Hans-Jörg Schulz, Bodo Urban, Uwe Freiherr von Lukas (Eds.)

Proceedings of the International Summer School on Visual Computing 2015

August 17-21, 2015 in Rostock, Germany

FRAUNHOFER VERLAG

Proceedings of the International

SUMMER SCHOOL on Visual Computing

2015

Hans-Jörg Schulz, Bodo Urban,
Uwe Freiherr von Lukas (Eds.)

Proceedings of the International SUMMER SCHOOL on Visual Computing 2015

August 17-21, 2015 in Rostock, Germany

Visual Computing
RESEARCH AND INNOVATION CENTER

Fraunhofer
IGD

Universität
Rostock
Traditio et Innovatio

Kontakt:
Herausgeber: Hans-Jörg Schulz, Bodo Urban, Uwe Freiherr von Lukas
Fraunhofer-Institut für Graphische Datenverarbeitung
Joachim-Jungius-Straße 11, 18059 Rostock
Telefon: +49 381 4024-110
Telefax: +49 381 4024-199
E-Mail: summerschool@igd-r.fraunhofer.de
URL: http://summerschool.igd-r.fraunhofer.de

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.
ISBN: **978-3-8396-0960-6**

Druck und Weiterverarbeitung:
IRB Mediendienstleistungen
Fraunhofer-Informationszentrum Raum und Bau IRB, Stuttgart

Für den Druck des Buches wurde chlor- und säurefreies Papier verwendet.

Fraunhofer-Informationszentrum Raum und Bau IRB
Postfach 800469, 70504 Stuttgart
Nobelstraße 12, 70569 Stuttgart
Telefon: +49 711 970-2500, Telefax: +49 711 970-2508
E-Mail: verlag@fraunhofer.de, URL: http://verlag.fraunhofer.de
Titelbild: © **OLIVER stockphoto - Fotolia.com**

Preface

The research field of *Visual Computing* encompasses everything graphical in computer science – from the synthesis and processing of graphical content to its human consumption. This broad spectrum includes multiple other fields that constitute research disciplines in their own right, such as perception, visualization, multimedia, virtual and augmented reality, as well as human-computer-interaction. The first International Summer School on Visual Computing held from August 17-21, 2015 at the Fraunhofer IGD Rostock, Germany aimed to give an overview of this broad field to graduate students from Rostock, Germany, and abroad. A week-long program of lectures and research talks by invited speakers introduced participating students into the subjects of visual perception and cognition, eye tracking, raster image databases, multimedia retrieval, computer vision, human-computer-interaction, mobile and wearable computing, and visual analytics. Each afternoon, the participating students had the opportunity to present their own research in posters and talks. Sessions with helpful tips and tricks on how to go about PhD level research, writing, and presentation, as well as an open lab tour rounded off the summer school program.

The post-conference proceedings at hand contain a selection of the research presented by the participants during that week. The 13 papers are grouped into three thematic sections: image generation, image analysis, and image usage. The following overview gives an impression of the breadth of topics they cover.

Part I: Image Generation contains five papers that concern themselves with technical issues and best practices of producing 3D and 2D images. In the first paper, S. Dübel et al. propose a novel flexible ray tracing architecture for terrain heightfields. Unlike conventional fixed pipelines, their architecture is able to renegotiate the tradeoff between rendering quality, rendering time, and available resources as needed. While this approach focuses on the technical aspects of rendering surfaces, the second paper by K. Furmanová addresses conceptual issues of visualizing two surfaces for their interactive comparison. In her case, these surfaces are facial scans that deviate in some parts and align in others, and she explores different means of superimposing these surfaces. The third paper in this part by K. Blumenstein et al. takes the challenge of display scalability to the number of screens and asks what to visualize on a second screen, such as a tablet, if one is available as an additional display device besides a regular TV set. There are a number of interesting technical questions involved in this setup, such as how to synchronize the TV's content with the content shown on the second screen. These are unique issues in the context of visualization and as such require novel solutions. The same holds true when employing visualization in different application domains, as it is exemplified by the fourth paper by C. Niederer et al. They surveyed the state of the art in visualizations for dynamic, weighted, directed, multimodal networks with a particular emphasis on visualizations used in data-driven journalism. In passing, their survey also updates existing surveys on dynamic graph visualization with the latest publications and developments in this area. Finally, J. Haider et al. give insight into best practices of developing visual analytics solutions from a comprehensive user study that was conducted in the UK. While the identified best practices were derived mainly for the scenario of comparative case analysis in criminal investigations, they are generalizable to the point of being valuable requirements that are applicable to the design of visual analytics solutions in other areas as well.

Part II: Image Analysis features four papers that contribute to the areas of image reconstruction, segmentation, restoration, and recognition. The first paper by T. Dolereit deals with refractive effects that impair the reconstruction of underwater structure from a stereo camera system. For doing so, the author infers additional constraints on the position and orientation of the refractive surface from the physically correct tracing of light rays. In addition to refraction, underwater images are often blurred, because of light scattering due to light attenuation and absorption. The second paper by F. Farhadifard aims to post-process such degraded images using a learned look-up scheme that does not require any prior knowledge about the scene or the water quality. The paper compares the effect of two different look-up schemes, so called dictionaries, that were generated for in-air images and underwater images, respectively. While these papers present image analysis techniques that operate on static images, the third paper by M. Radolko takes on the challenge of analyzing videos with the aim of separating foreground objects from a scene's background. To this end, it proposes an efficient implementation of a background subtraction algorithm that is evaluated with two different spatial models that incorporate assumptions about smooth regions in the scene. Lastly, the fourth paper in this part by A. Dadgar investigates how to detect hand gestures in image sequences. It gives an overview of Hidden-Markov-Model-based gesture recognition approaches and proposes two alternative approaches that hold the promise to overcome the difficulties that these approaches have with hand gesture recognition.

Part III: Image Usage is comprised of four papers that deal with the human factors of utilizing images for various tasks. The first paper by N. Flad et al. takes a measuring approach to gain insights into the information sampling and processing behavior of humans: The authors use eye-tracking and electroencephalography (EEG) to gather data about the sensation and cognition of visual stimuli. In their paper, the authors discuss a number of confounding factors in such data and in particular the side effects of the eye-tracking on the EEG results and how to clean the data from the resulting artifacts. Measurements play also central role in the second paper by J. Trimpop et al., which outlines a concept and architecture for a smart health support system that is based on sensor information gathered with a smartwatch. Depending on the use case scenario for this system, different functionalities are provided – e.g., emergency call features for the elderly, as well as fitness tracker features and accompanying visualizations for the younger generation. Different generations also play a role in the third paper by D. Matthies and A. Meier, which investigates the interaction between pedestrians and technology during navigation tasks. They find that even in the age of smartphones with GPS positioning, many people rely alternatively on landmarks and street signs, which are thus important features to consider when designing visual navigation aids. The paper by R. Alm and S. Hadlak concludes this part by showcasing a method for integrating and managing textual and pictorial annotations with a focus on manufacturing processes. Their method makes use of an ontological representation to derive contextually relevant annotations to show in certain situations.

For most participants, the paper they wrote for these proceedings was their first scientific paper. Yet from reading through them one could not tell. To a substantial part, this is due to an intensive revision cycle in which the board of reviewers has gone out of its way by providing quality feedback in a short time span, as well as the authors by incorporating the feedback to improve their papers. Together, these papers give an impressive overview of the excitement and incredible drive of the next generation of visual computing researchers that comes with new ideas and new technologies. We are proud that our summer school helped to further shape these ideas and to spark this excitement by giving input and fostering future cooperation between the participants. We wish them the best for their research careers!

Hans-Jörg Schulz, Bodo Urban, and Uwe Freiherr von Lukas

Participants, organizers, and guests of the International Summer School on Visual Computing 2015 in front of the Fraunhofer IGD Rostock.

Table of Contents

Part I - Image Generation: Rendering and Visualization

Part II - Image Analysis: Reconstruction, Segmentation, Restoration, and Recognition

Part III - Image Usage: Perception, Cognition, Interaction, and Annotation

Part I

Image Generation: Rendering and Visualization

A Flexible Architecture for Ray Tracing Terrain Heightfields

Steve Dübel[1], Lars Middendorf[2] Christian Haubelt, and Heidrun Schumann

[1] University of Rostock, Institute for Computer Science, D-18059 Rostock, Germany
steve.duebel@uni-rostock.de

[2] University of Rostock, Institute of Microelectronics and Data Technology, D-18119 Rostock, Germany lars.middendorf@uni-rostock.de

Abstract. High-quality interactive rendering of terrain surfaces is a challenging task, which requires compromises between rendering quality, rendering time and available resources. However, current solutions typically provide optimized strategies tailored to particular constraints. In this paper we propose a more scalable approach based on functional programming and introduce a flexible ray tracer for rendering terrain heightfields. This permits the dynamic composition of complex and recursive shaders. In order to exploit the concurrency of the GPU for a large number of dynamically created tasks with inter-dependencies, the functional model is represented as a token stream and is iteratively rewritten via pattern matching on multiple shader cores in parallel. A first prototype demonstrates the feasibility of our approach.

Key words: graphics hardware, terrain rendering, ray tracing

1 Introduction

With today's continuously growing amount of data and increased demand of quality, rendering complex terrain surfaces is a difficult task. Current heightfields can consist of hundreds of megabytes of raw data. To render them efficiently on current hardware, the heightfield needs to be triangulated, and appropriate level-of-details have to be defined. This results in data structures that easily increase the volume of raw data by an order of magnitude.

To considerably decrease this high memory consumption that in many scenario exceeds hardware capability, the heightfields can alternatively be rendered through ray tracing. In doing so, surface details are generated on-the-fly. Moreover, ray tracing allows for global illumination effects that significantly improve image quality. To enhance performance, ray tracing solutions use auxiliary data structures, but that increases memory consumption [1], or they compute approximations that decrease quality on the down side [2].

However, guaranteeing interactive frame rates requires an appropriate hardware support. This is mainly achieved by using multiple parallel processing units, e.g. clusters or many-core-systems [3, 4]. Common GPUs, above all, have been utilized for ray tracing. Tracing millions of individual rays in parallel by thousands of cores increases the performance of ray tracers significantly.

Since close interrelations between rendering quality, rendering time and available resources do exist, rendering approaches have to take these dependencies into account. In this way, changing requirements can be addressed. For example, quality can be prioritized before performance or vice versa: For instance, fully illuminated objects in the front need to be rendered in high quality, while objects in the dark or far away can be rendered with less effort to decrease rendering time.

Implementing a more flexible ray tracer comes along with problems for both options; GPU-based and CPU-based solutions. The recursive nature of ray tracing and the desired scalability of our approach do not fit well with the pipeline-based programming model of the GPU. Optix [5], a powerful and easy-to-use, general purpose ray tracing engine for the GPU, grant a better flexibility, but does not allow the user to fully customizing acceleration structures, buffer usage and task scheduling. On the other hand, the CPU permits high flexibility, but lacks high data parallelism. Thus, the performance of such CPU-based approaches is hardly sufficient. Hybrid solutions that run on CPU and GPU mostly suffer from the bottleneck of efficient communication. Hence, a novel approach is required.

In this paper, we propose a new rendering architecture for terrain visualization. The terrain is modeled as a mathematical function $f : \mathbb{R}^2 \to \mathbb{R} \times \mathbb{R}^3$ with $(x, y) \mapsto (z, (r, g, b))$ which provides an elevation ($z \in \mathbb{R}$) and a color value $(r, g, b) \in \mathbb{R}^3$ for each pair $(x, y) \in \mathbb{R}^2$ of terrain coordinates. The rendering architecture consists of three stages (Fig. 1).

The first stage (top of Fig. 1) is a flexible ray tracer that is made of two parts: *i*) a fixed, high efficient ray tracer kernel for terrain heightfields and *ii*) a modular extension unit. The ray tracing kernel utilizes beam tracing and fast intersection techniques to achieve real time frame rates. The output is a simple, colored image. The flexible, modular extension unit provides a set of enhanced rendering operators, which can be activated on demand to improve the image quality. Here, the generation of surface details (interpolation, microstructures, antialiasing) and advanced shading can be dynamically complemented by a set of appropriate modules. In this way, a trade-off between render time, render quality and available resources can be achieved.

The second stage provides a functional model (center of Fig. 1), defined as a network of executable tasks, including parallel or recursive parts. The basic ray tracing module, as well as the enhanced rendering operators are described as individual tasks, i.e. as nodes of a functional model. The functional model provides benefits like dynamic composition and recursion. However, it does not fit well into the data parallel execution model of current graphics hardware.

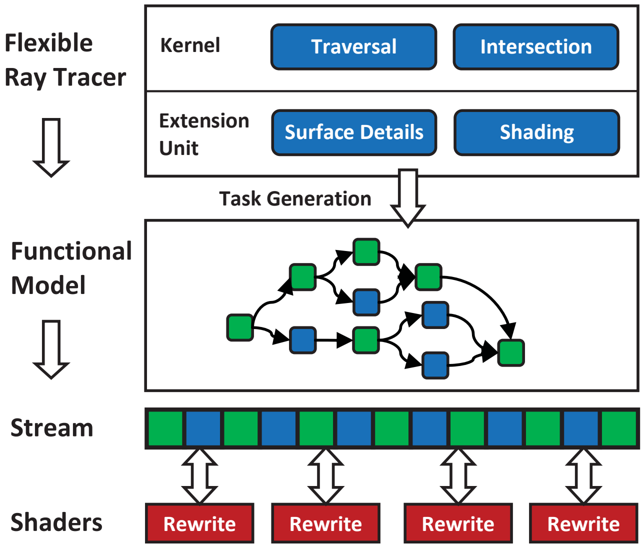


Fig. 1. Rendering architecture for flexible terrain ray-tracing.

Hence, with the *third stage* (bottom of Fig. 1) and in order to build a more flexible ray tracer, we propose a novel approach for dynamic task management of functional programs on the GPU. For this purpose, the functional model is encoded as a token stream and iteratively rewritten by several shader cores in parallel. In particular, invocations are represented by specific patterns in the stream that are replaced by the result of the corresponding functions. Therefore, all types of indirect and recursive functions can be evaluated in parallel.

In summary, the main contribution of this paper consists of a novel execution model for scheduling dynamic workload on the GPU and its application to the problem of terrain visualization. The most significant difference to related approaches like [6], [7], and [8], is the usage of pattern matching to resolve dependencies between tasks through local rewriting operations on the stream. The remainder of this work is structured as follows. In section 2 we present related work for interactive ray tracing and dynamic task scheduling. Next, the concept of the flexible terrain ray tracer will be introduced in section 3. The formal model and the parallel implementation for scheduling functional programs is subject of section 4 and 5, before we present a first prototype and respective results in section 6. Finally, section 7 concludes this report by a summary and hints for future research directions.

2 Related Work

Before going into the detail of our approach, we will briefly present related, state-of-the-art concepts of interactive terrain ray tracing on the one hand, and dynamic task management, on the other.

2.1 Interactive Ray Tracing

[9] describe three aspects to support interactive ray tracing: accelerating techniques, approximation and hardware.

Accelerating techniques To achieve acceptable frame rates, auxiliary data structures are necessary. Especially Bounding Volume Hierachies (BVH), kd-trees and grids [1] are used to accelerate the ray traversal and reduce intersection tests. BVH trees can be updated very fast, but do not adapt to the geometry as good as kd-trees do. In contrast to both, grids do not provide a hierarchy and no adaptability, but a fixed, uniform subdivision of the space [10]. However, this structure fits well to heightfield data. Our approach is based on [11], who extend the grid by a hierarchy, creating a so-called maximum mipmap data structure. This is similar to a quad-tree and allows for interactive ray casting on heightfields.

Other techniques accelerate ray tracing by exploiting ray coherence. That means, a number of rays are simultaneously traversed as ray packets [12] or beams [13]. Our approach utilizes a beam tracing based fast start, where a chunk of rays are traversed through the heightfield as one beam to calculate a map of starting points for the actual ray tracing (cf. Section 3.1).

Approximations Typically, intersection points and shading information can be determined by proper approximations. For instance, [14] proposed an efficient surface-ray-intersection algorithm for heightfields that is based on combination of uniform and binary search instead of an exact calculation. This might cause visible artifacts, but this problem is later solved through relaxed cone stepping [15, 16].

Moreover, global illumination can be approximated, since in outdoor scenes reflection and transparency do hardly contribute to shading. Such techniques were originally used to introduce global illumination effects to rasterization-based terrain rendering. A very simple approximation is ambient occlusion [17] that mainly estimates the locally limited distribution of ambient light by sampling the hemisphere through ray tracing. The average direction of unoccluded samples can additionally be used to include incident radiance e.g. through a look-up-texture (environment map). [18] extend this concept by additionally defining a cone, which aperture angle and alignment are based on the unoccluded samples and limits the incident light, considering a single light source. This technique is well-suited for outdoor terrain scenes, where the sun is the only light source, and considerably decreases render time. Our approach also supports different approximated illumination.

Hardware Numerous customized hardware solutions are developed [19, 20] to provide capable ray tracers. [21, 22] propose ray tracing approaches that are completely realized on the GPU. On the other hand, specialized hardware solutions address particular tasks of ray tracing, such as traversal and intersection, e.g. [20]. However, none of these architectures support flexible scaling between quality and performance or provide dynamic scheduling of tasks.

2.2 Dynamic Task Scheduling

There already exist several concepts for dynamic task scheduling for graphics processing.

Software Implementations Although the hardware architecture of modern GPUs is optimized towards data parallelism [23], dynamic scheduling of heterogeneous tasks can be implemented in software [6] and utilize work stealing for load-balancing [7]. Usually, a single kernel runs an infinite loop that consumes and processes tasks from queues in local or global memory [8]. However, if the tasks are selected via dynamic branching, irregular workloads can interfere with the single-instruction multiple-thread (SIMT) execution model of modern GPUs [23].

Our concept is compatible to these existing approaches, but additionally performs a pattern matching step to determine the readiness of a task. In particular, the relative execution order is controlled by data dependencies, which permit to efficiently embed complex task hierarchies into the stream, while both the creation and the completion of tasks are light-weight and local operations. [24]

Hardware Architecture Similar to our technique, the graphics processor proposed by [25] also stores the stages and the topology of a generic rendering pipeline as a stream. However, the scalability of the presented hardware implementation remains limited because the stream is decoded and reassembled sequentially. For comparison, our scheduling algorithm performs an out-of-order rewriting of the stream and keeps the tokens in the fast shared memory of a multiprocessor.

Programming Languages In addition, purely functional languages like *NOVA* were proposed for GPU programming [26] due to their applicability for automatic optimization techniques [27]. Predecessors like Vertigo [28] or Renaissance [29] are based on Haskell and allow composing complex objects from parametric surfaces and geometric operators in the shader. Similarly, the language *Spark* [30] introduces aspect-oriented shaders to permit a compact and modular description in the form of classes. While these approaches statically translate the source language into shaders, we introduce a runtime environment for the GPU, which retains the flexibility of functional programming at the expense of dynamic scheduling.

3 Flexible Terrain Ray Tracing

Our flexible rendering approach consists of two major components: A fixed ray tracing kernel to traverse the rays through a discrete terrain heightfield and a modular extension unit to control surface generation and shading. While the ray tracing kernel generates a simple, colored image, the extensions support flexibility to balance between rendering quality, rendering time and memory consumption. Both components are described in the following.

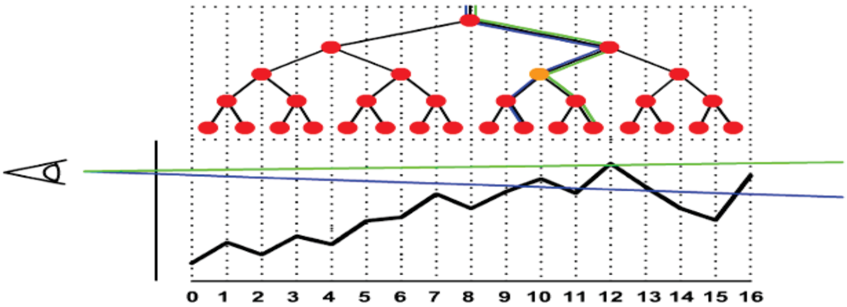


Fig. 2. Ray traversal in 2D for beam tracing. As long as the rays on the corners follow the same path through the tree, the whole beam visits the same node (orange). Otherwise, the beam needs to be subdivided.

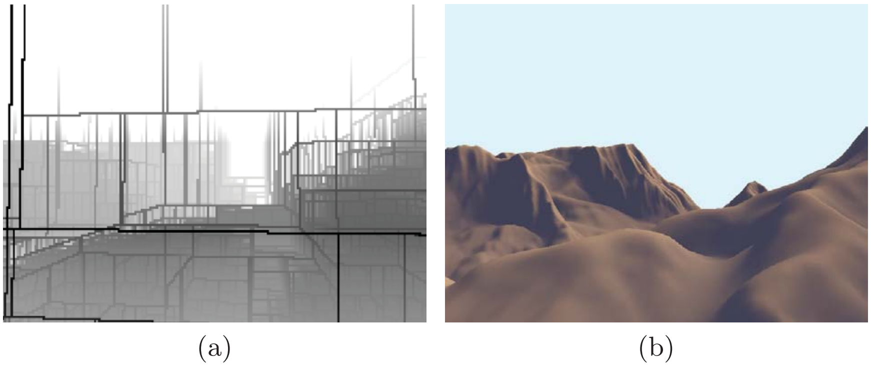
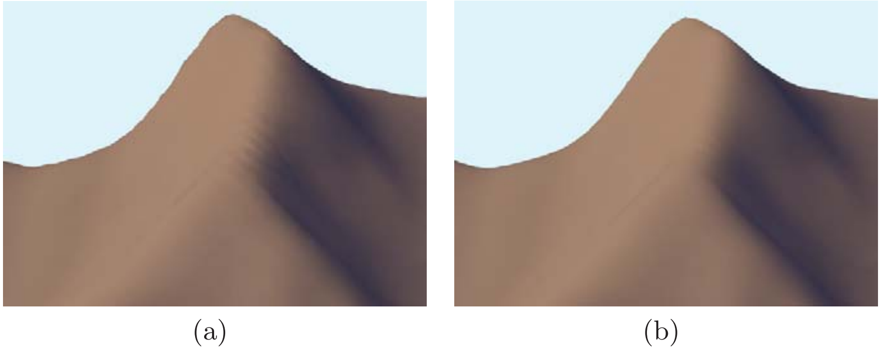


Fig. 3. Illustration of our beam tracing based fast start. (a) The result of beam tracing is the starting position of individual ray traversal encoded as a depth map. The grid structure reflects the split of them at the BV. (b) The final rendered image.

3.1 Ray Tracing Kernel

To ensure a high performance, we apply and combine sophisticated techniques from literature for both ray traversal and intersection tests. We utilize a grid-based bounding volume hierarchy, in particular the maximum mipmaps (MM) introduced by [11]. The maximum mipmap is structured as a quad tree that stores the maximum of all underlying height values at each node. The root node spans the whole heightfield, while a leaf node stores the maximum of four actual height values of the field. This structure can be constructed very fast. Since spatial information is stored implicitly, the increased memory footprint is very low (≈ 33%).

