such a system installed searching for a ride on a car-sharing website. Another user of such a system, who lives nearby the other one, searches for a ride to the same destination, too. So the systems could propose both to share a ride. The connection of many users and their produced data with a common analysis would be very beneficial for Ambient Assisted Living solutions.

The social media domain offers also opportunities to AAL. Social media platforms like Facebook, Twitter or Google+, which are more and more used also by elderly people, analyze their users and their social environment in detail. Because of the big amount of users producing a huge amount of data they are, for example, able to describe the relationships of the users by analyzing the frequency of contact to another user and the message contents. But they are also able to make predictions of, for example, trends or the result of upcoming elections, as described by Cameron Marlow, the sociologist from Facebook, in an interview in [11]. Additionally he said, that they examine also the data for improving their product (the Facebook platform and his features) itself.

III. Big Data and AAL

Big Data methods can be used in many subjects, for example, for home automation, for medical assistance (e.g. medication, therapy, diagnosis) or for social media (e.g. social integration through maintaining communication to contacts or building up new contacts). For the Ambient Assisted Living domain, these three subjects are mainly relevant. Through generation of knowledge out of the data of each single domain, it would be possible to improve existing AAL systems or to develop new services (see Figure 1). An example in the social media domain could be for example a social platform, where older adults can have contact to their village community to plan events together. Information about the message exchange and content, the interests of the users and the social relationships between them, could be used to help them by finding new contacts with the same interests, by getting new ideas of events or meetings, by motivating them to go out and meet people or by organizing the events. In the health domain there could be a remote therapy assistance. The physician could get information about the actual health status of the patient and could propose, for example, physiotherapeutic exercises, which could be observed by the Big Data AAL system.

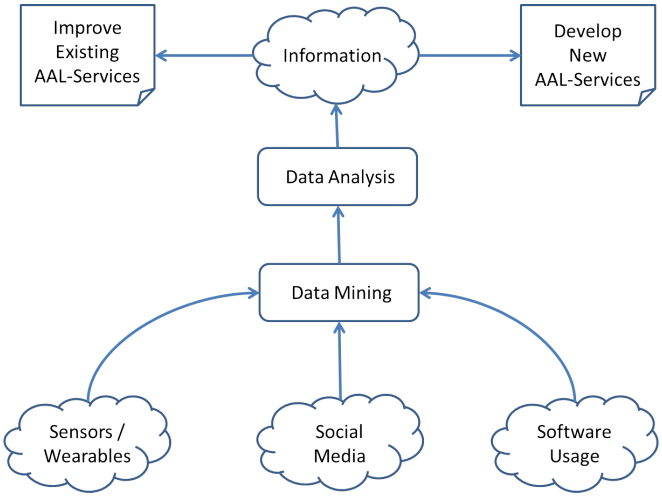


Fig. 1. Big Data improving AAL

To reach an even better assistance for elderly people it would be beneficial to combine the knowledge of all of the three domains. On the one hand each single person could be analyzed with his health and emotional condition, his social behavior and his environment in combination. So it would be possible to offer personalized services, which could be optimized all the time, assisting the person in his daily activities and his social integration. On the other hand data from many users could be anonymized and analyzed to improve the assistance systems for all of the users like Facebook did. It would be also possible to make predictions, for example, about health progressions by examining single groups of patients (dementia process, e.g.).

IV. Problems and Challenges

But there several problems to solve and questions to be answered before reaching such a Big Data usage in AAL. One of the largest problems to solve, when involving Big Data methods in AAL solutions, is to comply with the rules of data privacy. Some other challenges or questions rising up with the combination of Big Data and AAL are discussed with the help of the Big Data 4 V's.

A. Privacy

Above all there is the problem of data privacy (discussed by Weichert in [12]). In Germany, when personal data is analyzed, the person concerned must agree to this form of analysis (fundamental right to informational self-determination). When there is no need for direct personal reference, the data should be anonymized and aggregated in such a way that there is no possibility to assign the data to the person again. It must be considered that with growing quality of data and growing amount of characteristics in one data set the probability of possible assignments grows. Another aspect of data privacy is the accessibility of personal data. For this, it would be perhaps useful to classify the data in critical to safety levels to realize a finely granulated access control.

B. Volume

An example of a usage of such a Big Data AAL solution is in an assisted living center with many people to care. There it would be possible to install the same platforms and devices in every flat for each person. In such an environment there could be a huge amount of data which could be gathered and analyzed. So it would be possible to assist the individual person by analyzing personal data and to improve the services by analyzing relevant data from many persons.

C. Velocity

Velocity can be divided in two aspects. First it means the speed of data input, which is different from case to case. Some data like health information is perhaps sent more frequently than information about the window or door status. Second it means the speed of data processing which should perhaps depend on the priority of data. A dangerous situation should be recognized immediately and a change of user preferences can be treated less prioritised.