To decrease rendering time, we apply a beam tracing based fast start. The beams are pyramids with a base area of e.g. 8x8 pixels. The rays defined by the corners of the base area are traversed through the MM-tree. Figure 2 shows the traversal through the tree in 2D. Only if the four rays at the corner take the same path, it is guaranteed, that all rays within the beam will also take this specific path. If the rays visit different nodes or hit different sides of the



(a) (b)

Fig. 4. (a) Bilinear interpolation introduces artifacts, such as discontinuation within the silhouette and at shadow edges, whereas bicubic interpolation (b) provides smoother transitions.

bounding volume, beam tracing needs to be replaced by traversing individual rays. The result of the beam tracing is depicted in Figure 3.

An exactly calculation of the intersection points between the rays and the surface patch of the heightfield is time consuming. Therefore, we use uniform and binary search [14] to get an approximate intersection point. First the ray is subdivided into uniform line segments. The uniform search determines that line segment which intersects the patch of the heightfield. Second the binary search computes the approximated intersection point within this segment. An advantage of this method is the abstraction from the real structure of the patch. The intersection test is based only on the height values (z) at given coordinates (x,y). Therefore, the generation of patches themselves can be encapsulated by operators of the modular extension unit.

3.2 Modular Extension Unit

The modular extension unit consists of a set of operators that allow to improve image quality on demand. In this paper, we suggest enhanced operators for surface generation and shading, but further operators can be easily added.

Surface Generation The ray-patch-intersection computed by the ray tracing kernel is based solely on height values. Now the surface patches are generated by enhanced operators. They can be composed to adjust this part of the rendering process.

Interpolation The surface patches of a heightfield are generated through interpolation by a specific operator of the extension unit. Commonly, the patch is bilinear interpolated between four neighbored height values. However, bilinear interpolation can introduce artifacts at the silhouette and shadow edges (Fig. 4(a)). Hence, a different operator can be used to provide a smoother bicubic interpolation. This results in less artifacts (Fig. 4(b)), but also increases rendering time. A midway solution provides an approximated bicubic interpolation. Here the

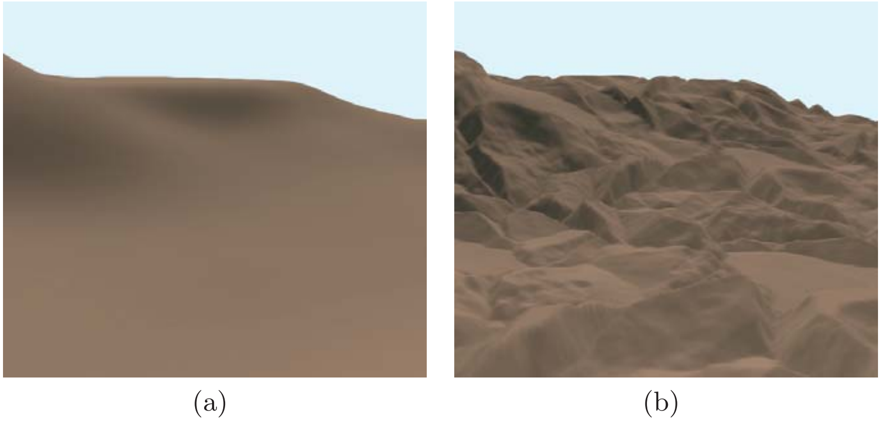

(a) (b)

Fig. 5. (a) The smooth interpolated surface of the heightfield lacks fine granular details. To increase realism, the surface can be enriched by microstructures that displace the height values along the z-axis (b).

surface is continuously subdivided by generating points through a bicubic function and is afterwards bilinear interpolated in between. This operator provides both, good quality and good performance.

Microstructures The resolution of heightfields is normally not sufficient to provide fine granular details. Hence microstructures are used to increase image quality. Different operators enrich a base heightfield by extra details. Either noise functions or additional micro-heightfields then describe the displacement along the z-axis (Fig. 5).

Antialiasing If only one ray per pixel is traversed, aliasing artifacts appear in the distance where the surface is under-sampled. Therefore, enhanced operators support antialiasing. On the one hand, an average mipmap allows for trilinear interpolation of the height values depending on the distance. This increases the memory footprint. Alternatively, multiple rays can be traversed per pixel. In this case, enhanced operators invoke supplementary ray tracing for surfaces in the distance. This increases rendering time.

Shading The ray tracing kernel assigns a material color to each visible point. To improve shading quality, the modular extension unit provides enhanced operators. A full recursive ray tracing is the most time consuming method. Simple texturing, e.g. with satellite images, however, may lead to low quality. Further, sophisticated illumination models, tailored to outdoor scenes, are supported. Especially Ambient Occlusion (AO) and Ambient Aperture Lighting (AAL) (cf. Section 2.1) apply well to terrain rendering. While AO is the fastest method, since only the quantity of self-occlusion is measured. AAL is more accurate and produces softer shadows, since the color and angle of incident light and even the diffuse scattered light of the sky are considered (Fig. 6). The higher quality leads to a higher rendering time. However, both techniques are based on preprocessing

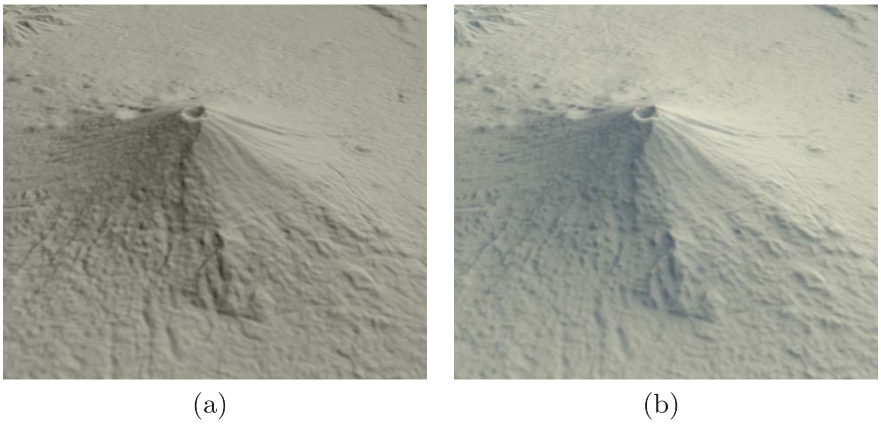


Fig. 6. Using (a) Ambient occlusion or (b) Ambient Aperture Lighting results in different levels of image quality. (Fuji region, elevation data source: ASTER GDEM, a product of METI and NASA.)

to reduce computation time during rendering. But this again increases memory consumption.

The described operators of the extension unit support the configuration of a ray tracer to different constrains. However, the introduced set of operators can easily be extended.

4 Functional Model

The functional model forms the interface between the flexible ray tracer and the dynamic execution model on the GPU. It is composed of individual tasks. Each task represents an operator of the ray tracing stage. The tasks are connected with regard to the processing flow.

They compose a network of sequential, parallel or recursive tasks. The basic configuration consists of a combination of ray traversal, intersection test and additional enhanced operators of the extension unit. The choice of enhanced operators determine the network topology. Recursive ray tracing, for instance, maps to a recursive network structure of the tasks, while the traversal of individual rays in parallel maps to a parallel structure.

To support the decision which enhanced operators should be used, we describe presets that either focus on rendering time, rendering quality or memory consumption. When rendering time is prioritized, simple bilinear interpolation and simple shading, for instance texturing, will be used. Whereas a quality-based configuration utilizes bicubic interpolation, adds details through microstructures and reduces artifacts in the distance through antialiasing. Moreover, high-quality shading can be activated, e.g. full recursive ray tracing. Memory-based set-ups, again, will omit precomputed shading operators and additional micro-heightfields to reduce memory consumption. When antialiasing is used, subsampling will be favored over additional average mipmaps.

The presets reflects a primary focus and define the primary functional model. However, varying data complexity and/or further constrains, such as ensuring minimal frame rates, might require adaptions of the functional model. Supporting dynamic insertion, removal or replacement of tasks on parallel hardware is a challenging issue. In the next section, we introduce a new task scheduling architecture that solves this problem.

5 Dynamic Task Scheduling

In this section, we present a formal model and a parallel implementation for scheduling the functional model (center of Fig. 1) on the GPU through parallel rewriting operations (bottom of Fig. 1). For this purpose, each invocation of a function is described as a task, whose dependencies are encoded into the stream.

5.1 Execution Model

The proposed execution model (Fig. 1) consists of a token stream, storing the current state of the functional program, and a set of rewriting rules, which are iteratively applied to modify the stream. In particular, we assume that the program is given as set of functions $F := \{f_1, \ldots, f_n\}$ and that the stream contains two different types of tokens to distinguish literal values from invocations.

Formally, a stream $s \in S$ can be described as a word from a language S with alphabet $\Sigma := \mathbb{Z} \cup F$, while each function f_i maps a tuple of n_i integers to m_i output tokens $f_i : \mathbb{Z}^{n_i} \to \Sigma^{m_i}$. The rewriting step is specified by a function $\mathit{rewrite} : S \to S$ that replaces the following pattern:

$$\langle a_1, \ldots, a_{n_i}, f_i \rangle \quad \text{with } a_1, \ldots, a_{n_i} \in \mathbb{Z}, f_i \in F$$

by the result of the invocation:

$$\langle r_1, \ldots, r_{m_i} \rangle \quad \text{with } (r_1, \ldots, r_{m_i}) := f_i(a_1, \ldots, a_{n_i})$$

In particular, an invocation pattern takes a list of literal arguments and a reference to the corresponding function $f_i \in F$. If the function token f_i is preceded by at least n_i arguments $(a_1, \ldots, a_{n_i})$, it is evaluated and replaces the original sub-stream. Hence, this scheme is equivalent to the post-order format also used in reverse Polish notations. Most important for a parallel GPU implementation, the rewriting affects only local regions of the stream and can be performed on different segments in parallel.

By starting with an initial stream s_0 and iteratively trying to replace invocations, we can construct a sequence of streams, whose limiting value can be considered as the final result:

$$\begin{aligned} s_{n+1} &:= \mathit{rewrite}(s_n) \\ \mathit{result}(s_0) &:= \lim_{n \to \infty} s_n \end{aligned}$$

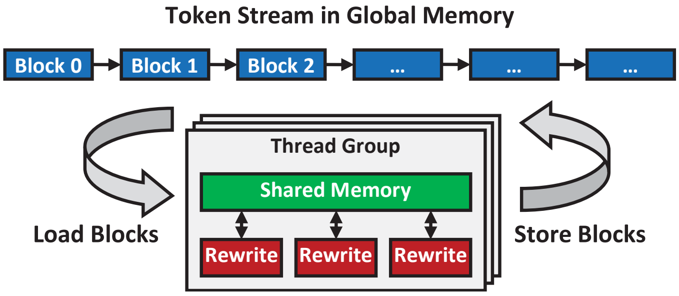


Fig. 7. Global scheduling of the segmented stream.

Due to the iterative rewriting, only at least one pattern must be replaced by the function *rewrite* in order to guarantee the monotony of this sequence. As a result, an implementation is not required to detect every pattern in the stream, so that it can be partitioned more freely for parallel rewriting. The following example illustrates the rewriting sequence for the expression $1 \cdot 2 + 3 \cdot 4$ with $F := \{f_1, f_2\}$, $f_1(x, y) := x \cdot y$ and $f_2(x, y) := x + y$:

$$\begin{aligned} s_0 &:= \langle \underbrace{1, 2, f_1}_{f_1(1,2)}, \underbrace{3, 4, f_1}_{f_1(3,4)}, f_2 \rangle \\ s_1 &:= \langle \underbrace{2, 12, f_2}_{f_2(2,12)} \rangle \\ s_2 &:= \langle 14 \rangle \end{aligned}$$

In iteration s_0 only the inner multiplications of f_1 can be evaluated in parallel, whereas f_2 waits for the intermediate result to become ready. Eventually, in the rewriting step from s_1 to s_2 the final sum is computed. In addition to literal values ($\mathbb{Z}$), also function tokens ($f_i \in F$) can be emitted to create recursive invocations.

Despite the simplicity of this model, which is entirely based on find-and-replace operations, an efficient GPU implementation has to solve several issues, which are discussed in the next two sections.

5.2 Parallel Implementation

According to the formal definition and the illustration in Fig. 1, the proposed algorithm can be parallelized by letting each core rewrite a different region of the stream. Also important, the partitioning of the stream into regions can be chosen almost arbitrarily. Further, we do not need to consider the contents of the stream because data dependencies are resolved by the pattern matching, and the model does not define an explicit execution order. Instead, control dependencies are also represented by local data dependencies. However, an efficient implementation also has to respect the architecture of modern GPUs, which are optimized for data parallel kernels and therefore require a large number of threads to reach optimal occupancy. In addition, threads are organized into groups, which are executed on the same processor and communicate via a small but fast shared memory. Since a function pattern creates and deletes a variable number of tokens in the stream, the length of the stream is continuously changing during the rewriting process, so that the data structure must be able to provide random access but also permit the fast insertion and removal of tokens.

As a consequence, the stream is partitioned into blocks of fixed size, which are stored as a linked list in global memory (Fig. 7). In correspondence to the two-level hierarchy of the graphics processor [31], we distinguish between the global scheduling of blocks at the system-level and the local rewriting of individual tokens, which is performed in the shared memory of each thread group. In particular, we utilize the concept of persistent threads, which run an infinite loop executing the following steps:

1. **Load Blocks** Depending on their size, one or two consecutive blocks are fetched from the stream and loaded into the shared memory of the multiprocessor. Due to the coherence of the memory access, the load operations can be coalesced to utilize the available bandwidth.
2. **Local Rewriting** The stream is rewritten locally and the results are stored in the shared memory (see Section 5.3). This process can be optionally repeated several times and requires multiple passes as well as random memory access.
3. **Store Blocks** The resulting tokens are written back into the global stream and up to three additional blocks are allocated. Also, the memory of empty blocks is released if necessary.

Most of these steps can be performed independently by several thread groups in parallel. Hence, especially the local rewriting but also the reading and writing of the stream can benefit from the parallel GPU architecture. However, the selection of a block in the linked list, the allocation, and the release of a block require exclusive access and must be protected by a global lock. The linked list provides the ability for fast insertion and removal of blocks, so that the memory layout of the stream is decoupled from its logical sequence, while the linearity within each block still facilitates coherent access of the global memory.

Formally, a block can be described as a tuple $(addr, next, size, active)$, storing the address of the tokens $(addr)$, a pointer to the next block $(next)$, the number of tokens in the block $(size)$, and a flag $(active)$ indicating whether the block is currently rewritten.

Each block has a fixed size in memory, so that the allocation and deallocation can be performed in constant time using a stack of free blocks. However, in order to compensate for varying lengths, the actual number of tokens in the block $(size)$ can vary during the rewriting process. Blocks, which are marked as *active*, are currently rewritten by a different thread group and must be ignored. In addition, we have to deal with the two cases of blocks causing *underflow* and *overflow* conditions. An *underflow* is reached if a block is too small, so that no valid pattern can be found, while an *overflow* occurs if the rewritten stream does not fit back into its original block. As a consequence, the system must be able to merge or split consecutive blocks on demand.

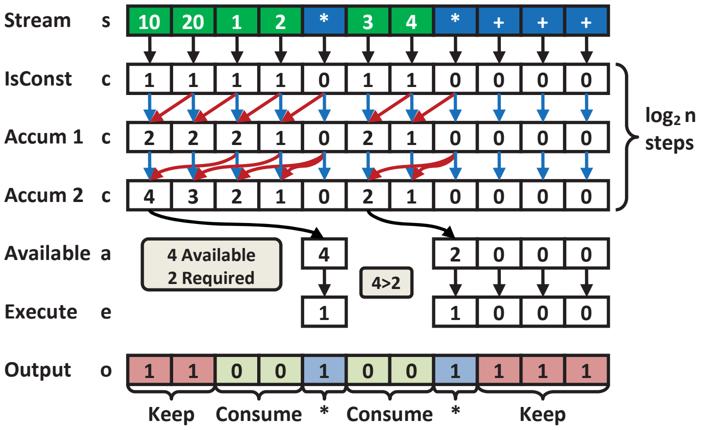


Fig. 8. Parallel decoding of the local stream.

In order to handle underflows, we always try to load two successive blocks as long as both fit into the shared memory. Therefore, small or empty blocks are automatically merged while the stream is rewritten. If the result can be stored back into a single block, the second one is released and removed from the linked list. On the other hand, if it becomes conceivable that the rewritten stream will not fit into the local memory, it is broken into four blocks. Since the next thread group loads at most two of these four blocks, a new overflow is less likely to happen and would subdivide the stream further until the results fit into the shared memory. Hence, the stream of blocks is expanded and reduced by several thread groups in parallel and out-of-order, so that compute-intensive regions of the stream do not delay the rewriting of faster blocks. In the next section, we will discuss the local rewriting in the shared memory.

5.3 Local Rewriting

Unlike the global rewriting process, which is partitioned across several thread groups, the local rewriting of a sub-stream must exploit the data parallelism of a single multiprocessor. Since each function consumes and emits a different number of tokens, a stream must be adjusted by copying it into a new array. In particular, for each of the two token types, we can identify two possible actions:

- **Literal values** are either removed from the stream, if the corresponding function is executed, or they are copied to the next iteration. Hence, a literal always creates one or zero outputs.
- **Function tokens** $f_i \in F$ either produce the specified number of m_i outputs if a sufficient number of literals are available or they are kept on the stream. Thus, a function token is rewritten into one or m_i output tokens.

Since the rewriting in shared memory should employ coherent control flow and data parallelism, it is restructured into three passes:

1. **Decode Stream** The stream is scanned for executable patterns and for each token, the number of outputs is computed according to the four cases above.

2. **Allocate Outputs** Depending on the number of outputs, the new position of each token is calculated using a prefix-sum.
3. **Execute Functions** The functions determined in the 1st step are executed and their results are stored at the positions computed in the 2nd step. Finally, the remaining tokens, which do not participate in an invocation, are copied to the output array.

Decode Stream The decoding pass is illustrated in Fig. 8 and assumes that the stream is given as a sequence of n tokens with $s := \langle t_1, \ldots, t_n \rangle$. In order to decide, if a literal $t_i \in \mathbb{Z}$ is an argument of an executable expression, the distance to the next succeeding function token in the stream is relevant. For this purpose, the number of literals c_i up to the next function token are counted. In particular, literals $t_i \in \mathbb{Z}$ are marked with $c_i = 1$, so that the second line of Fig. 8 (*IsConst*) contains a 1 for each literal and a 0 for each function token:

$$c_i := \begin{cases} 1 & \text{if } t_i \in \mathbb{Z} \\ 0 & \text{else} \end{cases}$$

Next, up to $log_2(n)$ iterative passes are required to converge c_i:

$$c_i := c_i + c_{i+c_i}$$

In this example, only two accumulation steps (*Accum 1, Accum 2*) are necessary to count up to four arguments. Hence, for each literal $t_i \in \mathbb{Z}$, the next function token can be found at position $i + c_i$ in the stream. When assuming that the stream s has a length of n tokens, the maximum number of preceding literals a_i of a token t_i can be computed as:

$$a_i := \max_{j \in [1,n]} \{c_j | c_j + j = i\}$$

As a result, an invocation $t_i = f_j$ with $f_j \in F$ is executable, which is indicated by $e_i = 1$, if the number of available arguments a_i are greater or equal than the number of required arguments n_j:

$$e_i := \begin{cases} 1 & \text{if } \exists j \in \mathbb{N} : (t_i = f_j) \wedge (a_i \geq n_j) \\ 0 & \text{else} \end{cases}$$

Hence, in the example shown by Fig. 8, only the multiplications are executable. Eventually, the number of generated outputs o_i of a token t_i is given by:

$$o_i := \begin{cases} 0 & \text{if } \exists j \in \mathbb{N} : (t_i \in \mathbb{Z}) \wedge (t_{i+c_i} = f_j) \\ & \wedge (e_{i+c_i} = 1) \wedge (c_i \leq n_j) \\ m_j & \text{if } \exists j \in \mathbb{N} : (t_i = f_j) \wedge (e_i = 1) \\ 1 & \text{else} \end{cases}$$

The first case checks if the literal t_i is consumed by the next successor function f_j, which requires the function to be executable ($e_{i+c_i} = 1$) and the literal to be within its argument list: $(c_i \leq n_j)$. Likewise, the second condition determines the execution of the current expression if it is a function $(t_i = f_j)$ and executable $(e_i = 1)$. Finally, the *else* branch corresponds to unused literals and unmatched functions, which are replicated and thus create exactly one output. In the presented example, each of the multiplications produces one result and their arguments are removed.

Allocate Outputs In order to compute the destination position of each token, the output size o_i is accumulated using a parallel prefix-sum [32]. If the resulting stream fits into the shared memory, it can be rewritten in the next step. Otherwise, it is sub-divided into four blocks and written back into global memory.

Execute Functions In the last pass, the previously collected functions are executed and their results are stored in the shared memory. Similarly, the remaining tokens, which do not participate in an invocation are copied to the resulting stream.

In the next section, we present an implementation of this technique.

6 Implementation

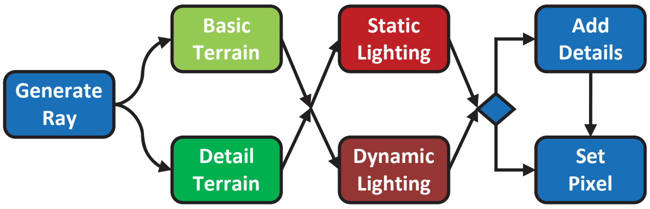


Fig. 9. Example function consisting of different stages for geometric and shading computations.

A prototype of our flexible architecture for terrain rendering has been evaluated on a GeForce GTX TITAN using CUDA 7.0. The rewriting process is started by a single kernel launch and performs the algorithm described in Section 5. For this purpose, the stream is divided into blocks of 512 tokens and each thread block stores at most two of them in the shared memory. In addition, there are 16 to 64 threads per thread block that decode the stream in parallel and execute the detected functions.

The structure of our example ray tracer is shown in Fig. 9 and consists of several stages which are described by the functional model (Section 4). In particular, we can switch between two detail levels of geometry and two lighting modes to vary quality, resource usage, and computation time. First, a fixed kernel generates the initial stream and creates a ray for each pixel. It either emits function calls to the '*Basic Terrain*' or '*Detail Terrain*' stages that compute the intersection point of the ray and the terrain through a combination of linear and

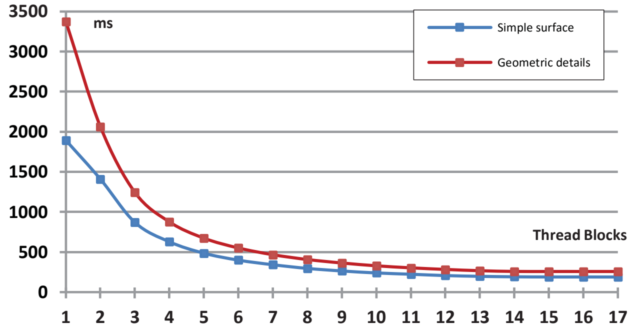


Fig. 10. Rendering time for different number of thread blocks and 16 threads per block on the GeForce GTX TITAN, which has 15 streaming multiprocessors.

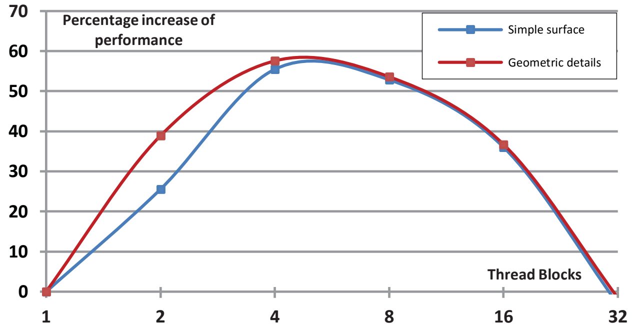


Fig. 11. Increase of performance when the number of thread blocks are doubled (16 threads per block).

binary search. Likewise, in the next stage, we can switch between pre-computed static or dynamic lighting models. In addition, the color is modulated by a microstructure and is interpolated for nearby pixels of the detail terrain ('*Add Detail*'). Finally, in the '*Set Pixel*' stage, the color data is written into the rendering target.

The ray tracer has been evaluated using four different heightmaps (*Terrain1*, *Terrain2*, *Fuji*, *Himalaya*) with a size of 512x512 samples and different configurations for geometry details and lighting models to show, in prinziple, feasability of our approach. Samples of the generated images shown in Fig. 12. For performance comparison, each test setup is used to draw 10 frames and the average rendering times per frames are listed in Table 1. Rendering times are nearly the same for all four data, since data size is equal. Also, the rendering times for static and dynamic lighting are comparable, so that this decision only affects the quality of the image. This would if more complex illumination models are used. In particular, it can be seen that the detailed rendering mode of all terrains requires more computational resources, but also creates more sophisticated images (Fig. 12 b).

Table 1. Rendering time for different configurations. Net Render Time is determined by the avg. render time without the system overhead (65ms) to handle stream rewriting

Terrain	Geometry	Lighting	Avg. Rendering Time	Net Render Time
Terrain1	Basic	Static	99.5 ms	34.5 ms
		Dynamic	101.1 ms	36.1 ms
	Detail	Static	136.5 ms	71.5 ms
		Dynamic	136.9 ms	71.9 ms
Terrain2	Basic	Static	98.8 ms	33.8 ms
		Dynamic	99.8 ms	34.8 ms
	Detail	Static	136.4 ms	71.4 ms
		Dynamic	136.3 ms	71.3 ms
Fuji	Basic	Static	99.6 ms	34.6 ms
		Dynamic	99.7 ms	34.7 ms
	Detail	Static	136.6 ms	71.6 ms
		Dynamic	137.0 ms	72.0 ms
Everest	Basic	Static	99.1 ms	34.1 ms
		Dynamic	100.0 ms	35.0 ms
	Detail	Static	136.4 ms	71.4 ms
		Dynamic	136.6 ms	71.6 ms

Tests have shown that a great portion of the rendering time ($\approx 65ms$) is consumed by the rewriting algorithm itself. Therefore additional tests to evaluate the core rewriting algorithm were performed. The scalability of the rewriting algorithm (Section 5) has been analyzed for the basic and detailed *Everest* terrain by varying the number of launched thread blocks. Since different thread blocks can run on distinct multiprocessors in parallel, a linear speed-up should be expected but there are two possible bottlenecks: First the linked list of blocks represents a global synchronization point and is protected by a mutex. However, each thread holds the lock only for a short amount of time. Second the stream is stored in global memory and must be copied into shared memory for rewriting. Though, it is accessed coherently, so that the available bandwidth can be maximized. As a result, the measurements indicate an continual decreasing render time, which stagnates at 15 thread blocks (Fig. 10). When the number of thread blocks is doubled, performance increases accordingly up to 16 blocks (Fig. 11) Since the GeForce GTX TITAN consists of 15 streaming multiprocessors (SMX) and similar tests on other graphic cards show the same coherence, we conclude that each thread block is mapped to a different multiprocessor and that each multiprocessor executes at most one thread block. Since on the Geforce GTX TITAN one thread block can execute at most 16 threads at once, for this test the number of threads was set to 16. However, the configuration resulting in the best performance, as seen in Table 1 utilized 8 SMX and 256 threads per block. This indicates that internally thread blocks and threads can be mapped differently depending on configuration and driver. Nevertheless our tests show the fundamental scalability of the parallel rewriting algorithm, but hardware limitation on graphic cards currently restricts scalability. Further improvements and optimiza-

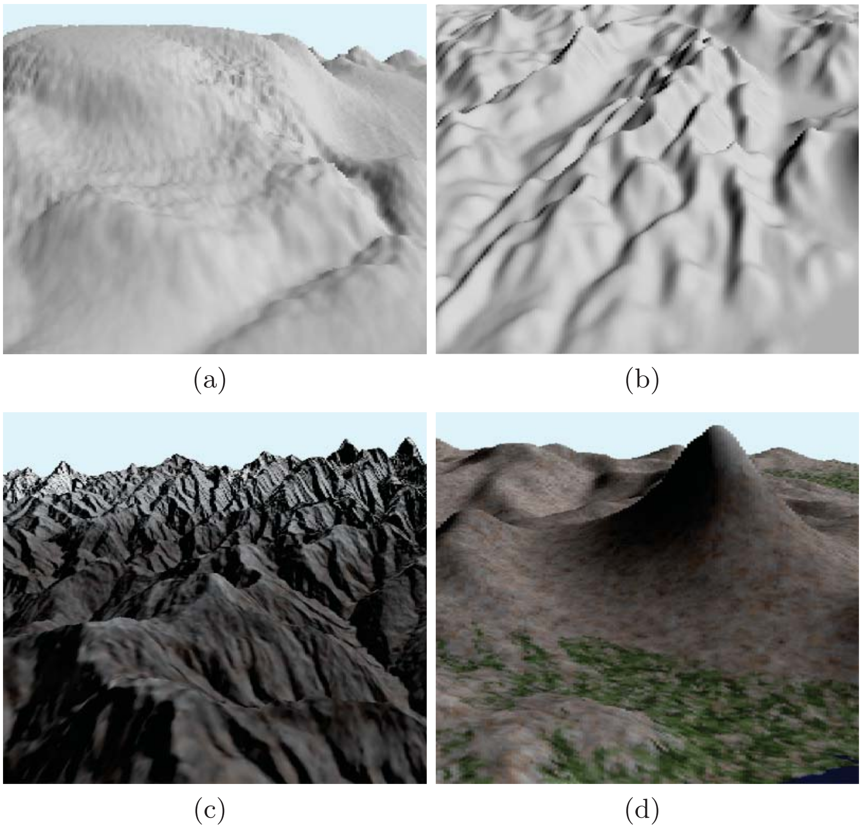

(a)
(b)
(c)
(d)

Fig. 12. Images rendered using our flexible terrain ray tracer. (Elevation data source: ASTER GDEM, a product of METI and NASA.)

tion should permit the execution of multiple thread blocks per multiprocessor to further scale our approach.

7 Conclusion

Balancing rendering time, rendering quality and resource consumption for ray tracing terrain surfaces is challenging. For this purpose, we presented a flexible ray tracing architecture. This is composed of a basic, high-efficient ray tracer and flexible, modular extensions to adjust the rendering process on demand with respect to time, quality and memory. To allow such a flexible approach to be mapped on the parallel hardware, we propose a novel execution model for scheduling dynamic workload on the GPU. A first prototype shows the feasibility. However, this is still subject to further development to increase the number of supported enhanced operators and to further facilitate the high parallelism of the GPU. Moreover the rewriting algorithm itself must be further optimized to the GPU to minimize execution time. Additionally, open questions still remain. Currently, manually composed presets determine the structure of the functional model. However, an automated adjustment by means of restrictions, e.g. minimum frame rate or minimum quality standards, has to be investigated. Moreover, how combinations of operators influence the total quality of an image, which is furthermore mostly subjective, is still subject of ongoing research. In the future, we will also investigate the possibility to support embedded visualization of data, such as movement and weather data. Since the functional model allows for freely adjusting the rendering process and the dynamic task scheduling can parallelize even strongly heterogeneous execution task, embedding data into the terrain in compliance with quality and performance constrains can be beneficial.

References

1. Ingo Wald, Solomon Boulos, and Peter Shirley. Ray tracing deformable scenes using dynamic bounding volume hierarchies. *ACM Trans. Graph.*, 26(1), January 2007.
2. László Szirmay-Kalos, Barnabás Aszódi, István Lazányi, and Mátyás Premecz. Approximate ray-tracing on the gpu with distance impostors. *Computer Graphics Forum*, 24(3):695–704, 2005.
3. Venkatraman Govindaraju, Peter Djeu, Karthikeyan Sankaralingam, Mary Vernon, and William R. Mark. Toward a multicore architecture for real-time ray-tracing. In *Proceedings of the 41st MICRO*, 2008.
4. Larry Seiler, Doug Carmean, Eric Sprangle, Tom Forsyth, Michael Abrash, Pradeep Dubey, Stephen Junkins, Adam Lake, Jeremy Sugerman, Robert Cavin, Roger Espasa, Ed Grochowski, Toni Juan, and Pat Hanrahan. Larrabee: A many-core x86 architecture for visual computing. In *ACM SIGGRAPH 2008 Papers*, SIGGRAPH '08, pages 18:1–18:15, New York, NY, USA, 2008. ACM.
5. Steven G Parker, James Bigler, Andreas Dietrich, Heiko Friedrich, Jared Hoberock, David Luebke, David McAllister, Morgan McGuire, Keith Morley, Austin Robison, et al. Optix: a general purpose ray tracing engine. *ACM TOG*, 29(4):66, 2010.
6. Markus Steinberger, Bernhard Kainz, Bernhard Kerbl, Stefan Hauswiesner, Michael Kenzel, and Dieter Schmalstieg. Softshell: dynamic scheduling on gpus. *ACM Trans. on Graphics*, 31(6), 2012.
7. Stanley Tzeng, Anjul Patney, and John D. Owens. Task management for irregular-parallel workloads on the gpu. In *High Performance Graphics*, HPG '10, 2010.
8. Markus Steinberger, Michael Kenzel, Pedro Boechat, Bernhard Kerbl, Mark Dokter, and Dieter Schmalstieg. Whippletree: Task-based scheduling of dynamic workloads on the gpu. *ACM Trans. Graph.*, 33(6):228:1–228:11, November 2014.
9. Ingo Wald and Philipp Slusallek. State of the art in interactive ray tracing. *State of the Art Reports, EUROGRAPHICS*, 2001.
10. Thiago Ize, Ingo Wald, Chelsea Robertson, and Steven G Parker. An evaluation of parallel grid construction for ray tracing dynamic scenes. In *Interactive Ray Tracing 2006, IEEE*, 2006.
11. Art Tevs, Ivo Ihrke, and Hans-Peter Seidel. Maximum mipmaps for fast, accurate, and scalable dynamic height field rendering. In *I3D '08*, 2008.
12. Ingo Wald, Philipp Slusallek, Carsten Benthin, and Markus Wagner. Interactive rendering with coherent ray tracing. In *Computer Graphics Forum*, pages 153–164, 2001.

13. John C Peterson and Michael B Porter. Ray/beam tracing for modeling the effects of ocean and platform dynamics. *Oceanic Engineering, IEEE Journal of*, 38(4):655–665, 2013.
14. Fábio Policarpo, Manuel M Oliveira, and João LD Comba. Real-time relief mapping on arbitrary polygonal surfaces. In *SI3D*, pages 155–162. ACM, 2005.
15. J Dummer. Cone step mapping: An iterative ray-heightfield intersection algorithm. *URL: http://www. lonesock. net/files/ConeStepMapping. pdf*, 2(3):4, 2006.
16. Fabio Policarpo and Manuel M Oliveira. Relaxed cone stepping for relief mapping. *GPU gems*, 3:409–428, 2007.
17. Matt Pharr and Simon Green. Ambient occlusion. *GPU Gems*, 1:279–292, 2004.
18. Christopher Oat and Pedro V Sander. Ambient aperture lighting. In *Proceedings of SI3D*, pages 61–64. ACM, 2007.
19. Won-Jong Lee, Youngsam Shin, Jaedon Lee, Jin-Woo Kim, Jae-Ho Nah, Seok-Yoon Jung, Shi-Hwa Lee, Hyun-Sang Park, and Tack-Don Han. SGRT: A mobile GPU architecture for real-time ray tracing. In *Proceedings of ACM High Performance Graphics 2009*, pages 109–119, 2013.
20. Jae-Ho Nah, Jeong-Soo Park, Chanmin Park, Jin-Woo Kim, Yun-Hye Jung, Woo-Chan Park, and Tack-Don Han. T&i engine: traversal and intersection engine for hardware accelerated ray tracing. In *ACM Transactions on Graphics (TOG)*, volume 30, page 160. ACM, 2011.
21. Timo Aila and Samuli Laine. Understanding the efficiency of ray traversal on gpus. In *High Performance Graphics 2009*, 2009.
22. Christian Dick, Jens Krüger, and Rüdiger Westermann. Gpu ray-casting for scalable terrain rendering. In *Proceedings of EUROGRAPHICS*, volume 50. Citeseer, 2009.
23. Samuli Laine, Tero Karras, and Timo Aila. Megakernels considered harmful: Wavefront path tracing on gpus. In *Proceedings of the 5th High-Performance Graphics Conference*, 2013.
24. Lars Middendorf, Christian Zebelein, and Christian Haubelt. Dynamic task mapping onto multi-core architectures through stream rewriting. In *SAMOS '13*, 2013.
25. Lars Middendorf and Ch Haubelt. A programmable graphics processor based on partial stream rewriting. In *Computer Graphics Forum*, volume 32, pages 325–334. Wiley Online Library, 2013.
26. Trevor L. McDonell, Manuel M.T. Chakravarty, Gabriele Keller, and Ben Lippmeier. Optimising purely functional gpu programs. *SIGPLAN Not.*, 48(9):49–60, September 2013.
27. Alexander Collins, Dominik Grewe, Vinod Grover, Sean Lee, and Adriana Susnea. Nova: A functional language for data parallelism. In *Proceedings of ACM SIGPLAN ARRAY'14*, 2014.
28. Conal Elliott. Programming graphics processors functionally. In *Proceedings of the 2004 Haskell Workshop*. ACM Press, 2004.
29. Chad Austin and Dirk Reiners. Renaissance: A functional shading language. In *Proceedings of Graphics Hardware*, 2005.
30. Tim Foley and Pat Hanrahan. Spark: modular, composable shaders for graphics hardware. In *ACM SIGGRAPH 2011*, 2011.
31. Erik Lindholm, John Nickolls, Stuart Oberman, and John Montrym. Nvidia tesla: A unified graphics and computing architecture. *IEEE MICRO*, 28(2):39–55, 2008.
32. Mark Harris, Shubhabrata Sengupta, and John D Owens. Parallel prefix sum (scan) with cuda. *GPU gems*, 3(39):851–876, 2007.

Visualization Techniques for 3D Facial Comparison

Katarína Furmanová

Masaryk University, Department of Computer Graphics and Design,
Faculty of Informatics, Brno, Czech Republic,
xfurman@fi.muni.cz

Abstract. Facial analysis and comparison form a substantial part of many research areas, such as security, medicine or psychology. 3D representations of facial images carry a lot of information beneficial to the researchers. However, comparing 3D shapes is not an easy task. Suitable visual representation of the data can help and simplify the work immensely. In this paper I therefore focus on visualization techniques for 3D facial data, specifically for its analysis and comparison. I present three different visualization methods suitable for pairwise comparison as well as for the analysis of large datasets. My methods target the typical tasks performed by the domain experts when conducting their research. The proposed visualizations were evaluated by experts working in the facial analysis area.

Key words: comparative visualization, facial analysis, nested surfaces, cross section, heat plot

1 Introduction

Many areas, where the facial analysis is used – such as criminal identification or authorization software – are nowadays quickly moving from 2D image to 3D representation. However, with higher dimensionality and complexity of the data also new challenges for its visualization appear. Researchers are posing different questions related to this topic, such as: How to visualize more than one facial surface without facing occlusions or losing track of data adherence? How to encode the measurements and visualize them to best convey their meaning? How to easily identify correlations between data?

The aim of this work is to deal with some of these challenges and present a complex visualization toolbox which could be used not only to visualize the results of the facial analysis but to aid the process itself. Depending on the purpose of facial analysis there are three main types of comparison – comparison of two facial models (e.g., for identity verification), comparison of one model with the entire dataset of models (e.g., for criminal identification) or comparison and analysis of an entire dataset of models (usually for research purposes). These three categories include also numerous subtasks, such as alignment verification, shape analysis, or variability analysis. To meet the demands posed by these tasks, I present three different visualization techniques, each of them targeting different subsets of these tasks. The evaluation of the results of my work was performed by a user study conducted among the domain experts in anthropology. This paper is based on the results of my diploma thesis [1].

2 Related Work

Research in the area of facial analysis is mostly focusing on technical aspects of a given task, such as the distance metrics and comparison algorithms. However, the area of 3D facial comparison can be considered a subfield of surface comparison. As such, there are already numerous techniques available. Some of them form the basis for my work.

According to Gleicher et al. [2], there are three main approaches to visual data comparison: juxtaposition, superposition, and explicit encoding. Juxtapositioning is scarcely used for 3D objects, as it is impractical and unintuitive, especially for objects that are very similar.

Superimposition is more suitable for detecting differences, however, for 3D shapes it tends to be too complex. Transparency plays an important role in this case – the proper level of opacity can improve the understandability of superimposed surfaces purely by modifying transparency values. The following examples modify the opacity of surfaces based on their geometric properties: Angle-based transparency [3], Normal variation transparency [4], or Geodesic fragment neighbors transparency [5], which also introduces surface contours to the image. Other techniques combine superimposition with explicit encoding and introduce features such as curvature glyphs [6], distance vectors, or fog simulation [7].

Another frequently used technique falling into the category of explicit encoding are the color maps. This method is often used as the default visualization method in many applications, including software tools for surface comparison [8], [9]. Related to color maps are also textures encoding additional information, e.g., stroke textures indicating curvature [10], [11].

Even with so many visual enhancements at hand, displaying big sets of 3D data all at once is ill-advised, due to the high complexity of images and a lot of visual clutter. A possible solution to this are cross-sectional views, an approach widely used in medical visualization for volumetric data – for example CT scan images – where a slice along a given plane is projected into 2D space [12]. A similar approach is the contouring of specific 3D object features followed by the projection of these contours into 2D space. This is often used when monitoring the temporal changes of a given feature, e.g., the width of a molecular tunnel [13].

Another example of data simplification by color encoding are heat plots and dense pixel displays [13–15]. In combination with interactive options such as thresholding, filtering, and data reorganization, they are very effective in discovering data relationships.

3 Proposed Visualization Techniques

I propose three techniques for visualizing facial models that have been designed specifically for facial comparison and analysis. In order to design the most suitable methods I conducted a study among anthropologists working in this field to identify the typical tasks performed during their work and the biggest drawbacks of the existing solutions they currently use. My visualization methods were crafted to provide solutions to three main detected drawbacks:

1. **The lack of shape information** The standard visualizations found in software used by anthropologists consist of color map variations. The color maps are typically mapped on one (primary) model from the processed dataset. Therefore, the shapes of the remaining models are omitted completely.
2. **The lack of local information** The color maps are computed on a global level (depending on the entire models and the entire dataset) and there is little or no possibility to limit the presented information only to a specific area and display it on a local scale.
3. **The limited view of data** The color maps provide one view of the acquired results. However, there are tasks where this technique is impractical as well as the numerical data that cannot be displayed in such way. Therefore, additional views and visualizations are needed.

3.1 Comparing Two Facial Models – Surface Superimposition

The first proposed technique serves for comparing two models. The superimposition principle has been selected for this task in order to address the first drawback (lack of shape information) and preserve the shape of both models. To illustrate the differences between two models properly, the following visual enhancements were added: surface splitting, fog simulation, shadow-casting glyphs, and intersection contours. This approach is based on the work of Busking et al. [7]. In the subsequent sections, the visual enhancements will be described in detail. Figure 1 shows an overview of the proposed techniques.

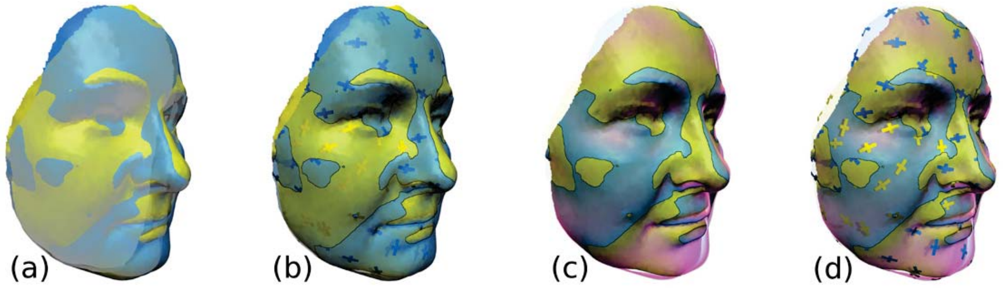


Fig. 1. Overview of the proposed visualization techniques. (a) Both surface models rendered with 50% opacity. (b) Opaque inner surface, transparent outer surface with shadow casting glyphs and intersection contours. (c) Simulation of fog between surfaces. (d) Combination of (b) and (c).

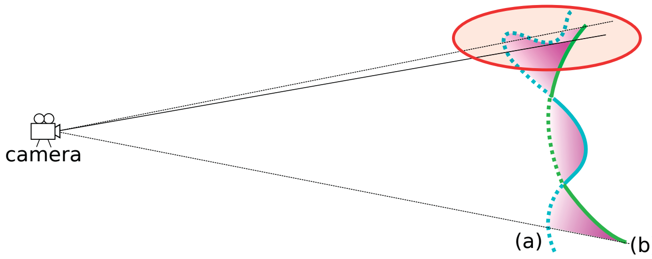


Fig. 2. Surface transparency and fog simulation scheme. The dotted parts of models (a) and (b) are considered *outer* and are rendered transparently, while the *inner* parts (solid line segments) are rendered opaque. The intensity of fog (pink color) depends on the distance between surfaces along the viewing direction. The area highlighted by red ellipse shows a special case when the second surface along the viewing ray belongs to the same model as the first one, and thus it is classified as the *outer* as well.