D. Variety

The variety of data in such a Big Data AAL solution is huge. There is high level and low level data, structured and unstructured data and different data sources. Low level data is mostly produced by sensors and wearables and must be enriched with semantics to reach a higher abstraction layer to gain information and knowledge from this data. There is also high level data which has a degree of meaning like social media data (content of messages, information requests, etc.). Categories are, for example, medical, social or environmental data. Variety is also given by the different stakeholders of such an AAL system (assisted persons, caring relatives, healthcare professionals). The problem of variety rises first of all because of the missing standards in the Big Data and AAL domain. At the moment, a working group at NIST (National Institute of Standards and Technology) [13] works on a report with recommendations in the Big Data domain.

E. Veracity

The quality and reliability of data has also to be considered. Mainly for sensors, a filter should be used, that sorts out wrong data. Data, that is entered by the user itself, for example, is normally reliable. In some cases a learning system is also a possibility, that can decide after a while, which data is reliable and which isn't.

V. Conclusion

This paper described some aspects of the usage of Big Data in the field of Ambient Assisted Living. It presented other work in this domain and pointed out some challenges rising up with Big Data and AAL combination. Most of the research in this area concentrates on health monitoring with sensors and the transmission of this information to the care-givers like caring relatives or physicians. Analyzing not only sensor data but also social and environmental data and combining the findings could lead to more personalized services. Aggregating and anonymising the data from many users would be beneficial to improve assistance systems and services in general. There is much information that could be extracted and many services that could be realized with the help of these knowledge but it should be examined what is realizable with conforming to data privacy. The described idea and the problems, like data privacy or the variety of data, in an Big Data AAL solution are just some examples and should be examined in detail in further research.

Acknowledgment

The project ZAFH-AAL ("Zentrum für Angewandte Forschung an Hochschulen für Ambient Assisted Living") is funded by the Ministry of Science, Research and the Arts of Baden-Württemberg, Germany. The funding program for the universities of applied science is called: Zukunftsoffensive IV "Innovation und Exzellenz" (ZO IV). The PCEICL project is a sub-project of the project ZAFH-AAL.

References

[1] M. Klein, A. Schmidt, and R. Lauer, "Ontology-Centred Design of an Ambient Middleware for Assisted Living: The Case of SOPRANO," in *In: Towards Ambient Intelligence: Methods for Cooperating Ensembles in Ubiquitous Environments (AIM-CU), 30th Annual German Conference on Artificial Intelligence (KI 2007)*, 2007.

[2] L. Litz and M. Gross, "Covering Assisted Living Key Areas based on Home Automation Sensors," in *Networking, Sensing and Control, 2007 IEEE International Conference on*, 2007, pp. 639–643.

[3] J. A. Botia, A. Villa, and J. Palma, "Ambient Assisted Living system for in-home monitoring of healthy independent elders," *Expert Systems with Applications*, vol. 39, no. 9, pp. 8136 – 8148, 2012.

[4] C. Fredrich, H. Kuijs, and C. Reich, "An Ontology for User Profile Modelling in the Field of Ambient Assisted Living," in *SERVICE COMPUTATION 2014, The Sixth International Conferences on Advanced Service Computing*. IARIA XPS Press, 2014, pp. 24–31.

[5] D. Laney, "3D Data Management: Controlling Data Volume, Velocity, and Variety," META Group, Tech. Rep., February 2001. [Online]. Available: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[6] Homepage of REMOTE project. REMOTE project. Retrieved: September, 2014. [Online]. Available: http://www.remote-project.eu

[7] F. Paganelli, E. Spinicci, and D. Giuli, "ERMHAN: A Context-aware Service Platform to Support Continuous Care Networks for Home-based Assistance," *Int. J. Telemedicine Appl.*, vol. 2008, pp. 4:1–4:13, Jan. 2008. [Online]. Available: http://dx.doi.org/10.1155/2008/867639

[8] A. Dohr, R. Modre-Opsrian, M. Drobics, D. Hayn, and G. Schreier, "The Internet of Things for Ambient Assisted Living," in *Information Technology: New Generations (ITNG), 2010 Seventh International Conference on*, April 2010, pp. 804–809.

[9] P. Jiang, J. Winkley, C. Zhao, R. Munnoch, G. Min, and L. Yang, "An Intelligent Information Forwarder for Healthcare Big Data Systems With Distributed Wearable Sensors," pp. 1–9, 2014.

[10] R.-A. Dobrican and D. Zampunieris, "Moving Towards a Distributed Network of Proactive, Self-Adaptive and Context-Aware Systems," in *ADAPTIVE 2014, The Sixth International Conference on Adaptive and Self-Adaptive Systems and Applications*. IARIA XPS Press, 2014, pp. 22–26.

[11] H. Geiselberger and T. Moorstedt, Eds., *Big Data: Das neue Versprechen der Allwissenheit*, 2nd ed., ser. edition unseld. Berlin: Suhrkamp, 2013.

[12] T. Weichert, "Big Data und Datenschutz," March 2013, retrieved: September, 2014. [Online]. Available: https://www.datenschutzzentrum.de/bigdata/20130318-bigdata-und-datenschutz.pdf

[13] (2014, January) ISO/IEC JTC 1 Study Group on Big Data. National Institute of Standards and Technology. Retrieved: September, 2014. [Online]. Available: http://jtc1bigdatasg.nist.gov/home.php