Surface Splitting One of the popular techniques for improving the understandability of transparent surfaces is the modulation of transparency values based on the placement of the surface according to other surfaces. A variation of this technique suitable for cases of pairwise comparison splits the surfaces into outer parts (parts of the surface which are the closest to the camera) and inner parts, which are hidden behind the outer parts. This classification of surfaces takes place in an image space, which allows easy handling of special cases such as the one highlighted in Figure 2.

Fog Simulation The modulation of transparency, although beneficial, is not particularly helpful for conveying the distance between the surfaces. As a visual clue for this task I come with two techniques. First of them is the fog simulation. The aim of this technique is to simulate a partially transparent volume – fog. Its color has to be different from the colors of the models. The fog fills the space between the two surfaces (see Figure 2). The limitation of the real case scenario is that the fog is accumulated along the viewing ray – the result is therefore view dependent. Another problem is that the outer surface needs to be nearly completely transparent when we want the fog to be visible. To deal with this issue, I devised three different fog simulation methods based on the real case scenario (see Figure 3):

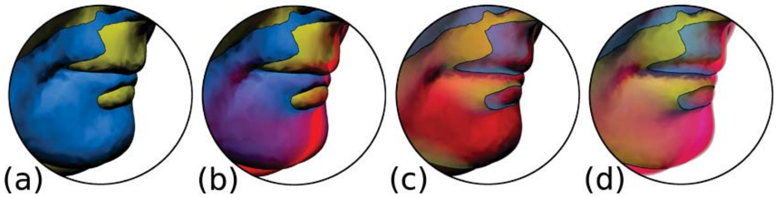


Fig. 3. Different fog simulation techniques. (a) Models rendered with full opacity. (b) Color overlay – notice the illusion that the blue surface lies behind the red one. (c) Transparency mapping on the outer surface. (d) Color mapping on the inner surface.

- **Color overlay.** This method modifies the color of the outer surface based on the distance between the surfaces – the distance serves as the ratio between the original color of the surface and the color of the fog. However, it might create misleading illusions about the surface adherence to models.
- **Transparency mapping on outer surface.** With this method the entire outer surface is colored by the color of the fog. The distance is then mapped onto the opacity values of the outer surface – the bigger the distance, the higher the opacity. This method yields nice visual results, but the interpretability of the surface adherence to models is reduced by coloring the entire outer layer by one color.
- **Color mapping on inner surface.** This method modifies the color of the surface similarly to the first method. It mixes the color of the model with the fog according to the distance – only this time the color is mapped onto the inner layers.

Shadow-Casting Curvature Glyphs According to several studies [16], [17], shadows aid the human perception of depth and shape. Therefore, as the second method dedicated to improving the interpretability of the distances between surfaces I incorporate the shadow-casting glyphs, a method based on work of Itterante et al. [11] and Weigle et al. [6]. These glyphs are mapped onto the outer parts of the surfaces and cast shadows on the inner surfaces. The light source position is fixed with respect to the models, so when the users rotate the scene, they can explore the shadows from various angles. The glyphs are evenly distributed across the surface. The color of the glyphs matches the color of the surface so they are only visible when the outer surface is not fully opaque. The glyph shape is derived from [6]. It is a plus sign of constant size elongated in one direction which depicts the maximal principal curvature at the center of the glyph.

Intersection Contours The last enhancement technique is the contouring of surface intersections. When using the transparency and glyphs, sometimes the intersections of surfaces are not very prominent or are not visible at all (see Figure 4). On the other hand, the bump edges and occlusions may be wrongly interpreted as intersections.

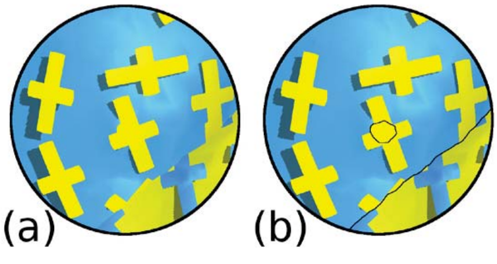


Fig. 4. Example of situation where the small intersection is hidden due to glyph placement. (a) Without intersection contours. (b) With intersection contours.

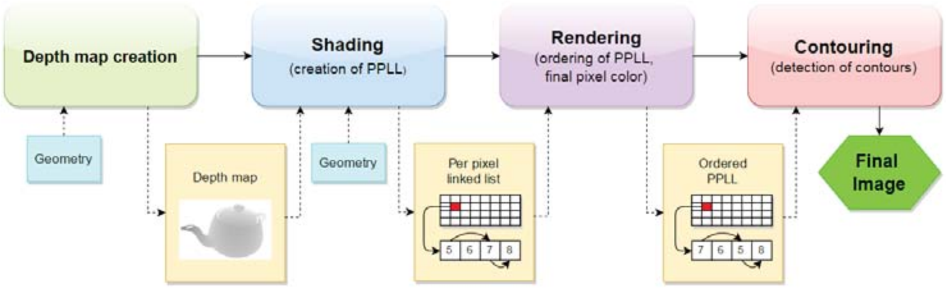


Fig. 5. Proposed rendering pipeline.

For the implementation of the surface superimposition-based techniques I designed a rendering pipeline displayed in Figure 5. The pipeline consists of four basic steps. After the creation of the depth map necessary for shadow-casting glyphs and glyph placement, the shading of fragments, creation of linked lists, and glyph placement follows. The rendering phase consists of ordering the linked lists, color and opacity modulations for visualization, and computation of final color per pixel. The last contouring phase serves for the detection and rendering of the intersection contours.

3.2 Comparing Many Facial Models – Cross Sections

The above described methods are not applicable in situations when processing of large datasets is necessary. Here preserving the information about the shape variation, especially the local shape variation displayed on a local scale, is desirable.

When dealing with such complex data, the projection, or rather reduction of the 3D data into 2D space is a popular approach. The cross section method I propose here is inspired by the technique typically used for visualization of volumetric data. A slicing plane is used, its intersection with 3D dataset is computed and then displayed in 2D.

In my case I assume that all models in the dataset are spatially aligned and that one facial model is selected as primary – typically it is the averaged model of the dataset. The primary model is displayed in 3D space along with the slicing plane (Figure 6 (a)). The user can move and rotate the slicing plane in the space of the primary model to get the desired intersection position. Then, the intersections of the plane with every model in the dataset are computed. The intersection with the primary model (primary intersection) is sampled and the variability at each sample point is determined.

There are three main options for what the user can display:

- **Intersections with all faces** (Figure 6(b)). It enables to observe how well the models are aligned, especially with interactive manipulation with the slicing plane. However, with the increasing number of models the interpretability of the final image decreases.

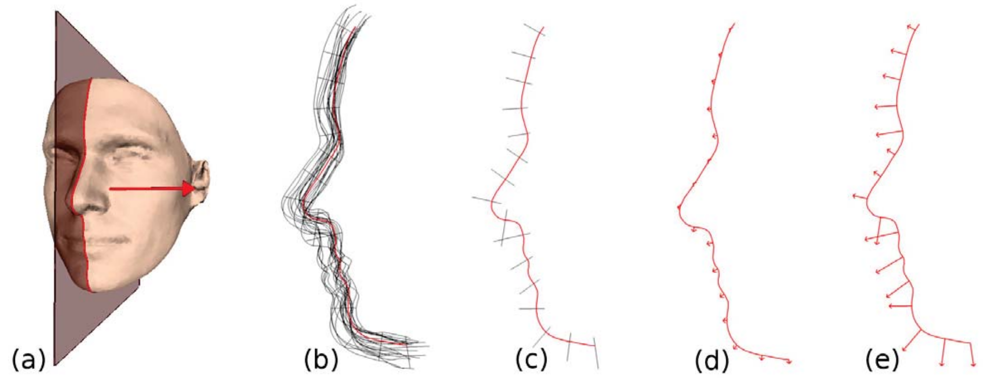


Fig. 6. Cross Sections. (a) Reference picture of the primary face with the slicing plane specifying the cross-sectional slice. (b) Red – intersection of the slicing plane with the average face. Black – intersections with all faces in the dataset. (c) Visualization of the distance span. (d) Vectors indicating the average distance. (e) Same as (d) with enhanced vector sizes.

- **Distance span** (Figure 6(c)). It indicates the interval of distances from a given sampling point to the intersection curves of each model from the dataset in the direction of the normal vector to the primary intersection curve at the sampling point.
- **Average distance** (Figure 6(d,e)). It shows the average distance from a given sampling point to the intersection curves of each model from the dataset, again using the normal direction.

3.3 Plotting Numerical Results

The last proposed method displays the numerical results computed during the analysis of datasets with more than hundred models. Heat plot is a typical way of displaying large sets of numerical data with color encoding. By filtering and reordering of the data, correlations may be discovered more easily than by exploring the numerical values.

Here, I propose two versions of heat plots, one showing results from the pairwise comparison of models in one dataset and one focusing on detailed results of the comparison of one model with a given dataset.

Pairwise Comparison The results from the pairwise comparison of models in one dataset consist of a table of $n \times n$ values representing the measurement between each pair of models in the dataset consisting of n faces. These measurements can represent the maximal or minimal distance between the two faces, variance, geometric mean, etc. These values are displayed in the heat plot represented by the $n \times n$ matrix where each matrix cell shows the measurement between two models (Figure 7). The users can filter the lowest and/or the highest values on an interactive scale. As an additional feature, a histogram illustrating the distribution and variability of the values in the dataset may be displayed.

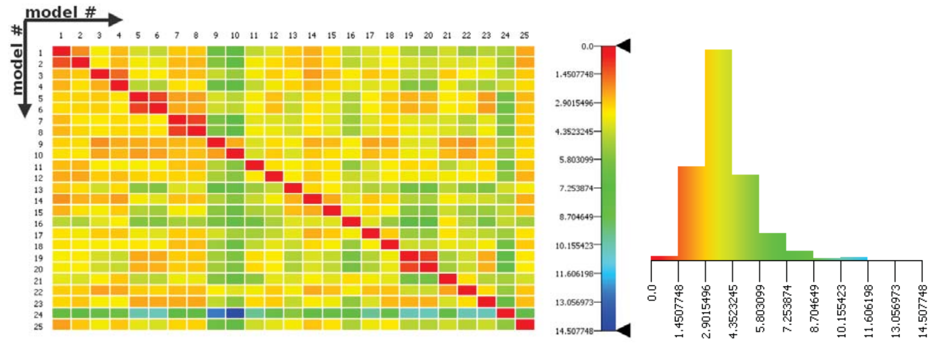


Fig. 7. Heat plot for visualization of pairwise comparison results and the accompanying histogram.

Auxiliary Results In cases when one model is compared with a given dataset, for each vertex of this primary model the distances to the closest vertex on each model in the dataset is computed. The auxiliary results for model M consisting of m vertices and the dataset consisting of n models would contain $m \times n$ values. One row of the heat plot then depicts the distances from vertices of the primary model to the nearest vertices of one model in the dataset. A vertical slice at position x then represents the distances from the x-th vertex of the primary model to the nearest vertex of each model from the dataset.

The models typically contain thousands of vertices. In case the values do not fit onto the screen, the neighboring vertices are aggregated, their values averaged and the zooming window is added. The individual values of the area covered by the zooming window are displayed under the heat plot – each value as a vertical line segment of one pixel width (see Figure 8).

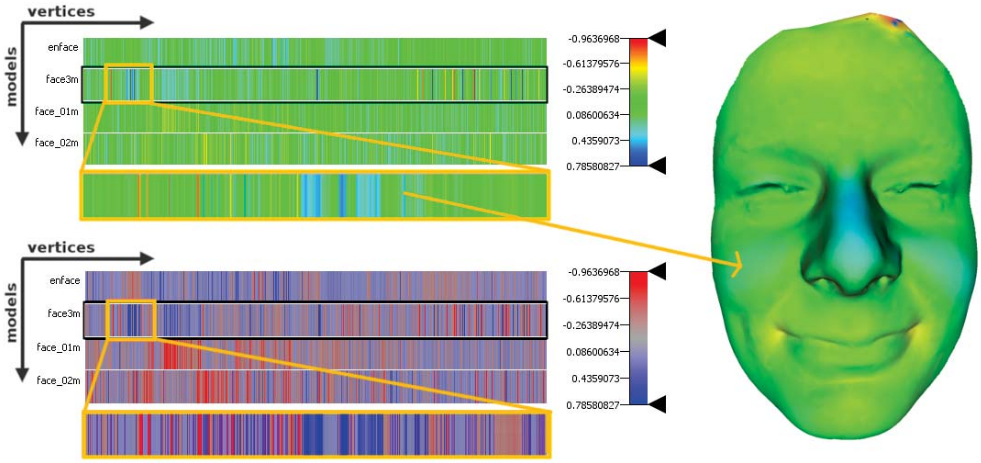


Fig. 8. Two color scheme variants of the heat plot for auxiliary results of the average face computed from the dataset of four models. Each row depicts the distances between the vertices of the average face and one model of the dataset. The orange highlight shows the zooming window and the detailed view of individual values at the bottom.

4 Results and Evaluation

To evaluate the usability of my visualization techniques I conducted a user study with four scientists working in a field of facial morphometry and analysis. The researchers were asked to fill out a questionnaire in which they assessed the selected techniques and compare them to the standard color maps present in the currently available tools. The questionnaire consisted of four parts.

Surface Superimposition In this part eight selected combinations of techniques were presented (see Figure 9): (a) standard color map, (b) shadow-casting curvature glyphs, (c–e) different fog simulation techniques, (f–h) shadow-casting curvature glyphs combined with fog simulation techniques. For each of these techniques the users were asked how well the visualizations convey the shapes and differences between the surface models.

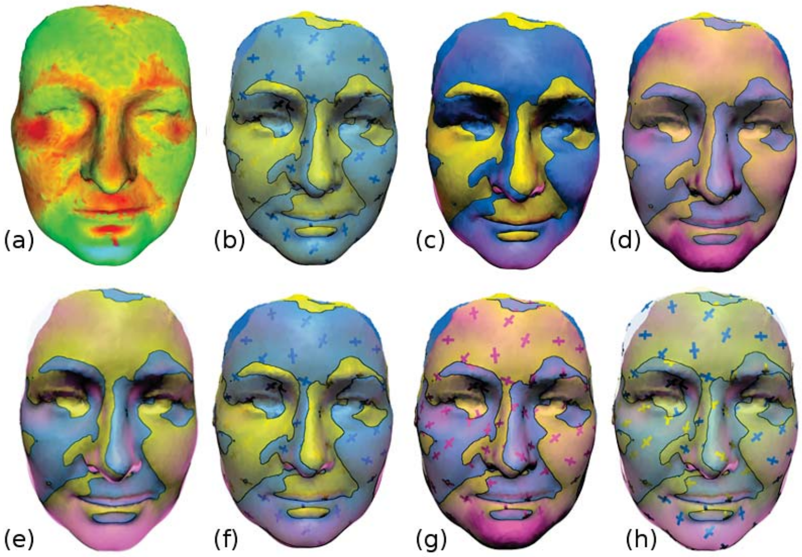


Fig. 9. Eight visualizations presented for evaluation.

The averaged results of their evaluation can be seen in Figure 10. The results show that while the standard visualization (a) – ranks high in showing the differences, it ranks low on conveying shapes in comparison with other methods. On the other hand, the newly proposed methods, the visualizations (b) – shadow-casting glyphs and (h) – the combination of shadow-casting glyphs and color mapping on inner surface rank high in both case, balancing both – need for illustration of the shape and imparting the differences between models.

Cross Sections The scientists were asked to evaluate the contribution of the cross-sectional slices to the variability readability in a set of models. The cross section based technique was proclaimed fairly demonstrative for interpreting local variability and with respect to this task placed above the standard visualization.

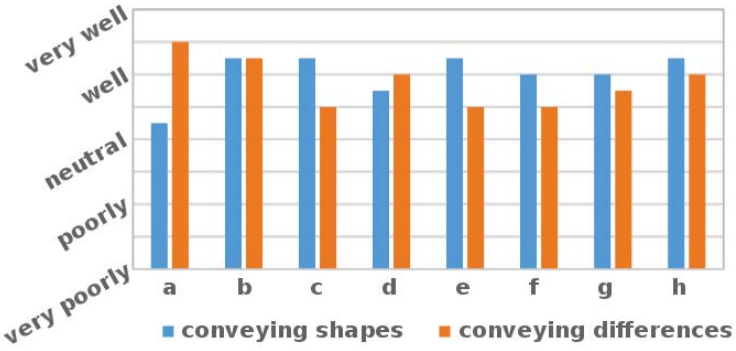


Fig. 10. Averaged evaluation of surface superimposition visualization techniques.

Plots The third part aimed to evaluate the heat plots with respect to conveying the variability in a set of models and the contribution to the analysis of facial dataset. The visualization of pairwise comparison results ranked high in both aspects. As for detailed results, the scientists found it less demonstrative.

Survey Summary In the last part of the questionnaire I asked the respondents to choose the best suited visualization method for several tasks to see if and where my methods were usable. For verifying the alignment of models and analyzing the shape of models the two selected visualizations were surface superimposition method with shadow-casting glyphs and cross-sectional slices. For the task of analyzing local variability the preferred methods were color map on model, cross-sectional slices, and heat plot for pairwise comparison results. In case of analyzing a set of models all scientists uniquely settled on heat plot for pairwise comparison results.

5 Conclusion

The aim of my work was to design several visualization methods for facial comparison on different levels of data sizes. I analyzed the needs of scientists working in this area and designed three different visualization approaches. After implementing these techniques I conducted a user study with the scientists to evaluate the contribution of my visualizations to their work. The results revealed the most contributory representations.

The evaluation revealed possible extensions for the future work. It was suggested that the light position should not be fixed with respect to orientation of the models in order to achieve moving shadows when the models are rotated. It was also noted that the fog simulation would be more beneficial if it were view independent. Finally, concerning the cross sections, it was suggested to add the option of displaying absolute variability values (as opposed to currently used relative, which take into account orientation of vectors).

References

1. Furmanová, K.: Visualization techniques for 3D facial comparison. Masaryk University (2015).
2. Gleicher, M., Albers, D., Walker, R., Jusufi, I., Hansen, C., Roberts, J.: Visual comparison for information visualization. Information Visualization. 10, 4, pp. 289–309 (2011).
3. Hummel, M., Garth, C., Hamann, B., Hagen, H., Joy, K.: IRIS: Illustrative Rendering for Integral Surfaces. IEEE Transactions on Visualization and Computer Graphics. 16, 6, pp. 1319–1328 (2010).
4. Born, S., Wiebel, A., Friedrich, J., Scheuermann, G., Bartz, D.: Illustrative Stream Surfaces. IEEE Transactions on Visualization and Computer Graphics. 16, 6, pp. 1329–1338 (2010).
5. Carnecky, R., Fuchs, R., Mehl, S., Jang, Y., Peikert, R.: Smart Transparency for Illustrative Visualization of Complex Flow Surfaces. IEEE Transactions on Visualization and Computer Graphics. 19, 5, pp. 838–851 (2013).
6. Weigle, C., Taylor, R.: Visualizing intersecting surfaces with nested-surface techniques. In: Visualization, 2005. VIS 05. IEEE. pp. 503–510. (2005).
7. Busking, S., Botha, C., Ferrarini, L., Milles, J., Post, F.: Image-based rendering of intersecting surfaces for dynamic comparative visualization. The Visual Computer. 27, 5, pp. 347–363 (2010).
8. CloudCompare (version 2.6.2). EDF R&D, Telecom ParisTech (2015).
9. Schmidt, J., Preiner, R., Auzinger, T., Wimmer, M., Grőller, M., Bruckner, S.: YMCA – Your Mesh Comparison Application. In: Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on. IEEE. pp. 153-162. (2015)
10. Diewald, U., Preusser, T., Rumpf, M.: Anisotropic diffusion in vector field visualization on Euclidean domains and surfaces. IEEE Transactions on Visualization and Computer Graphics. 6, 2, pp. 139–149 (2000).
11. Interrante, V., Fuchs, H., Pizer, S.: Conveying the 3D shape of smoothly curving transparent surfaces via texture. IEEE Transactions on Visualization and Computer Graphics. 3, 2, pp. 98–117 (1997).
12. Friese, K., Blanke, P., Wolter, F.: YaDiV–an open platform for 3D visualization and 3D segmentation of medical data. The Visual Computer. 27, 2, pp. 129–139 (2010).
13. Byška, J., Jurčík, A., Gröller, M., Viola, I., Kozlíková, B.: MoleCollar and Tunnel Heat Map Visualizations for Conveying Spatio-Temporo-Chemical Properties Across and Along Protein Voids. Computer Graphics Forum. 34, 3, pp. 1–10 (2015).
14. Ivanisevic, J., Benton, H., Rinehart, D., Epstein, A., Kurczy, M., Boska, M., Gendelman, H., Siuzdak, G.: An interactive cluster heat map to visualize and explore multidimensional metabolomic data. Metabolomics. 11, 4, pp. 1029–1034 (2014).
15. Zhai, Y., Huang, X., Chang, X.: Combining least absolute shrinkage and selection operator (LASSO) and heat map visualization for biomarkers detection of LGL leukemia. In: Systems and Information Engineering Design Symposium (SIEDS). pp. 165–170 (2015).
16. Erens, R., Kappers, A., Koenderink, J.: Perception of local shape from shading. Perception & Psychophysics. 54, 2, pp. 145–156 (1993).
17. Mamassian, P., Knill, D., Kersten, D.: The perception of cast shadows. Trends in cognitive sciences. 2, 8, pp. 288–295 (1998).

Interactive Data Visualization for Second Screen Applications: State of the Art and Technical Challenges

Kerstin Blumenstein[1], Markus Wagner[1], Wolfgang Aigner[1],
Rosa von Suess[1], Harald Prochaska[1], Julia Püringer[1],
Matthias Zeppelzauer[1], and Michael Sedlmair[2]

[1] St. Poelten University of Applied Sciences, Austria,
[first].[last]@fhstp.ac.at
[2] University of Vienna, Austria,
[first].[last]@univie.ac.at

Abstract. While second screen scenarios - that is, simultaneously using a phone, tablet or laptop while watching TV or a recorded broadcast - are finding their ways into the homes of millions of people, our understanding of how to properly design them is still very limited. We envision this design space and investigate how interactive data visualization can be leveraged in a second screen context.
We concentrate on the state of the art in the affected areas of this topic and define technical challenges and opportunities which have to be solved for developing second screen applications including data visualization in the future.

Key words: Information visualization, second screen, multi screen, mobile device, touch

1 Introduction

With the continuous proliferation of accessible computational devices, the media consumption behavior of millions of people is significantly changing. While traditionally medial content was consumed with one device at a time, multi device setups become more and more common. One specific instance of a multi device setup are second screen (2^{nd} screen) scenarios in which a secondary device is used to access information while simultaneously watching television or a recorded broadcast on a large screen. While many studies show the rapid increase of 2^{nd} screen usage [26, 36, 37], dedicated applications for it are still in its infancy and very little is known on how to properly design them.

Often numbers, data and graphics are used in broadcasts. Because of limited time, editors have to reduce those data and cannot give an extended description of the content. Data visualization can help here to provide an easy to understand detailed description of the content [40]. Therefore, integrating interactive data visualization in a 2^{nd} screen application seems to be a promising approach.

Target devices for 2^{nd} screen applications are mainly laptop, smartphone and tablet [37]. Because of increasing sales figures for tablets and smartphones and decreasing sales figures for laptops, which are projected by the International Data Corporation (IDC) [20] until 2018, the focus in this research is on mobile touch devices like tablets and smartphones.

We will give an overview of the state of the art in the affected areas of developing 2^{nd} screen applications with visualization. This endeavor comprises related aspects from different disciplines in computer and media science. We therefore will take into account aspects not only from the still small research field on 2^{nd} screen applications, but also from a technical perspective (interactive visualization on mobile touch interfaces, multi device environments and their synchronization) and from a content perspective (TV formats). Afterwards we describe technical challenges which have to be solved to develop visualization for 2^{nd} screen applications as well as opportunities for such scenarios.

2 State of the Art

To provide a broad overview about the topic of 2^{nd} screen applications with visualization we investigate the following areas: (2.1) Interactive TV & 2^{nd} Screen, (2.2) TV Formats for 2^{nd} Screen Scenarios, (2.3) Visualization on Touch Screens, (2.4) Multi Screen Environments and (2.5) Device Synchronization.

2.1 Interactive TV & 2^{nd} Screen

Since smartphones and tablets have appeared on the market, the behavior of watching TV has changed [11]. Obrist et al. [29] emphasize that: *"Television still plays an important role in everyday life, but the way we consume and interact with TV content has changed dramatically."* A survey by ARD/ZDF (German public-service broadcasters) [6] found that 56.6% of TV users also access online content via 2^{nd} screen devices simultaneously to the TV, supporting the statement of Proulx and Shepatin [32] that *"The internet didn't kill TV! It has become its best friend"*.

With the proliferation of such 2^{nd} screen scenarios, research in the field of TV is now increasingly focusing on human-computer interaction in the sense of developing new interaction concepts for domestic environments [17]. However, one of the major challenges is that the audience switches attention between a TV and one or multiple mobile devices. The recent experience of watching TV is far beyond the 'lean back and do nothing' ethos from the past, but it's challenging to heel the audience to action. The recent trend is not to create an alternative to watching TV, which might distract the users from TV's content, but to support the users' immersion and the program's enhancement by using additional information about the content of the TV broadcast and about user-generated content through back channel solutions [12]. While 2^{nd} screen applications clearly open up a whole new space of possibilities, they are still heavily underutilized [11]. Although broadcasters recognized the potential and have started to provide dedicated 2^{nd} screen applications, the knowledge about what works is still limited [16]. Based on an analysis of Twitter messages during a live broadcast, Lochrie and Coulton [24] found out that smartphones are heavily used as 2^{nd} screens, but that the audience mostly create their own forums for inter-audience interaction using (social media) platforms such as Twitter or Facebook that are disconnected from the primary content channel.

Bubble-TV (see Figure 1) is one of the few existing examples of an innovative solution that embeds Twitter feeds as dynamic visualization in the background of a TV studio during live discussions [18]. Bubble-TV goes far beyond showing single tweets as the audience makes decisions and can intervene immediately at several points of the show.

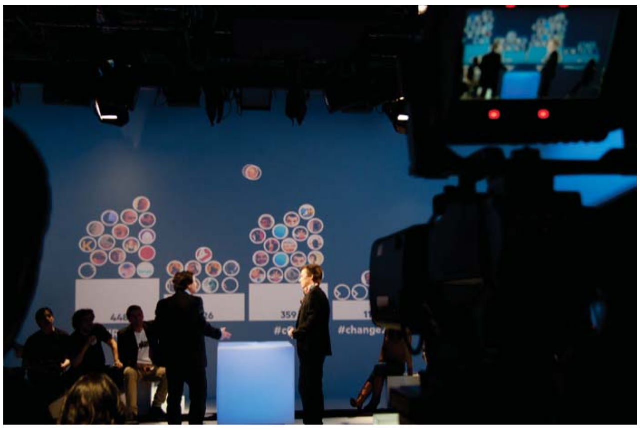


Fig. 1: Bubble-TV: Dynamic visualization in the back of a TV stage [18].

In their survey 'In Front of and Behind the Second Screen', Geerts et al. [16] defined five critical success factors of a 2^{nd} screen application: ease of use, timing or live synchronization, social interaction, attention and added value. In addition, Obrist et al. [29] came up with the four key areas of research in interactive television: content, recommendations, device ecosystems and user feedback.

2.2 TV Formats for 2^{nd} Screen Scenarios

Interaction and the supply of information via several channels are key components for successful future television. There is a wide range of usage possibilities of these features depending on the format and topic of the content [6]. In a survey conducted at St. Pölten University of Applied Sciences [38], annual data about the state of the art usage of 2^{nd} screen applications in the field of information television was investigated. The results showed that there are five layers of interaction in TV programs that reflect how intense additional applications can be used to enrich a broadcasted show: (1) *social media platforms* offer a space to share opinions and discuss the TV broadcast independently in real-time; (2) *moderated social media* where the broadcaster is part of the social media content production; (3) *responded social web activities* where viewers can intervene/influence first screen (1^{st} screen) content via back channels; (4) *cross media storytelling* where the user has several options to follow the story and multiple platforms offer additional content and information; (5) *user-generated content* where users themselves are contributing material.

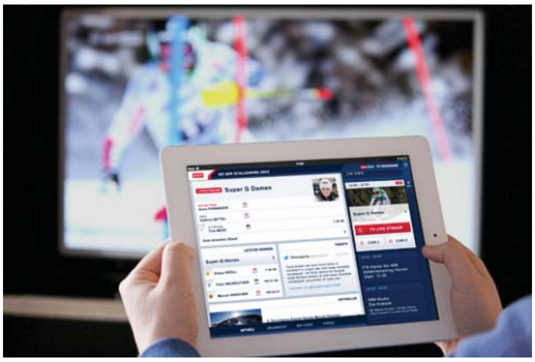


Fig. 2: 2^{nd} screen application published by the Austrian public broadcaster ORF for the skiing world championship 2013[1].

Following the half year report of Goldmedia Custom Research's TV Monitor [1] the late night show 'Circus Halligalli' counts the most web and social media activities on the German TV market. Big sports events and in general all kinds of live events achieve a high level of interaction on social media platforms. During the skiing world championship 2013 in Schladming, the Austrian public broadcaster ORF offered a successful 2^{nd} screen application (see Figure 2). The user was able to switch between several additional camera angles, an instant live standing was available and background stories were offered. Further a strong social media support was provided[2].

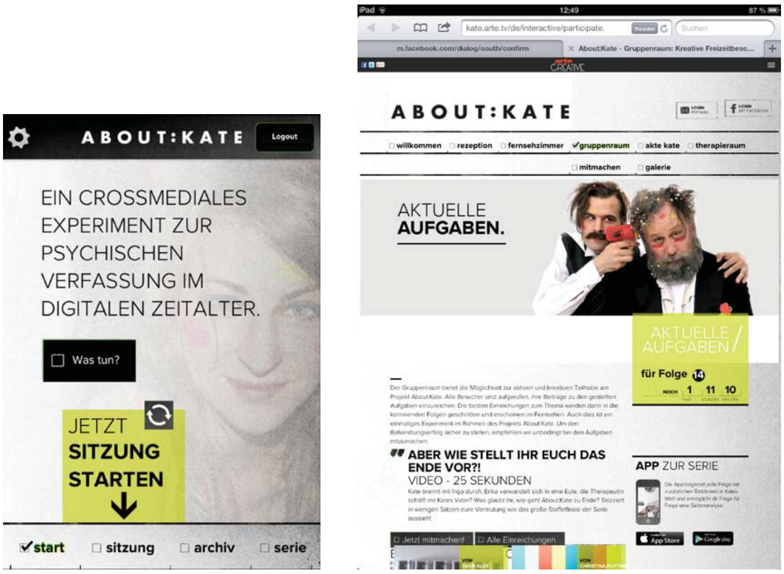
According to Würbel et al. [43], the creation of a nonlinear multi stream video show in real-time, which changes to the interests of the consumer instantly, is the future of interactive TV. They have tested such a concept with 489 users during the Olympic games in Beijing in 2008. During this test, explicit and implicit feedback has been collected and analyzed to adapt the program to the audience's needs.

Apart from these examples for specific events, there are a number of notable examples of TV formats that offer innovative solutions of integrating 2^{nd} screen applications. 'About:Kate'[3] for instance, is an innovative TV series, produced by ulmen.tv on behalf of ZDF/ARTE. This series is state of the art in the field of cross media storytelling with the usage of different 2^{nd} screen applications like a smartphone app and a web platform to upload user-generated video clips

[1] `http://bit.ly/1i92xFa`, accessed September, 2015.

[2] `http://bit.ly/1i92xFa`, accessed January, 2015.

[3] `http://bit.ly/1iod8wH`, accessed December, 2014.



(a) Home screen of the smartphone app where you can start the session.

(b) Home screen of the website opened on an iPad.

Fig. 3: Cross media storytelling for the TV series 'About:Kate'.

related to the narrative of the TV series[4] (see Figure 3). The production company created a virtual character called Kate that users follow via TV and different social media platforms. Users can watch Kate's blog and, via the smartphone application, Kate will randomly call them during the live broadcast.

Another example for 2^{nd} screen applications is the 'Red Bull Signature Series' that Red Bull produces in collaboration with Shazam and the US TV station NBC[5]. This format includes a snowboard live broadcast and a 2^{nd} screen application. The user can watch the sport event from another point of view, for instance the ego perspective of the world's most progressive riders, synced to the live image of the TV set. The synchronization is performed via the audio signal of the TV broadcast. Shazam also connects automatically to social media platforms. The viewer is able to follow all the riders and the event organization during the event[6].

While these are all innovative concepts, there is still a wide space of opportunities that has not been explored so far. Most of the existing work was developed and approved for narrative content, sport events and game shows because these are entertaining events with a high level of community response. However, in the segments of TV magazines, documentaries and live broadcasts the challenge is to visualize and distribute more complex data sets synchronized to the live broadcast on 2^{nd} screen applications. Moreover, there has been very little focus on

[4] http://bit.ly/1J655cQ, accessed January, 2015.

[5] http://bit.ly/1i92Y2g, accessed January 2015.

[6] https://www.youtube.com/watch?v=7ftyEUIYcJ8, accessed January 2015.

representations of the content that can be adapted to the preferences and needs of different viewers, for example via personalization or location-aware features.

Another interesting aspect is the differentiation between traditional TV broadcasting and recorded content as well as online video. Taking into account other viewing habits (e.g., watching a whole season at once) the type of secondary content differs. In this case the 2^{nd} screen application could contain the content for the whole season and not only for one episode.

2.3 Visualization on Touch Screens

Visualization is as much of importance for the informed citizen as it is for expert users. For a long time, expert users have been the main target group in visualization research. Only recently, interest in broader audiences grew by activities in the areas of *visualization for the masses* [44], *casual information visualization* (e.g., ManyEyes[7], Tableau Public[8]) or *data journalism* [7]. Both, Al Gore's Nobel prize-winning campaign on global warming and Hans Rosling's work on sustainable global development[9] demonstrate how visualization can be successfully applied to educate broad audiences. However, utilization of data visualization in regular TV formats is relatively uncommon apart from classical infographics and maps in news broadcasts with relatively low information density. Moreover, the mentioned examples are incorporated for storytelling in 1^{st} screen contexts that are fully controlled by the broadcaster and keep the viewer in a passive role. During the last years, natural user interfaces (NUIs) have become increasingly relevant for visualization [34, 23]. Pike et al. [30] point out the need for better understanding and novel forms of interaction via a so-called 'science of interaction'. Along these lines, Elmqvist et al. [15] demand the concept of fluid interfaces that lets users touch and manipulate elements directly instead of interacting indirectly with interface widgets, which can be seen as extension to the concept of direct manipulation introduced with the advent of graphical user interfaces (GUIs) in desktop operating systems [19].

However, the context of mobile devices introduces a diverse set of challenges and opportunities for visualization. For example, Chittaro [10] summarized that *"[...] visualization applications developed for desktop computers do not scale well to mobile devices"*. PRISMA Mobile is an Android-based information visualization tool for tablets with treemaps, zoom (with pinch gesture), filters and details-on-demand [13]. The mobile tourism information analysis tool by Pinheiro et al. [31] is JavaME based and shows hierarchical data as treemaps, georeferenced maps and filters.

There is also some first research about using touch gestures in data visualization on mobile devices. Baur et al. [5] presented TouchWave (touchable stacked graphs, see Figure 4) with kinetic manipulations and integrated interaction without complex gestures. Drucker et al. [14] compared a non-touch-centric WIMP

[7] `http://www.ibm.com/software/analytics/manyeyes/`, accessed January, 2015.

[8] `http://public.tableausoftware.com`, accessed January, 2015.

[9] `http://www.gapminder.org/world/`, accessed January, 2015.

(window, icon, menus and pointer) interface and a touch-centric fluid interface on a tablet in a user test with 17 participants and showed that users prefer the fluid interface. Willett et al. [42] investigated user-elicited selection gestures on a non-mobile device (32" multi touch display). They found a strong preference for simple one-hand gestures, which is also relevant for implementing data visualization in 2^{nd} screen applications.

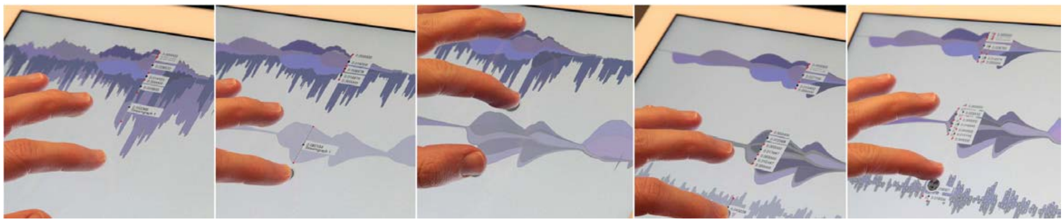


Fig. 4: TouchWave: Visualization for hierarchical stacked graphs [5].

Isenberg and Isenberg [21] published a survey on visualization on interactive surfaces. They have systematically analyzed 100 interactive systems and tools for small and big displays. The overview shows that most research work is done on multi touch tabletop devices. Smartphones are only used in 6% of the analyzed research projects although smartphones are disseminated widely. What's more, none of them is related to 2^{nd} screen applications.

2.4 Multi Screen Environments

In a 2^{nd} screen scenario, content is shared over two spatially disconnected displays. We are therefore facing a specific form of a multi display environment (MDE) with a 2^{nd} screen device and a bigger screen (e.g., TV or computer). Stemming from the different intrinsic characteristics of devices and the need to switch attention between them, MDEs pose unique challenges to the design of interactive interfaces [28].

While there has been much work in the human-computer interaction (HCI) and the computer supported cooperative work (CSCW) communities to better understand these design characteristics, there is only little work on how data visualization works in such environments. One of the few projects that intrinsically focused on visualization in MDEs is WeSpace [41] where users could bring their own laptops and share visual content on a larger screen. A similar approach was followed by Sedlmair et al. [35] when studying visual MDE applications in the automotive industry. In addition a smart view management concept for smart meeting rooms was developed by Radloff et al. [33]. This approach combines and displays views of different systems considering the dynamic user positions, the view directions as well as the semantics of views to be shown. Recently, Badam et al. [2] suggested a middleware framework for implementing visualization applications in such MDE environments.

Research in MDEs with visualization and 2^{nd} screen applications is still in its infancy. An important aspect seems to be linking and brushing across distant displays to get the users attention on the right screen for the right time.

2.5 Device Synchronization

2^{nd} screen applications require a robust device synchronization of all participating devices to manage the interactive visualization of additional content. A number of techniques have been developed so far: (1) manual sync, (2) time code sync, (3) direct link, (4) closed captions, (5) visual sync and (6) audio sync.

In (1) manual synchronization a visible (or audible) trigger is embedded into the broadcasted content and the user needs to actively push a button or select a position on a timeline to sync [4]. While this technique is easy to implement, it cannot maintain synchronization when the stream is paused.

Alternatively, (2) time codes can be used or devices can be (3) linked directly, e.g., using WIFI or a web server [27, 25]. However, these techniques require special hardware, which limits the broad applicability in different scenarios.

Another source of information for sync are (4) closed captions but a lot of content does not provide those.

(5) Visual sync triggers (QR-Codes as well as natural features) can also be used for synchronizing 2^{nd} screen devices which however is computationally expensive and heavily depends on the lighting situation. Ultimately, the audio channel provides robust features for synchronization.

Two general approaches in (6) audio-based synchronization are watermarking and fingerprinting. In watermarking, a time code for synchronization as well as data for the 2^{nd} screen (e.g. a URL) are embedded into the audio signal in a way that it cannot be perceived by humans but can be reconstructed robustly from the modified signal [22, 9]. This can be done completely at the client and the complete data is directly embedded into the primary stream. However, there are also a number of problems: the original content must be modified in advance (or at broadcasting time), licensing may prohibit watermarking for certain types of content and watermarks may become audible which is annoying for the user. In fingerprinting [39] a short audio snippet is recorded at the client, indexed, transformed into a compact signature and matched against pre-indexed content at the server with high accuracy. Fingerprinting does not change the content and is more flexible than watermarking but requires a pre- or real-time indexing of the broadcasted content. It has originally been developed for music identification [39] and has recently gained increasing attention for media synchronization in 2^{nd} screen applications due to its high precision and low latency [3, 8].

2.6 Summary

Studies confirm that smartphones and tablets are used as secondary devices whilst using the TV or computer [26, 36, 37]. However, these devices change the behavior of TV viewers. The synchronized usage opens up new possibilities (e.g., providing a 2^{nd} screen application with additional content related to the specific broadcast). Currently, there are no generalized rules for designing 2^{nd} screen applications. TV stations and producers are searching for suitable concepts, testing applications in genres like narrative content, sport events and game shows. However, other segments like TV magazines, documentaries and live broadcasts provide an excellent basis for integrating visualization of more complex data sets to substantiate the content of the broadcast.

From a technical point of view, the topic of device synchronization is well covered in research. However, concerning visualization on smaller touch screens and MDEs more research has to be carried out. Interesting aspects are for example linking and brushing across distant displays and developing visualizations for different screen sizes and operating systems. In addition, current visualization research concentrates on expert users. With the integration of interactive data visualization in 2^{nd} screen applications the target audience will become more general and diverse.

3 Technical Challenges & Opportunities

Based on the findings of the state of the art (see Section 2) we derive technical challenges and opportunities to come for data visualization in 2^{nd} screen applications.

- **Visualization for the masses:** To bring visualization to the general public, we have to think about visualizations that allow to explore the data in an intuitive and understandable way. Users should not have to learn the visualization.
- **Visualization and cross device compatibility:** In relation to the different devices which could be used for the data exploration, it will be necessary to use a framework which supports cross platform compilation. In addition, an automated (semantic) scaling for the presented visualization will be needed in relation to the big variety of device screen sizes.
- **Visualization and touch interaction:** There should be a generalized set of gestures which work for a wide range of commonly used visualization techniques. According to these techniques there is a second interesting issue depending on the different screen sizes and resolutions of the devices. For example a small device with high DPI number is more sensitive for gestures than the same device with lower DPI (e.g., lower DPI → bigger gesture; higher DPI → smaller gesture).
- **Linking and brushing across distant displays:** This aspect depends on studying and advancing the process of visual synchronization (e.g., linking and brushing) for heterogeneous displays with different types of content (broadcasted TV and interactive visualization).
- **Recording the 2^{nd} screen exploration:** In relation to the new possibilities of this upcoming technology, it will be very helpful to give the user the ability to record his/her exploration depending on the broadcast to reconstruct his/her newly gained insights. This can be very helpful for schools for example:

A teacher watches an interesting documentary and explores the data on the 2^{nd} screen device. With the record function, he/she gets the ability to show this record in the next lecture. A further idea could be a picture in picture feature for broadcast and recorded explorations.

– **Crowd sourced commenting:** Following the Polemic Tweet project[10], 2^{nd} screen applications can be used for crowd sourced commenting. By integrating such applications into a live broadcast as back channel, users can participate directly with the TV content (e.g., political discussions).

4 Conclusion

As we have shown, interactive visualization of data for and from 2^{nd} screen devices is a complex and multi-faceted endeavor that touches both, technological as well as content-related aspects. Recent technical developments allow for new perspectives on the TV of tomorrow and mobile devices such as smartphones and tablets with interactive surfaces are ubiquitous and already applied as 2^{nd} screen devices today. However, these approaches are not well integrated and are mostly constrained to pointing to supplementary information or social media platforms focusing on text, image and video material. Toward application of future visualization integrated in 2^{nd} screen applications we defined technical challenges and opportunities which are not solved yet.

Acknowledgement

This work was supported by the Austrian Science Fund (FWF) via the KAVA-Time project no. P25489 and the Austrian Ministry for Transport, Innovation and Technology (BMVIT) under the ICT of the future program via the VALiD project no. 845598.

References

1. Social TV Monitor, Halbjahresauswertung 2014. In *Social TV Monitor*. Goldenmedia Research Group, Berlin, Germany, 2014.
2. S. K. Badam, E. Fisher, and N. Elmqvist. Munin: A Peer-to-Peer Middleware for Ubiquitous Analytics and Visualization Spaces. *IEEE Transactions on Visualization and Computer Graphics*, 21(2):215–228, Feb. 2015.
3. R. Bardeli, J. Schwenninger, and D. Stein. Audio fingerprinting for media synchronisation and duplicate detection. *Proc. MediaSync*, pages 1–4, 2012.
4. S. Basapur, H. Mandalia, S. Chaysinh, Y. Lee, N. Venkitaraman, and C. Metcalf. FANFEEDS: evaluation of socially generated information feed on second screen as a TV show companion. In *Proceedings of the 10th European conference on Interactive tv and video*, EuroiTV '12, pages 87–96, New York, NY, USA, 2012. ACM.

[10] `http://polemictweet.com/index.php`, accessed September, 2015.

5. D. Baur, B. Lee, and S. Carpendale. TouchWave: kinetic multi-touch manipulation for hierarchical stacked graphs. In *Proceedings of the 2012 ACM international conference on Interactive tabletops and surfaces*, pages 255–264. ACM Press, 2012.
6. K. Busemann and F. Tippelt. Second Screen: Parallelnutzung von Fernsehen und Internet. *Media Perspektiven*, 7-8:408–416, 2014.
7. A. Cairo. *The Functional Art: An introduction to information graphics and visualization*. New Riders, Aug. 2012.
8. C. Castillo, G. De Francisci Morales, and A. Shekhawat. Online matching of web content to closed captions in IntoNow. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '13, pages 1115–1116, New York, NY, USA, 2013. ACM.
9. S. Chauhan and S. Rizvi. A survey: Digital audio watermarking techniques and applications. In *2013 4th International Conference on Computer and Communication Technology (ICCCT)*, pages 185–192, Sept. 2013.
10. L. Chittaro. Visualizing information on mobile devices. *Computer*, 39(3):40–45, Mar. 2006.
11. C. Courtois and E. D'heer. Second Screen Applications and Tablet Users: Constellation, Awareness, Experience, and Interest. In *Proceedings of the 10th European Conference on Interactive Tv and Video*, EuroiTV '12, pages 153–156, New York, NY, USA, 2012. ACM.
12. M. Dabrowski. Emerging technologies for interactive TV. pages 787–793. IEEE, Sept. 2013.
13. J. de Jesus Nascimento da Silva Junior, B. Meiguins, N. Carneiro, A. Meiguins, R. da Silva Franco, and A. Soares. PRISMA Mobile: An Information Visualization Tool for Tablets. In *2012 16th International Conference on Information Visualisation (IV)*, pages 182–187, July 2012.
14. S. M. Drucker, D. Fisher, R. Sadana, J. Herron, and m. schraefel. TouchViz: A Case Study Comparing Two Interfaces for Data Analytics on Tablets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2301–2310, New York, NY, USA, 2013. ACM.
15. N. Elmqvist, A. V. Moere, H.-C. Jetter, D. Cernea, H. Reiterer, and T. Jankun-Kelly. Fluid interaction for information visualization. *Information Visualization*, 10(4):327–340, Oct. 2011.
16. D. Geerts, R. Leenheer, D. De Grooff, J. Negenman, and S. Heijstraten. In front of and behind the second screen: viewer and producer perspectives on a companion app. In *Proceedings of the 2014 ACM international conference on Interactive experiences for TV and online video*, pages 95–102, New York, NY, 2014. ACM Press.
17. J. Hess, B. Ley, C. Ogonowski, T. Reichling, L. Wan, and V. Wulf. New technology@home: impacts on usage behavior and social structures. In *Proceedings of the 10th European conference on Interactive tv and video*, pages 185–194, New York, NY, USA, 2012. ACM Press.
18. S. Huron, R. Vuillemot, and J.-D. Fekete. Bubble-TV: Live Visual Feedback for Social TV Broadcast. In *ACM CHI 2013 Workshop : Exploring and enhancing the user experience for television*, Paris, France, Apr. 2013.
19. E. L. Hutchins, J. D. Hollan, and D. A. Norman. Direct Manipulation Interfaces. *HumanComputer Interaction*, 1(4):311–338, Dec. 1985.
20. IDC. Tablets, PCs und Smartphones - Prognostizierter Absatz bis 2018 | Statistik, 2014. Retrieved 2014-12-05, from `http://de.statista.com/statistik/daten/studie/183419/umfrage/prognose-zum-weltweiten-absatz-von-pcs-nach-kategorie/`.

21. P. Isenberg and T. Isenberg. Visualization on Interactive Surfaces: A Research Overview. *I-COM*, 12(3):10–17, Jan. 2013.

22. A. W. Jones, M. R. Reynolds, D. Bartlett, I. M. Hosking, D. G. Guy, P. J. Kelly, D. R. E. Timson, N. Vasilopolous, A. M. Hart, and R. J. Morland. System and method for shaping a data signal for embedding within an audio signal. Patent, 12 2008. US 7460991 B2.

23. D. J. Kasik. The Third Wave in Computer Graphics and Interactive Techniques. *IEEE Computer Graphics and Applications*, 31(4):89–93, July 2011.

24. M. Lochrie and P. Coulton. Sharing the viewing experience through second screens. In *Proceedings of the 10th European conference on Interactive tv and video*, pages 199–202, New York, NY, USA, 2012. ACM Press.

25. R. Martin, A. Santos, M. Shafran, H. Holtzman, and M. Montpetit. neXtream: A Multi-Device, Social Approach to Video Content Consumption. In *2010 7th IEEE Consumer Communications and Networking Conference (CCNC)*, pages 1–5, Jan. 2010.

26. Micorosft Advertising. Cross-screen engagement, 2013. Retrieved 2014-09-20, from `http://advertising.microsoft.com/international/WWDocs/User/Europe/ResearchLibrary/CaseStudy/Cross_ScreenWhitepaper.pdf`.

27. J. Murray, S. Goldenberg, K. Agarwal, T. Chakravorty, J. Cutrell, A. Doris-Down, and H. Kothandaraman. Story-map: iPad companion for long form TV narratives. In *Proceedings of the 10th European conference on Interactive tv and video*, EuroiTV '12, pages 223–226, New York, NY, USA, 2012. ACM.

28. M. A. Nacenta, S. Sallam, B. Champoux, S. Subramanian, and C. Gutwin. Perspective cursor: perspective-based interaction for multi-display environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 289–298, New York, NY, USA, 2006. ACM Press.

29. M. Obrist, P. Cesar, D. Geerts, T. Bartindale, and E. F. Churchill. Online Video and Interactive TV Experiences. *interactions*, 22(5):32–37, Aug. 2015.

30. W. A. Pike, J. Stasko, R. Chang, and T. A. O'Connell. The Science of Interaction. *Information Visualization*, 8(4):263–274, Dec. 2009.

31. S. Pinheiro, B. Meiguins, A. Meiguins, and L. Almeida. A Tourism Information Analysis Tool for Mobile Devices. In *Information Visualisation, 2008. IV '08. 12th International Conference*, pages 264–269, London, UK, July 2008. IEEE.

32. M. Proulx and S. Shepatin. *Social TV: How Marketers Can Reach and Engage Audiences by Connecting Television to the Web, Social Media, and Mobile.* John Wiley & Sons, Jan. 2012.

33. A. Radloff, M. Luboschik, and H. Schumann. Smart Views in Smart Environments. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, L. Dickmann, G. Volkmann, R. Malaka, S. Boll, A. Krger, and P. Olivier, editors, *Smart Graphics*, volume 6815, pages 1–12. Springer Berlin Heidelberg, 2011.

34. H. Reiterer. New forms of Human-Computer Interaction for Visualizing Information. In A. Kerren, C. Plaisant, and J. T. Stasko, editors, *Information Visualization*, Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2010. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.

35. M. Sedlmair, D. Baur, S. Boring, P. Isenberg, M. Jurmu, and A. Butz. Requirements for a MDE system to support collaborative in-car communication diagnostics. In *CSCW Workshop on Beyond the Laboratory: Supporting Authentic Collaboration with Multiple Displays*, 2008.

36. SevenOne Media. Der Second Screen als Verstärker, June 2013. Retrieved 2014-09-20, from `https://wirkstoff.tv/docs/default-source/second_screen_verstaerker-pdf`.
37. United Internet Media and InteractiveMedia CCP GmbH. Catch Me If You Can - Grundlagenstudie zur Multi-Screen-Nutzung, 2013. Retrieved 2013-09-30, from `http://www.multi-screen.eu/`.
38. R. von Suess, K. Blumenstein, J. Doppler, G. Kuntner, A. Schneider, and J. Brunner. Priticop 3.0 - Formatentwicklung: Wissenschaftsmagazin im Fernsehen 3.0, 2013. Retrieved 2015-09-18, from `https://www.dropbox.com/sh/2p0704lzy89l4vn/AACYpqzwSnYYEoAm2UzptYkIa?dl=0`.
39. A. Wang. An Industrial Strength Audio Search Algorithm. In *ISMIR*, pages 7–13, 2003.
40. M. Ward, G. G. Grinstein, and D. Keim. *Interactive data visualization: foundations, techniques, and applications*. A K Peters, Natick, Mass, 2010.
41. D. Wigdor, H. Jiang, C. Forlines, M. Borkin, and C. Shen. WeSpace: the design development and deployment of a walk-up and share multi-surface visual collaboration system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1237–1246, New York, NY, USA, 2009. ACM Press.
42. W. Willett, Q. Lan, and P. Isenberg. Eliciting Multi-touch Selection Gestures for Interactive Data Graphics. In *EuroVis 2014 - The Eurographics Conference on Visualization*, Swansea, Wales, UK, 2014. Eurographics.
43. V. Wrbel, S. M. Grnvogel, and P. Krebs. An analysis of new feedback methods for parallel multi-stream productions. In *Proceedings of the seventh european conference on European interactive television conference*, pages 141–144, New York, NY, USA, 2009. ACM Press.
44. R. N. Zambrano and Y. Engelhardt. Diagrams for the Masses: Raising Public Awareness From Neurath to Gapminder and Google Earth. In G. Stapleton, J. Howse, and J. Lee, editors, *Diagrammatic Representation and Inference*, number 5223 in Lecture Notes in Computer Science, pages 282–292. Springer Berlin Heidelberg, 2008.

Survey on Visualizing Dynamic, Weighted, and Directed Graphs in the Context of Data-Driven Journalism

Christina Niederer, Wolfgang Aigner, and Alexander Rind

St. Pölten University of Applied Sciences, Austria
Christina.Niederer@fhstp.ac.at
Wolfgang.Aigner@fhstp.ac.at
Alexander.Rind@fhstp.ac.at

Abstract. Data journalists have to deal with complex heterogeneous data sources such as dynamic, directed, and weighted graphs. But there is a lack of suitable visualization tools for this specific domain and data structure. The aim of this paper is to give an overview of existing publications and web projects in this area by classifying the works in a systematic characterization that adapts existing characterizations for a focus on Data-Driven Journalism (DDJ). The survey highlights a lack of work in visualizing dynamic, directed, and weighted graphs, albeit individual aspects of dynamic graphs are well explored in the graph visualization literature. The results of this survey show that Sankey diagrams and chord diagrams occur frequently in web projects. A further popular method is the animated node-link diagram. The representation of a flow (directed and weighted) is typically illustrated as lines giving the direction of the relationship and width of lines showing the weight.

Key words: dynamic graphs, data-driven journalism, network, graph visualization, quantitative flow

1 Introduction

Today we live in a world in which it is increasingly important to understand different complex phenomena to facilitate well-informed decisions. Traditionally, journalists play an important role in uncovering hidden patterns and relationships to inform or entertain readers. In addition, the amount of available data is growing and, thus, it becomes crucial for journalists to use data science in their investigative work. This trend led to the advent of *Data-Driven Journalism (DDJ)* [1]. The journalists' workflow now includes dealing with complex heterogeneous datasets. Such datasets comprise multiple variables of different data types that often stem from different sources and are sampled irregularly and independently from each other. Moreover, specialized data types need to be managed and analyzed. Often, the data structure of dynamic, weighted, and directed graphs appears such as the Austrian Media Transparency Database [2] showing the flow of money over time between governmental organizations and media companies. Because of the complex data structure and the lack of software tools especially for journalists, there is a need for analysis of existing work in dynamic, weighted, and directed graph visualization.

A *graph* can be defined as a set of objects, called *vertices* (nodes), and their relationships, called *edges* (links) [3]. In contrast to a *static* graph, a *dynamic* graph evolves over time. As von Landesberger et al. [4, p. 1721] emphasize "[t]ime-dependent changes may affect the attributes of nodes and edges, the graph structure or both". A *weighted graph* assigns a numeric attribute, called weight, to each edge. Graphs are often classified into undirected and directed [5]. In *directed graphs* the vertices of an edge are ordered. Many visualization techniques have been introduced in the field of dynamic graph visualization [4]. A number of surveys [4, 6, 7, 8] of visual techniques and also task taxonomies [9, 10, 11, 12, 13] exist in the literature. The focus of these papers lies on dynamic graphs and the categorization of visual approaches. The currently available literature lacks a survey that addresses the specific and complex data structure of dynamic, weighted, and directed graphs. The paper at hand aims to extend existing surveys by providing an overview of approaches for dynamic graphs visualizations showing quantitative flows (directed and weighted edges).

In Section 2 we discuss related work. In Section 3 we outline the systematic characterization of the dynamic, weighted, and directed graphs. Section 4 contains the description of the results, we reflect on the outcomes in Section 5, and Section 6 proposes directions for future work.

2 Related Work

Various surveys, state of the art reports, and design space papers exist to provide an overview of dynamic/temporal graph visualizations. Von Landesberger et al. [4] conducted an analysis of large graphs with the focus on the aspects of *visual representation, user interaction,* and *algorithmic analysis*. The graphs are classified according to whether they are static or dynamic (attribute change, structural change, or both) and by graph structure (tree, generic graphs, and compound graphs). In 2014 Beck et al. [6] surveyed the state of the art in dynamic graphs by classifying visualization techniques in a structured hierarchy of three layers: *animation, timeline* and *hybrid techniques*. In the same year (2014) Kerracher et al. [7] presented the work of mapping the design space of techniques for temporal graph visualization. They identified two dimensions according to which the existing visualization techniques can be classified: *graph structural encoding* and *temporal encoding*. Hadlak et al. [8] created a meta survey, which is built on existing graph visualization surveys and identifies the four common facets of *partitions, attributes, time,* and *space*.

All these state of the art reports and surveys aimed for a categorization and classification of visualization techniques in the field of dynamic graphs visualizations. To sum up, the characteristics of temporal and graph structure are considered in all papers. However, we could not identify overview literature that focuses on directed and weighted flows in dynamic graphs in particular.

Also various task taxonomies in the field of dynamic graph visualizations exist in the literature. The design space of visualization tasks by Schulz et al. [9] and the multi-level topology of abstract visualization tasks by Brehmer and Munzner [10] are general but can to some extent be applied for graph visualization. In the field of dynamic/temporal graph visualizations, the work of Lee et al. [11], Ahn et al. [12], and Kerracher et al. [13] provide more specific task taxonomies. Together, these papers provide important insight into the field of dynamic graph visualization and tasks the users perform.

The aim of this survey is to provide an update by adding more recent publications and web projects in the context of DDJ and techniques for directed and weighted graphs to the body of work presented in the existing surveys. The focus hereby lies on dynamic, weighted, and directed graphs.

3 Systematic Characterization

Our characterization of work on graph visualization consists of three groups of categories: general categories, categories relating to time (dynamic graphs), and categories relating to flows (directed and weighted edges).

General. A first categorization is done by the *application domain* (e.g., economy, science etc.) of the project or publication. Then, a categorization by *visualization technique* will give an overview of the most common representations for these graphs. In addition, we will look at the *arrangement* of nodes. For every project and publication, the status of conducting an evaluation is documented as the type of *evaluation* (qualitative, quantitative, no evaluation, or unknown).

Time. Our categories relating to time are based on existing taxonomies in the literature on dynamic graph visualization: Based on the data sources behind the visualizations we distinguish between works for either dynamic or static graphs [4]. For work supporting time, we adopt the categorization of graph structure and time component from Beck et al. [6] and categorize time interaction additionally. For *graph structure,* they distinguish between animation, timeline, and hybrid techniques on the first level of their characterization. Animation is a time-to-time mapping, this means that the different timestamps are illustrated as an animated representation. If the representation of the graph can be drawn onto a timeline in a time-to-space mapping the categorization of the visualization is a timeline. Using animation in combination with for example a static timeline Beck et al. speak about hybrid approaches [6]. Also the categorization for *time component* into either superimposed or juxtaposed are based on the survey of Beck et al. [6]. In addition, we analyze interaction techniques and interface elements used to navigate time in the category *time interaction.*

Flows. Likewise, we classify publications based on the underlying data sources showing the *characterization* of directed or undirected and weighted or unweighted graphs. We study whether direction and weights are shown on edges, nodes, or both. This survey also examines the representations of quantitative flows such as using colors or width of a line showing the direction or weight of the relationship.

Literature Search. To collect relevant publications for this report, we started to work through publications in the state of the art report by Beck et al. [6], von Landesberger et al. [4], Hadlak et al. [8]. In parallel, we used different search engines such as Google Scholar, IEEE Xplore, ACM digital library, Springer Link, and Google.

At first we defined keywords to use such as “dynamic graph visualization”, “flow visualization”, “weighted and directed graph”, “multimodal graph visualization”, and also different combinations of them. Also keywords in the area of data driven journalism “journalism” or “data driven journalism” are used. These keywords were also used to find online material and projects in this domain. The examples presented in this report are appropriate to the domain of data driven journalism with the focus on quantitative flow (directed, weighted graphs).

4 Results

Six web projects in the domains of education [14], politics [15, 16, 17], sport [18], and economy [19] were found, that show quantitative flows. 10 publications, which are relevant for this report in the domains of neuroscience [20, 21], science [22, 23, 24], ecosystem [25, 26], and social networks [27, 28] can be identified. Four of the found publications have their focus on the development of visualization techniques for no specific domain. The overview of all publications and web projects is shown in Table 1.

4.1 Visualization Techniques

Seven of the found works are classical node-link diagrams [17, 20, 21, 22, 23, 29, 30]. Besides that, hybrid representations that adapt and combine techniques are popular. For example, Google+Ripples combines node-link diagrams and treemaps [27]. Further, the node-ring representation merges node-link diagrams with the inspiration of concentric circles [31]. Etemad et al. [25] presented Eco-Spiro Vis, a visualization specifically designed for ecological networks. The representation uses the circular character of the chord diagram in combination with aspects of Spirographs to visualize directed, weighted graphs. Farragui et al. [32] introduced a visualization method for dynamic graphs inspired by the tree rings of a tree, showing the age of a tree and the amount of new growth of a tree in a year. Greilich et al. [24] published a visualization method for visualizing weighted, directed compound digraphs called TimeArcTree (Fig. 1). Based on node-link diagrams they aligned the nodes of a graph vertically for each timestamp. Two publications use a matrix-based approach for dynamic graphs. The matrix visualizations are integrated in a multiple view layout or are part of a study comparing two visualization techniques [20, 21]. Only one paper uses a

Table 1. Classification of recent publications and web projects.

	Application Domain	Visualization Techniques	Arrangement	Evaluation	Static & Dynamic	Time Component	Time Interaction	Graph Structure	Graph Characterization	Direction/ Weights on	Quantitative Flow
Woher die Daten stammen [17]	politics	node-link diagram	circular	unknown	static	–	–	–	directed	–	line
Soccer Transfer Window [18]	sport	chord diagram	circular	unknown	static	not mentioned	–	–	weighted + directed	edge	flashing balls + animation
Media Transparency Database [15]	politics	sankey diagram	–	unknown	static	–	–	–	weighted + directed	edge	width of lines
Parteispenden [16]	politics	sankey diagram	–	unknown	static	–	–	–	weighted + directed	edge	width of lines
International Trade Flow [19]	economy	chord diagram	circular	unknown	dynamic	superimposed	check box	animation	weighted	edge + node	width of lines
VisualCinnamon [14]	education	chord diagram	circular	unknown	dynamic	–	–	animation	weighted + directed	edge + node	width of lines
GraphDiaries [30]	–	node-link diagram	–	qualitative	dynamic	superimposed	thumbnails	animation	–	–	–
GraphAEL [23]	science	node-link diagram	vertical	no	dynamic	juxtaposed + superimposed	–	animation	weighted + directed	edge + node	color + size
Tree-ring Layouts [32]	ego networks	Tree-ring	circular	qualitative	dynamic	juxtaposed	–	animation	directed	–	–
TimeArcTree [24]	computer science	TimeArcTree	–	no	dynamic	juxtaposed	–	timeline	weighted + directed	edge	color
DiffAni [29]	–	node-link diagram	–	qualitative	dynamic	superimposed	selection	hybrid	–	–	–
Visual Adjacency List [33]	–	matrix-based	–	qualitative	dynamic	juxtaposed	–	timeline	weighted + directed	node	–
egoSlider [28]	ego networks	glyph-based	–	qualitative	dynamic	–	–	hybrid	undirected + weightec	edge	color, width of line
Weighted Graph Comparison [20]	neuroscience	node-link + matrix	–	qualitative	static	–	–	–	weighted	edge	color, type of line
Visualization of Energy System [26]	energy system	sankey diagram	map	qualitative	dynamic	superimposed	slider	animation	weighted	edge	width of lines
Google+Ripples [27]	social network	node-link + Treemap	–	no	dynamic	superimposed	slider	animation	directed	–	curved arrows
Node-Ring Visualization [31]	not mentioned	Node-Ring	rearrangeable	no	not mentioned	–	–	–	weighted + directed	edge + node	color, size
Maps of Random Walks [22]	science	node-link diagram	–	no	dynamic	not mentioned	–	–	weighted + directed	edge + node	width of line, color (edges) color, size (nodes)
Spirograph Inspired Visualization [25]	ecosystem	chord diagram + Spirograph	circular	no	static	–	–	–	weighted	edge + node	width of lines + new technique
Network Flow [21]	neuroscience	node-link & matrix	circular	qualitative	dynamic	superimposed	slider	animation	directed	–	invisible path + animation

The grey area at the top of the table denotes the found web projects.

matrix-based approach to visualize dynamic graphs in the form of a visual adjacency list [33]. Sankey diagrams [15, 16] and chord diagrams [14, 18, 19] showing quantitative flows occur particularly often in web projects (e.g., Fig. 2 and 3). Alemasoom et al. [26] used Sankey diagrams to generate visualization of flows and correlations in an energy system. EgoSlider by Wu et al. [28] uses a glyph-based diagram in combination with a multiple view layout, giving more insights into the data of ego networks.

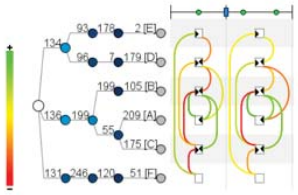


Fig. 1. TimeArcTree [24]

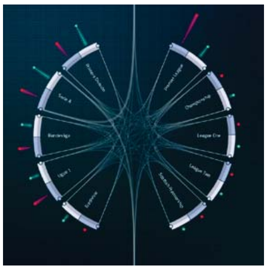


Fig. 2. Transfer Window [18]

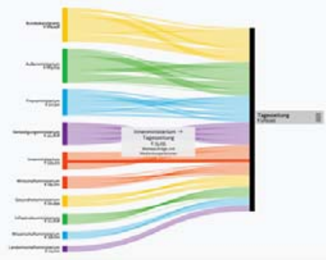


Fig. 3. Sankey diagram "Medientransparenz" [15]

4.2 Quantitative Flows

Quantitative flows are mainly depicted in the form of colors, width of lines, arrows, transparencies, animations, or flash metaphors as shown in Fig. 4. The most common representation of flows is the width of lines illustrating the weight of an edge, especially because this representation is used by visualization techniques such as Sankey diagrams and chord diagrams as well as node-link representations.

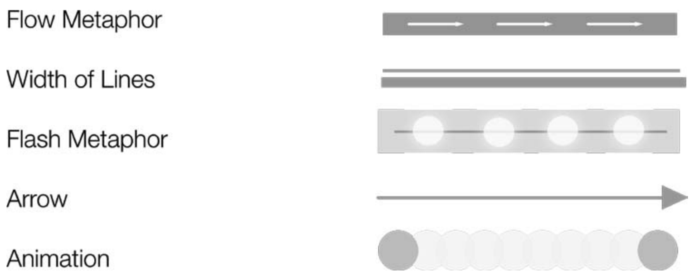


Fig. 4. Examples showing quantitative flows

4.3 Time

Two of the investigated publications use the timeline approach [24, 33] to show the time aspect. Also, two hybrid approaches can be found in the literature

[28, 29]. The most common visual representation are animations showing the changes over time [14, 19, 21, 26, 27, 30]. Erten et al. [23] and Farrugia et al. [32] use small multiples to visualize different timestamps and their changes in it. The visualization tools provide different interaction possibilities to navigate through time. The most common form are sliders [21, 26, 27]. GraphDiaries [30] integrated thumbnails giving an overview of the changes over time. Users are able to use check boxes to navigate through the visualization getting insight into changes over time [19]. To represent time, superimposition is most frequently used [19, 21, 23, 26, 27, 29, 30]. Moreover, a number of approaches use juxtaposition to give insight into the time changes [23, 24, 32, 33].

4.4 Evaluation

The most common evaluation methods are qualitative studies. Eight of 14 publications perform qualitative methods to evaluate their developed visualization technique. More than half of them do not integrate an evaluation in their research process. The evaluation status of web projects is not known.

5 Discussion

Only nine of 20 publications and web projects explored the possibilities to visualize weighted and directed graphs. The underlying data structure of three online projects are static, weighted, and directed graphs. Five scientific papers work on the problem of visualizing this special data type of dynamic, directed, and weighted graphs. Most of these developed tools use combinations of existing techniques based on node-link diagrams. Von Landesberger et al. [4] define the main challenge in using node-link representations to produce a readable layout. This includes no overlapping of nodes and less edge crossings as well as homogeneous edge lengths. It also seems to become important to find visualization methods addressing this problem and find possibilities to represent dynamic, directed and weighted graph data. The analyzed projects and publications are implemented in various domains like science, politics, social networks, sports, education, and neuroscience. Tools or visualization techniques for the domain or the special use case of DDJ are not explored in the literature. An interesting finding is that five of six online visualizations are based on Sankey [15, 16] or chord diagrams [14, 18, 19]. Only one uses a node-link diagram [17]. Often the weight of edges is shown as line width and represented together with direction by animation, which should give a visual metaphor for flow. Moreover, it became apparent that a circular arrangement of network nodes is quite common and has been found seven times. The results also show that the most common possibility for showing changes over time are animations, small multiples, and juxtaposed alignments. The interface elements to navigate over time are often sliders or checkboxes. More than half of the publications do not perform evaluations, the rest qualitative studies.

6 Conclusion & Future Work

This survey presents an overview of research literature in the area of visualization of dynamic, weighted, and directed graphs in the domain of DDJ. We updated the existing literature of dynamic graph visualization with recent publications and web projects. In addition, we focused on weighted and directed graphs due to their relevance for DDJ. The found publications are investigated along a consistent characterization that is derived from existing overview literature.

Further research should be undertaken to investigate the suitable representation of weighted and directed graphs changing over time. Alemasoom et al. [26], Chaoyu et al. [21] and Alper et al. [20] integrate existing visualization techniques as node-link or matrices in their tools. Other publications tried to combine different aspects of visualization techniques such as node-link based approaches [31] and chord diagram with Spirograph aspects [25]. For investigative exploration of graphs, new interaction possibilities for graphs should be developed. Other interaction techniques beside sliders and checkboxes are possible for interaction along the time aspect. Further interaction support for the flow aspect of graphs is needed.

Even though they were out of the scope of this survey, *multimodal graphs* are another aspect of graphs to be considered in DDJ. A multimodal graph can have different types of vertices [34] such as government organizations and media companies in the Austrian Media Transparency Database [2].

Large and complex heterogeneous datasets are the basis for most of the visualizations. The web projects often concentrate on a specific aspect, trying to communicate a certain message to the end user. The range of suitable visualization techniques is wide. So the investigation of those visualization methods to other domains and use cases could be part of further research.

The work of data journalists, which is related to the problem of working with complex data structure of dynamic, weighted and directed graphs, is not explored in depth. Most of the web projects address the end user, the consumer of online newspapers.

Acknowledgements This work was supported by the Austrian Ministry for Transport, Innovation and Technology under the initiative "ICT of the future", via the VALiD project no. 845598.

References

1. Ausserhofer, J.: "Die Methode liegt im Code": Routinen und digitale Methoden im Datenjournalismus. In Maireder, A., Ausserhofer, J., Schumann, C., Taddicken, M., eds.: Digitale Methoden in der Kommunikationswissenschaft. Digital Communication Research, Berlin (2015) 87–111
2. RTR: Bekanntgabepflichten nach dem Medienkooperations- und -förderungs-Transparenzgesetz. [Online]. Available: `https://www.rtr.at/de/m/Medientransparenz`. [Accessed: 16-09-2015].

3. Diestel, R.: Graph Theory. Springer, New York (2000)
4. von Landesberger, T., Kuijper, A., Schreck, T., Kohlhammer, J., van Wijk, J.J., Fekete, J.D., Fellner, D.W.: Visual analysis of large graphs: State-of-the-art and future research challenges. Computer Graphics Forum **30**(6) (2011) 1719–1749
5. Herman, I., Melançon, G., Marshall, M.S.: Graph visualization and navigation in information visualization: A survey. IEEE Transactions on Visualization and Computer Graphics **6**(1) (2000) 24–43
6. Beck, F., Burch, M., Diehl, S., Weiskopf, Daniel: The state of the art in visualizing dynamic graphs. In: Proceedings of the Eurographics Conference on Visualization – State of The Art Report, EuroVis STAR. (2014) 83–103
7. Kerracher, N., Kennedy, J., Chalmers, K.: The design space of temporal graph visualisation. In Elmqvist, N., Hlawitschka, M., Kennedy, J., eds.: EuroVis'14 Short Paper Proceedings, Eurographics (2014) 7–11
8. Hadlak, S., Schumann, H., Schulz, H.J.: A survey of multi-faceted graph visualization. In Borgo, R., Ganovelli, F., Viola, I., eds.: Proceedings of the Eurographics Conference on Visualization – State of The Art Report, EuroVis STAR. (2015) 1–20
9. Schulz, H.J., Nocke, T., Heitzler, M., Schumann, H.: A design space of visualization tasks. IEEE Transactions on Visualization and Computer Graphics **19**(12) (2013) 2366–2375
10. Brehmer, M., Munzner, T.: A multi-level typology of abstract visualization tasks. IEEE Transactions on Visualization and Computer Graphics **19**(12) (2013) 2376–2385
11. Lee, B., Plaisant, C., Parr, C.S., Fekete, J.D., Henry, N.: Task taxonomy for graph visualization. In: Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization, BELIV, Venice, Italy, ACM (2006) 81–85
12. Ahn, J.w., Plaisant, C., Shneiderman, B.: A task taxonomy for network evolution analysis. IEEE Transactions on Visualization and Computer Graphics **20**(3) (2014) 365–376
13. Kerracher, N., Kennedy, J., Chalmers, K.: A task taxonomy for temporal graph visualisation. IEEE Transactions on Visualization and Computer Graphics **21**(10) (2015) 1160–1172
14. Bremer, N.: Hacking a chord diagram to visualize flows. VisualCinnamon. [Online]. Available: `http://www.visualcinnamon.com/2015/08/stretched-chord.html`. [Accessed: 14-10-2015]. (2015)
15. Lang, F.: Ein Jahr Medientransparenz – die Summen sind gewaltig. [Online]. Available: `http://www.paroli-magazin.at/607/`. [Accessed: 15-09-2015]. (2013)
16. Beutelsbacher, S., Zschäpitz, H., Pauly, M., Merz, F.: Parteispenden – Wer von wem wie viel Geld bekommt. Welt Online. [Online]. Available: `http://www.welt.de/wirtschaft/article138941661/Parteispenden-Wer-von-wem-wie-viel-Geld-bekommt.html`. [Accessed: 15-09-2015]. (2015)
17. Pietsch, C., Matzat, L., Gassner, P., Wiederkehr, B., Gruhnwald, S.: Woher die Daten stammen: Visualisierte Interessenbindungen. Neue Zürcher Zeitung. [Online]. Available: `http://www.nzz.ch/schweiz/die-daten-hinter-der-visualisierung-1.18255344`. [Accessed: 15-09-2015]. (2014)
18. Signal | Noise: Transfer Window. [Online]. Available: `http://transferwindow.info`. [Accessed: 14-10-2015].

19. Hall, S.: International Trade Flows. [Online]. Available: `http://projects.delimited.io/experiments/chord-transitions/demos/trade.html`. [Accessed: 14-10-2015]. (2014)
20. Alper, B., Bach, B., Henry Riche, N., Isenberg, T., Fekete, J.D.: Weighted graph comparison techniques for brain connectivity analysis. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '13, New York, NY, USA, ACM (2013) 483–492
21. Chaoyu, Y., Handa, A., Nelson, G.L., Nied, A.C.: Network Flow: Visualizing Neurological Connectivity. Data visualization course final paper, University of Washington, Washington, DC, USA (2014)
22. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences of the United States of America **105**(4) (2008) 1118–1123
23. Erten, C., Harding, P.J., Kobourov, S.G., Wampler, K., Yee, G.: GraphAEL: Graph animations with evolving layouts. In Liotta, G., ed.: Graph Drawing. LNCS, vol. 2912. Springer, Berlin (2004) 98–110
24. Greilich, M., Burch, M., Diehl, S.: Visualizing the evolution of compound digraphs with TimeArcTrees. Computer Graphics Forum **28**(3) (2009) 975–982
25. Etemad, K., Carpendale, S., Samavati, F.: Spirograph inspired visualization of ecological networks. In: Proceedings of the Workshop on Computational Aesthetics. CAe '14, New York, NY, USA, ACM (2014) 81–91
26. Alemasoom, H., Samavati, F.F., Brosz, J., Layzell, D.: Interactive visualization of energy system. In: 2014 International Conference on Cyberworlds (CW), IEEE (2014) 229–236
27. Viégas, F., Wattenberg, M., Hebert, J., Borggaard, G., Cichowlas, A., Feinberg, J., Orwant, J., Wren, C.: Google+Ripples: a native visualization of information flow. In: Proceedings of the 22nd international conference on World Wide Web. (2013) 1389–1398
28. Wu, Y., Pitipornvivat, N., Zhao, J., Yang, S., Huang, G., Qu, H.: egoSlider: Visual analysis of egocentric network evolution. IEEE Transactions on Visualization and Computer Graphics (2015) to appear/early access. doi:10.1109/TVCG.2015.2468151
29. Rufiange, S., McGuffin, M.: DiffAni: Visualizing dynamic graphs with a hybrid of difference maps and animation. IEEE Transactions on Visualization and Computer Graphics **19**(12) (2013) 2556–2565
30. Bach, B., Pietriga, E., Fekete, J.D.: GraphDiaries: Animated transitions and temporal navigation for dynamic networks. IEEE Transactions on Visualization and Computer Graphics **20**(5) (2014) 740–754
31. Etemad, K., Carpendale, S., Samavati, F.: Node-ring graph visualization clears edge congestion. In: Proceedings of the IEEE VIS 2014 Arts Program, VISAP'14. (2014) 67–74
32. Farrugia, M., Hurley, N., Quigley, A.: Exploring temporal ego networks using small multiples and tree-ring layouts. In: Proceedings of the Fourth International Conference on Advances in Computer-Human Interactions, ACHI. (2011) 23–28
33. Hlawatsch, M., Burch, M., Weiskopf, D.: Visual adjacency lists for dynamic graphs. IEEE Transactions on Visualization and Computer Graphics **20**(11) (2014) 1590–1603
34. Hansen, D., Shneiderman, B., Smith, M.A.: Analyzing Social Media Networks with NodeXL: Insights from a Connected World. Morgan Kaufmann (2010)

Supporting Sense-Making and Insight Processes in Visual Analytics by Deriving Guidelines from Empirical Results

Johanna Haider[1], Margit Pohl[1], Chris Pallaris[2], B.L. William Wong[3]

[1] Vienna University of Technology, Austria,
{johanna.haider,margit}@igw.tuwien.ac.at,
[2] i-intelligence, Zurich, Switzerland, c.pallaris@i-intelligence.eu,
[3] Middlesex University London, UK. w.wong@mdx.ac.uk

Abstract. We need analytics systems that are optimised to human sense-making and reasoning due to the amount of data that is available for intelligence analysis. The purpose of Visual Analytics is to enable and discover insights in complex data through automatic computation and interactive visualisation. Sense-making describes the process of giving meaning to one's experience in order to understand events. To do so interactivity is the essential key to make sense out of data because the limits of human capacities are reached quickly. It is, however, not sufficiently clear how analysts derive knowledge from information systems. In this work, we investigate interaction processes with visual analytics tools to improve the understanding on how this meaning is generated. Additionally, guidelines based on empirical research and user requirements are presented.

Key words: insight model, decision making, design recommendations

1 Introduction

Digital data has the potential to inform us in many ways, but reasoning about discovered clusters, trends, and outliers requires contextualised human judgements. Yet, the main problem is that we do not know how knowledge is derived from analytical tools and, therefore, it is not clear how those systems should be designed. The R&D project VALCRI - funded by the EU under the FP7 Programme - develops a new system prototype for information exploitation by intelligence analysts working in law enforcement agencies. In this kind of analysis, an increasing amount of information is collected. With this amount of data it is sometimes difficult to make sense of the data and to derive valid inferences [1]. There are several different factors which influence this problem. One aspect is the organisation of the information and the design of the system, so that perception and reasoning processes of the analysts are appropriately supported. It is still not very well understood how users, especially how analysts, derive new knowledge from information systems. Nevertheless, there is some empirical research in human computer interaction (HCI) and cognitive psychology that can help to inform the design of intelligence analysis systems. On the basis of empirical results, guidelines and recommendations for the design of such systems need to be developed to support the work of analysts. Especially in visual analytics, a lot of research is being conducted which tries to clarify how analysts can be supported by visualisations. Empirical studies to understand and better support sense-making and insight processes have been conducted, which is why we outline the benefits of empirically derived design guidelines and work on practical sense-making guidelines for the work of analysts. The main problem is that the utilisation of empirical results is not a straightforward process, see Figure 1. Sometimes, it is not clear how basic research from cognitive psychology might be applicable because of its abstract nature. In addition, psychological research is often formulated in a way which cannot be understood easily when coming from a different domain, such as technical engineering and software development. Therefore, a process of translating this research is needed.

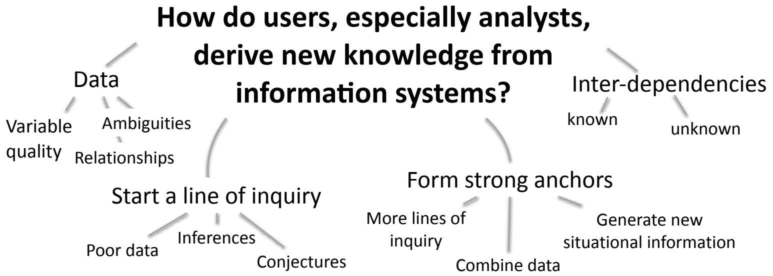


Fig. 1. Influencing factors in the retrieval of knowledge from information systems

Another challenge is to work with terms that share no common definition and, hence, are used for various tasks and processes in the literature. Sense-making has been described as "the reciprocal interaction of information seeking, meaning ascription and action" [2, p. 240], and more generally speaking by Klein as "the deliberate effort to understand events" [3, p. 114]. Zhang et al. [4] point out that sense-making is most present when we face problems where we have insufficient knowledge. For the purpose of our research Klein's definition suits our emphasis on the human as the driving force to understand data.

One of the key problems in the sense-making process is to "connect the dots" or to quickly find the fragments of relevant information from very large datasets, such as systems containing big data, and to piece them together so that it is possible to draw a sensible, reasonable, and justifiable conclusion. VALCRI focuses on Comparative Case Analysis (CCA) to identify series of linked events which can be distinguished from other events to evaluate crimes that have already happened. Much of this process is very labour intensive and inefficient. Part of the project's goals involve the development of a set of guidelines and recommendations to help the analysts in their work and to design visualisations providing a comprehensive overview of the data. In a first step a comprehensive requirements analysis was conducted to assess the users' needs. Subsequently, tentative guidelines have been developed according to the characteristics of the criminal intelligence analysis process as it is described in the literature.

In this paper, we describe this process as well as part of the resulting sense-making guidelines. Furthermore we show how they relate to the requirements analysis and to the users' needs and point out the relation to the more theoretical models of the intelligence analysis process.

2 Related Work

The classic model of sense-making for intelligence analysis by Pirolli and Card [5] organises the process of foraging and sense-making in two major loops. The foraging loop incorporates search activities whereas the sense-making loop tries to interpret the found information and make sense out of it to develop a consistent mental model. But humans do not only perceive information by user-driven processes, like searching for information, but also by visualisation-driven processes, like salience or Gestalt laws. The information processing of Pirolli and Card's model supports that, so that processing can be driven from data to theory (bottom-up) or from theory to data (top-down).

However, they point out that their model brings a cost structure with it which results in the trade-off between wide exploration and detailed exploitation of the information. Their model is not suitable to explain user interaction with information visualisations because it misses the tool and perceptual interaction that are relevant in this context. Additionally, there is no distinction between processes that are driven by the information provided and those that are driven by the user's prior knowledge. From a cognitive perspective, processes like *schematize* or *build case* should rather be classified as top-down, knowledge-driven processes than bottom-up processes [6].

According to Klein [3] we have to shift an anchor in our prior knowledge to gain new insights. He describes *insight* as an "aha" experience when all of a sudden a new understanding of events gets created. He observed five cognitive triggers that change the way we think and that lead to insights: *making connections, finding coincidences, emerging curiosities, spotting contradictions,* and *being in a state of creative desperation.*

Due to automatic computation and interactive visualisation we can work with complex data and discover new insights with visual analytics systems. Still, contextualised human judgement is necessary to reason about discovered clusters, trends, and outliers. Therefore, visualisations need to transform available data into representations that can be processed by individuals in the best possible way. The investigation of how intelligence analysts interact with visual analytics systems gives some indication of the sense-making processes users engage in while they work with the system. Attfield et al. [7], for example, argue that understanding how users work like, e.g, legal staff performing e-discovery in real life, informs us about how to support them by visual analytics tools.

3 Methodology

Within the VALCRI project we were able to get to know the processes of intelligence analysis in the environment of the intelligence analyst partners, who are the end users of the developed system. On four occasions we were introduced to the intelligence work by police analysts at their working places to get to know their day-to-day job. On this basis several bi-lateral meetings between different project partners followed, which led to a collection of more than 700 user stories describing the analysts' work including current as well as anticipated procedures within an improved system. Intelligence analysts engage in fluid and rigorous thinking and reasoning processes during an intelligence analysis task [8]. To generate further information on the cognitive processes common to analytic work, interviews with end users were conducted in addition.

Two higher goals for an improved system resulted from the requirement analysis and end user interviews. First a system that incorporates multi-dimensional data from different exploitation systems should result in efficient procedures that in the end lead to very valuable time savings but at the same time, improve the human reasoning processes and reduce the possibility for human failures, such as cognitive bias or misusing the system. The development of intelligence systems, consequently needs to consider human issues (e.g. management of task uncertainty, experience level, etc.) to achieve these high level goals. A human issues framework [9] is being developed for that reason, which covers the themes *evidential structuring and reasoning, advances in sense-making and insight, cognitive bias mitigation* and *legal, ethical and privacy aspects.*

In the next section we describe important design recommendations from the sense-making and insight point of view. Spence [10] argues for a structured approach to translate the existing information from cognitive psychology and HCI to interaction design. In the systematic form of design actions, empirical research from cognitive psychology and HCI can be made useful for information visualisation. We adopted this approach and developed a set of 17 guidelines, of which we will elaborate on 7 in the following.

4 Sense-Making Guidelines

The following guidelines consider general sense-making processes that are relevant as they relate to the requirements analysis and to the users' needs. By way of example we chose the best documented guidelines about areas, where a lot of research exists, which can be reformulated in an applicable, generic way to inform and in further consequence improve the design of visual analytics tools.

4.1 Provide Different Perspectives on the Data

Data from the end user interviews show that analysts are actively looking for contradictions in a hypothesis. They challenge their hypotheses against contradicting data and they go after each piece of evidence to form a coherent story.

To support this a system should provide different perspectives on the data and let the user work with multiple visualisations.

A design that supports relationship mapping in the data and form a coherent mental model can be achieved by multiple coordinated views. Multiple coordinated views combine different views on the data in one visual representation like, e.g., Microsoft's Windows Explorer combines an outliner view of the folders, a tabular view of the files in the selected folder and a quick details view of the selected file. Empirical data shows that the ability to see data in multiple linked views significantly speeds up easy as well as difficult tasks in comparison to a single scrolling view [11]. The requirement analysis further shows that analysts want to work with different visualisations that support cross filtering. They want to have a visual representation of spatial events on maps and additional information of the events in the investigated area to get ideas in which direction an inquiry can go. A multiple view approach supports that because unexpected or less expected ideas can be shown as well.

4.2 Provide an Open Exploration and Allow the Redefinition of the Goal and the Methods

Research in everyday reasoning indicates that so-called wicked problems are solved differently than clear-cut problems. Wicked problems have no clear method or path to the solution and it is sometimes not clear how the solution might look like. It is typical for such problems that users explore the problem space and often redefine the goal of analysis while working on the solution [3,12]. A trade-off decision needs to be done so that exploration gets supported in a way which does not lead users astray but helps to brainstorm on the one hand and to focus at the end of the exploration process on the other hand. The higher objectives for a better system are saving time and reducing uncertainties and errors. A more concrete goal would be to identify relations between crimes. A general hypothesis is that a crime is seldom a single event.

Search is especially in the context of CCA a particularly interesting topic due to its breadth in functionality. One might be interested to find similarities and look for patterns or find oddities that stand out from the others. Control over threshold parameters should be given to the user as one can more openly explore the data and is able to bring in one's experience in that way. Analysts have different privileges at different systems and sometimes have to ask for access to different databases which results in delays of different magnitudes. Analysts feel the barrier that this procedure brings to the investigation process. Though restricted, a vast amount of data is available and uncertainties as well as gaps (missing data) need to be represented in equal terms. This needs to be addressed in distinct guidelines from other perspectives but also is considered in these sense-making guidelines.

4.3 Support Holistic Sense-Making Processes by Structuring Information in a Coherent Manner

Gestalt psychology indicates that sense-making is a holistic process where structure plays an important role and empirical evidence shows that everyone develops his or her own structure [13]. Hence, users should be able to generate their own structures of knowledge where this is possible, e.g., be able to structure network representations (e.g. social networks) themselves if it makes sense in the context of their work. However, anything too complicated and time-consuming is unlikely to be used in an environment characterised by significant time-pressure. The requirement analysis supported this because the analysts prefer to start from scratch rather than remodel an auto-generated representation. In addition to this, program response times are important for sense-making and reasoning. Analysts stated that slow programs are very frustrating. Responses, like e.g. search results, should ideally be available to the user at the speed of thought. Otherwise the slower response might influence reasoning, as one might have no time to follow an idea or forgets a sparking idea in the meantime [14].

4.4 Support Reuse and Provide Graphical Histories

An investigation can last over days and weeks and an analyst might end up looking for the same thing over and over again. Therefore, users should be able to reuse done work and get a fluid transition between different visualisations to save time. A keyword search on text documents should be transferable onto other data so that results can be texts but for example also spatial data on a map.

To provide a graphical history on the one hand improves a quick reuse of a previous search but also provides an overview of the investigation work. Additionally to facilitating analysis and communication, graphical histories may help in teaching analysis by example [15]. The problem with visual histories is that they do not scale up. It might be faster to do a new search than to look up an entry in a long search history. For example, the history in a web browser is a simple list that lists every site you visited that gets very long in a short time which makes it hard to search for a specific site. An entry in the history must be identifiable and for reusing a previous search it must be visible what the result set means. To provide a graphical history is a challenging task but users appreciate the feature if they can reuse done work.

4.5 Provide Interaction Possibilities

Providing interaction techniques supports sense-making. A single image can only answer a small set of question. Visual analysts gain insight by repeatedly looking at the data, exploring, refining the views iteratively and hence developing insights [14]. The theory of Distributed Cognition assumes that interaction with cognitive tools (e.g. the VALCRI system) plays an important role for sense-making processes [16].

Semantic zooming is one interaction technique that should be provided for temporal and spatial data to enable the exploration of more details as a chart is enlarged in the area where the zoom happens. Granularity in time can vary from years to weeks to days, time of day, etc., the granularity of the spatial data occurs in different levels as well, for example zip code, grid, and address, whereas address is the most common practise in crime analysis. Geocoding is a critical issue because the visualisation reveals information that is not obvious by looking at a street name or address. However, using coordinates from GPS data is the most accurate and reliable way to locate incident data, especially when they happen in parks or on vacant land [17].

4.6 Enable the Users to Cluster Similar Cases

Another interaction technique is to offer an easy clustering and hot spots generation possibility to show similarities in the data. There are various mathematical methods to cluster data that can support analysts in finding similarities, e.g., in MOs (modus operandi) of criminal gangs. The results of the mathematical methods can be visualised on the screen so that analysts can see at a glance which data is similar. It should be possible for the users to influence the clustering process, for example to choose how many clusters should be presented. As a result we need to distinguish between machine and human created clusters.

Cluster visualisation can be used for rapid identification of relevant documents. Allan et al. [18] support the cluster hypothesis and show how feedback techniques enhance the effect to help users separate relevant from non-relevant documents. Data can be clustered in different ways, depending on the criteria for the clustering procedure. It is sometimes difficult to identify a clustering procedure that helps the user to gain insights. For example, in node-link diagrams (networks) the users could cluster persons with similar attributes or the users could cluster crimes with a similar MO. It might sometimes be difficult to identify the criteria for what constitutes a similar MO which indicates that the same group of persons is responsible for a series of crimes. In alignment with the guideline to support reuse users ideally should be able to save a cluster selection for future analysis.

4.7 Do Not Overwhelm the User with Too Many Connections

Displays that visualise too many individual data points become cluttered which affects analysis, because trends and patterns are harder to spot. There are different techniques to measure and reduce clutter in a visualisation [19]. Crime maps need to convey a lot of attributes in different granularities, like the incident numbers per year (as well as per hour a day, etc.), the crime types, boroughs or boundaries. It is sometimes difficult to identify what constitutes clutter. How much information is needed might depend on the data or the task. Possible solutions for this problem are to enable filtering or highlighting interesting data.

The representation of large datasets can be simplified with the help of aggregates, which represent a group of data points so that fewer markers are needed.

A challenge in the construction of aggregates is to choose the right granularity. This is an important decision for the effective visualisation and depends on the application domain and on the task at hand [20, 21]. The problem with different aggregation levels is that anomalies are possibly hidden [22]. On the other hand effective maps have to be highly generalised so that important trends and features emerge [23]. To reduce this threat the system needs to give control to the user and raise curiosity to explore different levels. To address the problem of the unit of analysis exploratory tools should allow the selection and combination of aggregates and give several perspectives on the data.

Further, in the context of map visualisations spatial exaggeration is a relevant factor. The data of interest needs to be exaggerated in order to make connections and contradictions visible. An example that came up during the requirement analysis is that telephone records and the routes of buses might be connected and an exaggeration of antenna regions may reveal that the mobile phone data is following a certain bus route.

5 The Intelligence Analysis Process - Identification of Meaningful Patterns

An important goal in the VALCRI project is to support analysts in CCA. In this context, the identification of meaningful patterns plays a significant role. We want to give a brief outline of how this process of pattern identification is described in the literature. Crimes often follow distinct geographical patterns that can be uncovered using maps because the spatial and temporal dimensions of criminal activity can be clarified easily through such visualisations [17]. In the analysis of patterns analysts have to recognise differences and similarities in the cases of interest. This is done in an iterative process of manipulating, searching and sorting the characteristics of crimes. The United Nations Report on Drugs and Crime 2011 [24] distinguishes between four different analysis techniques: *link analysis, event charting, flow analysis*, and *telephone analysis*. In our context, link analysis is especially interesting. The authors point out that links between entities (persons, locations, ...) should be represented graphically to clarify their relationships. The Major Incident Manual 2006 [25] describes Crime Pattern Analysis as a method to identify patterns and trends in data on crime and incidents. Among the methods that this report suggests for this analysis are also visualisations (bar charts, mapping).

Several of the guidelines on sense-making developed for VALCRI address issues raised in the literature on the intelligence analysis process. The guidelines in general emphasise the importance of the identification of patterns. We describe how analysts can be induced to perceive connections as well as contradictions. Other recommendations indicate that it is important to support interaction. Interaction helps analysts in the manipulation, sorting and searching of data [17]. We also suggest assisting the analysts in processes of open exploration, although care should be taken to help the analyst to reach a conclusion. The guidelines also mention the use of graphical representations (e.g., maps, node-link diagrams,. . .) to support this process because some patterns can be identified more easily through visualisations. In this context, multiple views showing data in different forms of graphical representation play an important role. Multiple views also help analysts to identify unexpected connections between the data, but also contradictions.

6 Conclusion

In this paper we give an overview of our work in progress of design guidelines for visual analytics systems in the field of criminal intelligence analysis. We discuss part of the guidelines developed in the course of the VALCRI project. Their applicability results from a detailed requirement analysis as well as their relations to the theoretical models of the intelligence analysis processes described in the literature. In this way, we want to outline the practical relevance of the guidelines. In future work we will refine the guidelines on the basis of the developer feedback and adapt them further to the users' requirements and add additional guidelines as we get to know more insights into their working processes.

Acknowledgement The research leading to these results reported here has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) to Project VALCRI under the EC Grant Agreement N° FP7-IP-608142 awarded to Middlesex University and partners.

References

1. Thomas, J.J., Cook, K.A., eds.: Iluminating the Path: The Research and Development Agenda for Visual Analytics. IEEE CS Press (2005)
2. Thomas, J.B., Clark, S.M., Gioia, D.A.: Strategic sensemaking and organizational performance: Linkages among scanning, interpretation, action, and outcomes. Academy of Management journal **36**(2) (1993) 239–270
3. Klein, G.: Seeing What Others Don't: The Remarkable Ways We Gain Insights. PublicAffairs, a Member of the Perseus Book Group, New York, USA (2013)
4. Zhang, X., Qu, Y., Giles, C.L., Song, P.: Citesense: Supporting sensemaking of research literature. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '08, New York, NY, USA, ACM (2008) 677–680
5. Pirolli, P., Card, S.: The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. Volume 5. (2005) 2–4
6. Pohl, M., Smuc, M., Mayr, E.: The user puzzle - explaining the interaction with visual analytics systems. IEEE Transactions on Visualization and Computer Graphics **18**(12) (2012) 2908–2916
7. Attfield, S., Blandford, A.: Discovery-led refinement in e-discovery investigations: sensemaking, cognitive ergonomics and system design. Artificial Intelligence and Law (2010) 1–23
8. Wong, B.L.: How analysts think (?): Early observations. In: IEEE Joint Intelligence and Security Informatics Conf. (JISIC). (2014) 296–299

9. Haider, J., Pohl, M., Hillemann, E.C., Nussbaumer, A., Attfield, S., Passmore, P., Wong, B.L.W.: Exploring the challenges of implementing guidelines for the design of visual analytics systems. In: Proceedings of the Annual Meeting of Human Factors and Ergonomics Society, Los Angeles (2015, in print)
10. Spence, B.: The broker. In Ebert, A., Dix, A., Gershon, N., Pohl, M., eds.: Human Aspects of Visualization, Springer (2011) 10–22
11. North, C., Shneiderman, B.: Snap-together visualization: A user interface for coordinating visualizations via relational schemata. (2000) 128–135
12. Davidson, J., Sternberg, R., eds.: The Psychology of Problem Solving. Cambridge University Press, Cambridge (2003)
13. Pretz, J.E., Naples, A.J., Sternberg, R.J.: Recognizing, Defining, and Representing Problems. In: The Psychology of Problem Solving. Cambridge University Press (2003) 3–30
14. Heer, J., Shneiderman, B.: Interactive dynamics for visual analysis. Commun. ACM **55**(4) (April 2012) 45–54
15. Heer, J., Mackinlay, J.D., Stolte, C., Agrawala, M.: Graphical histories for visualization: Supporting analysis, communication, and evaluation. IEEE Transactions on Visualization and Computer Graphics **14**(6) (2008) 1189–1196
16. Hollan, J., Hutchins, E., Kirsh, D.: Distributed cognition: Toward a new foundation for human-computer interaction research. ACM Transactions on Computer-Human Interaction **7**(2) (June 2000) 174–196
17. Boba Santos, R.: Crime Analysis with Crime Mapping. Sage, Los Angeles, London, New Delhi (2013)
18. Allan, J., Leouski, A., Swan, R.: Interactive cluster visualization for information retrieval. In Proceedings of ECDL'98 (1997)
19. Peng, W., Ward, M.O., Rundensteiner, E.a.: Clutter reduction in multi-dimensional data visualization using dimension reordering. IEEE Symposium on Information Visualization (2004) 89 – 96
20. Fredrikson, A., North, C., Plaisant, C., Shneiderman, B.: Temporal, geographical and categorical aggregations viewed through coordinated displays: A case study with highway incident data. Workshop on New Paradigms in information Visualization and Manipulation (1999) 26–34
21. Weisburd, D., Bernasco, W., Bruinsma, G., eds.: Putting Crime in its Place. Springer New York (2003)
22. Robertson, G., Fernandez, R., Fisher, D., Lee, B., Stasko, J.: Effectiveness of animation in trend visualization. IEEE Transactions on Visualization and Computer Graphics **14**(6) (2008) 1325–1332
23. Harrower, M.: Tips for designing effective animated maps. Cartographic Perspectives **0**(44) (2003) 63–65
24. United Nations New York: United Nations Office on Drugs and Crime. Criminal Intelligence. (2011)
25. Association of Chief Police Officers in Scotland (ACPOS), Centrex: Major Incident Manual. (Revised Edition, 2006)

Part II

Image Analysis: Reconstruction, Segmentation, Restoration, and Recognition

Concepts for Underwater Stereo Calibration, Stereo 3D-Reconstruction and Evaluation

Tim Dolereit[1][2]

[1] Fraunhofer Institute for Computer Graphics Research IGD, Joachim-Jungius-Str.
11, 18059 Rostock, Germany
[2] University of Rostock, Institute for Computer Science, Albert-Einstein-Str. 22,
18059 Rostock, Germany
, tim.dolereit@igd-r.fraunhofer.de

Abstract. Handling refractive effects in computer vision disciplines like underwater stereo camera system calibration or 3D-reconstruction is a major challenge. Refraction occurs at the borders between different media on the way of the light and introduces non-linear distortions, that are dependent on the imaged scene. In this paper, concepts will be proposed for the calibration of a stereo camera system including a set of additional refractive parameters, for underwater stereo 3D-reconstruction and for evaluation of the computations.

Key words: Underwater Imaging, Underwater Camera Calibration, Underwater 3D-Reconstruction, Stereo Camera Systems

1 Introduction

The application of imaging devices in underwater environments has become a common practice. They can be installed on autonomous underwater vehicles (AUV), remotely operated vehicles (ROV) and also divers can be equipped with them. The non-destructive behavior toward marine life and its repeatable application makes underwater imaging an efficient sampling tool. Underwater imaging is confronted with quiet different constraints and challenges than imaging in air. The camera's constituent electric parts have to be protected against water. This leads to setups where cameras are looking through a viewing window like an aquarium or to cameras being placed inside a special waterproof housing. All of these setups are subject to refraction of light passing bounding, transparent interfaces between media with differing refractive indices (water-glass-air transition). Refractive effects lead to objects seeming to be closer to the observer and hence bigger than they actually are. The effects are non-linear distortions, that depend on the incidence angle of light rays onto the refractive interface. These non-linear magnifications are a problem for gaining metric information from images like 3D-reconstruction with conventional in air approaches.

For gaining metric 3D-reconstructions, using a stereo-camera-system is a common practice. The cameras' intrinsic and relative extrinsic parameters have to be calibrated. Since the imaging behavior of a camera in air can be well approximated using the linear pinhole camera model of perspective projection, it forms the foundation of most calibration algorithms. Additionally, refractive effects have to be handled in underwater environments. Underwater images have multiple viewpoints [15], hence, calibration of cameras in underwater usage is theoretically not possible with the pinhole camera model. Due to the fact that the imaging model does not match the imaging conditions, it is an acknowledged approach to account for refractive effects by modeling them explicitly. Therefore, the pose of the refractive interface towards the cameras has to be calibrated as well. Afterward, a physically correct tracing of light rays can be utilized for 3D-reconstruction.

The parameters representing the pose of the refractive interface will be referred to as refractive parameters. These parameters comprise the orientation between a camera's optical axis and a refractive interface's normal (retrieval will be referred to as *axis determination*), as well as the distance of the camera's center of projection along this normal (retrieval will be referred to as *distance determination*).

In the following, some concepts will be presented on how to perform stereo 3D-reconstruction underwater, ranging from system calibration to evaluation of gained results. Most of the concepts are based on earlier works of the author on virtual object points - proposed to be actually seen by the cameras - in underwater imaging [2]. A model can be utilized to relate the location of these virtual object points non-ambiguously to the real object points. The main contributions of this work are to show

- **that axis determination in system calibration can be performed independently of knowing refractive indices of the participating media as well as interface thickness.**
- **that 3D-reconstruction can be done by utilizing virtual object points and can be simultaneously used as a constraint for system calibration.**
- **how evaluation concepts like generation of ground truth data, refractive reprojection error or computation of correspondence curves can be realized.**

2 Related Work

In this section a brief overview on handling refraction in relation to camera calibration is given. A comprehensive overview on camera models in underwater imaging can be found in [12].

Many works are founded on the pinhole camera model alone. This means, refractive effects are either completely ignored [5, 8, 14] or expected to be absorbed by the non-linear distortion terms [13, 10]. Further similar approaches using in-situ calibration strategies are mentioned in [6, 11].

A second way to handle refraction is by approximation. *Ferreira et al.* [4] assume only low incidence angles of light rays on the refractive surface. *Lavest* *et al.* [9] try to infer the underwater calibration from the in air calibration in form of an approximation of a single focal length and radial distortion. It is also based on the pinhole camera model.

The applicability of the pinhole model in imaging through refractive media is said to be invalid by many authors [15, 1, 16, 7]. Since the pinhole camera model cannot handle refractive effects, approaches were developed handling these explicitly. Hence, refractive effects are modeled physically correct and are incorporated into the camera model and calibration process. The camera model is extended by refractive parameters.

Since this is the only way to handle refractive effects physically correct, the proposed concepts also aim at a solution to calibrate additional refractive parameters.

3 System Design and Restrictions

The concepts to be presented are for now restricted to stereo cameras in a single underwater housing with a flat interface. This design leads to some useful simplifications, which will be explained in section 4. The cameras can be arbitrarily oriented towards the refractive interface and each other. The stereo camera system has to be calibrated for intrinsic and relative extrinsic camera parameters in air and both cameras are supposed to have a constant focal length. Both cameras' non-linear distortion terms in air are supposed to be calibrated as well. It is expected that lens distortion is not inuenced by refraction and hence can be eliminated by standard in air distortion correction algorithms in advance. The way of the light is characterized by a water-glass-air transition. The indices of refraction of the involved media are expected to stay constant. The refractive index of water is supposed to be equal to 1.33 and of air equal to 1. Furthermore, the refractive interface's thickness and its refractive index are known as well, since they can be usually determined manually.

4 Calibration and 3D-Reconstruction

The calibration of the underwater stereo camera system is performed in two phases. The first phase is the determination of all the described parameters in the previous section 3 in a pre-process. This is done in air. The second phase comprises the determination of the refractive parameters. These parameters are illustrated in Fig. 1. During the so called *axis determination* the orientation between the left camera's optical axis and the interface's normal is computed. This can be parametrized in 3D space by spherical coordinates. Afterward, the last refractive parameter is computed during the so called *distance determination*. It is the distance between the center of projection of the left camera and the water-sided interface border along the determined axis.

In the following, concepts for this refractive calibration for the earlier specified setup will be presented.

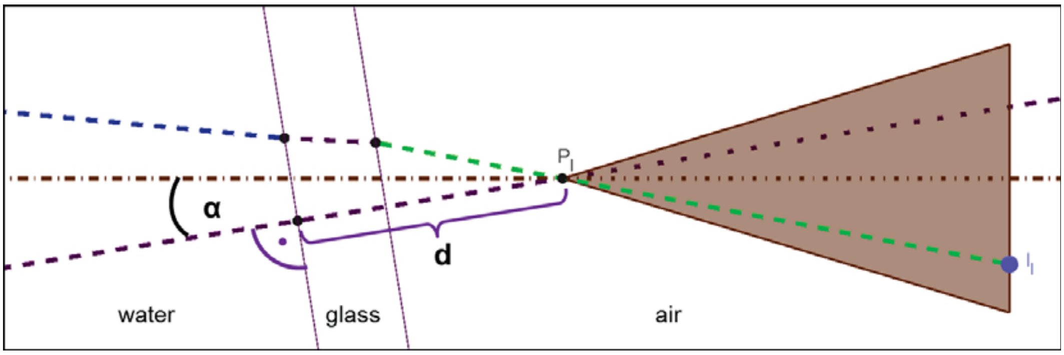


Fig. 1. Simplified illustration of the left camera with a center of projection P_l. A ray arriving in pixel I_l is refracted twice at the interface between water and air. The interface is parametrized by the angle α between the camera's optical axis and the interface's normal and the distance d along this normal.

4.1 Independence of Axis Determination

The specified setup of a stereo camera and a single flat refractive interface leads to some useful simplifications. As is known from physics, refraction always happens in a plane - the so called plane of refraction. A single plane of refraction is spanned by two vectors. The first is the image ray from a pixel through the center of projection and the second is the refractive interface's normal.

Hence, we get two planes of refraction for a corresponding pixel pair in a stereo camera system. Since we have a single refractive interface, both, the left plane of refraction as well as the right plane of refraction are spanned by the same interface normal and the corresponding image ray (see Fig. 2). This leads to the fact, that both planes of refraction have to intersect in a line with the same direction as the interface's normal, as long as both planes are not parallel. Since the planes of refraction can be determined without knowing the entire way of the light through all participating media, the *axis determination* is independent of the values of the refractive indices, the interface's thickness and the distance to the interface.

Axis determination can be performed, for example, by deriving some constraints with the aid of known calibration targets like in [3]. An error metric is minimized followed by a second minimization for *distance determination* based on the resulting axis. This means, if the resulting axis is erroneous, the resulting distance will incorporate this error. Another way is to do a two-dimensional search over the possible spherical coordinate space for the axis. Hence, one minimization process can be replaced due to the independence of *axis determination*. In the following, a concept will be proposed on how to perform *axis-* and *distance determination* simultaneously by connecting such a search and 3D-reconstruction.

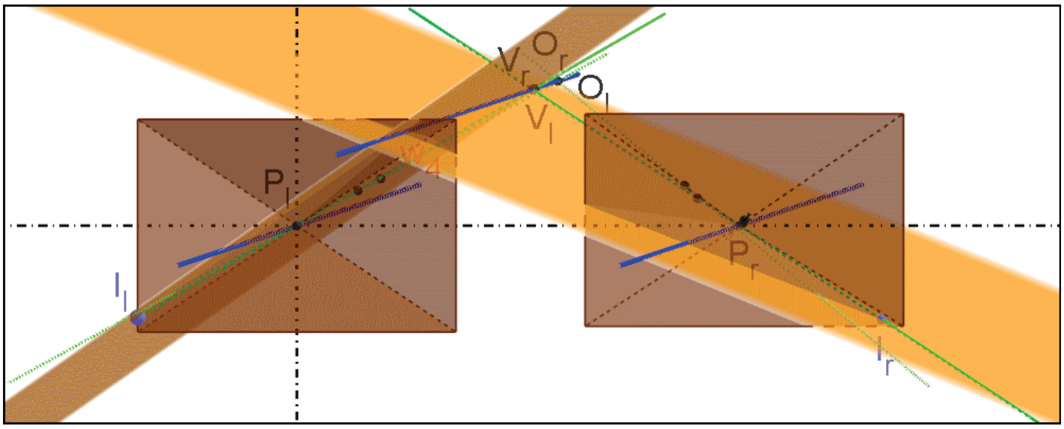


Fig. 2. Rear view of the stereo camera system in 3D. The image rays (green lines) of a corresponding pixel pair I_l and I_r together with the refractive interface's normal (blue line) span the two planes of refraction (orange planes). Both planes intersect in a line which has the same direction as the interface's normal.

4.2 3D-Reconstruction as a Constraint

Suppose, we already know the true axis. This enables us to compute the planes of refraction for every corresponding pixel pair, as well as the corresponding line of intersection of both these planes. The real object point has to lie on this line. The non-refracted left and right image ray (extended into water - see solid lines in Fig. 3 right) also have to intersect this line. These intersection points are called virtual object points in the following, since they are proposed to be seen virtually by the cameras. The left and right virtual object points only coincide if the incidence angles of the non-refracted rays are equal. On the contrary, the left and right refracted rays always coincide in the real object point (see dashed lines in Fig. 3 right). Hence, the resulting left and right virtual point as well as the real object point lie on the line of intersection (see Fig. 2 & Fig. 3 right). The overall closest virtual point gives us a maximal distance at where the refractive interface can be placed. If the interface is placed at the true distance d, tracing a pair of corresponding rays by explicitly considering refraction until both of them meet the line of intersection should result in a single real point (see Fig. 3 right). The final placement of the interface is done by minimizing the difference between the left and right real point for all corresponding rays (compare [3]).

In this way, the distance to the interface can be determined simultaneously with 3D-reconstruction of the corresponding pixel pairs. There is no need for a known calibration target. The 3D-reconstruction forms the constraint for the error metric. In combination with the above proposed two-dimensional search over the possible spherical coordinate space for the axis and by finding the overall minimal error value, the refractive calibration can be fulfilled.

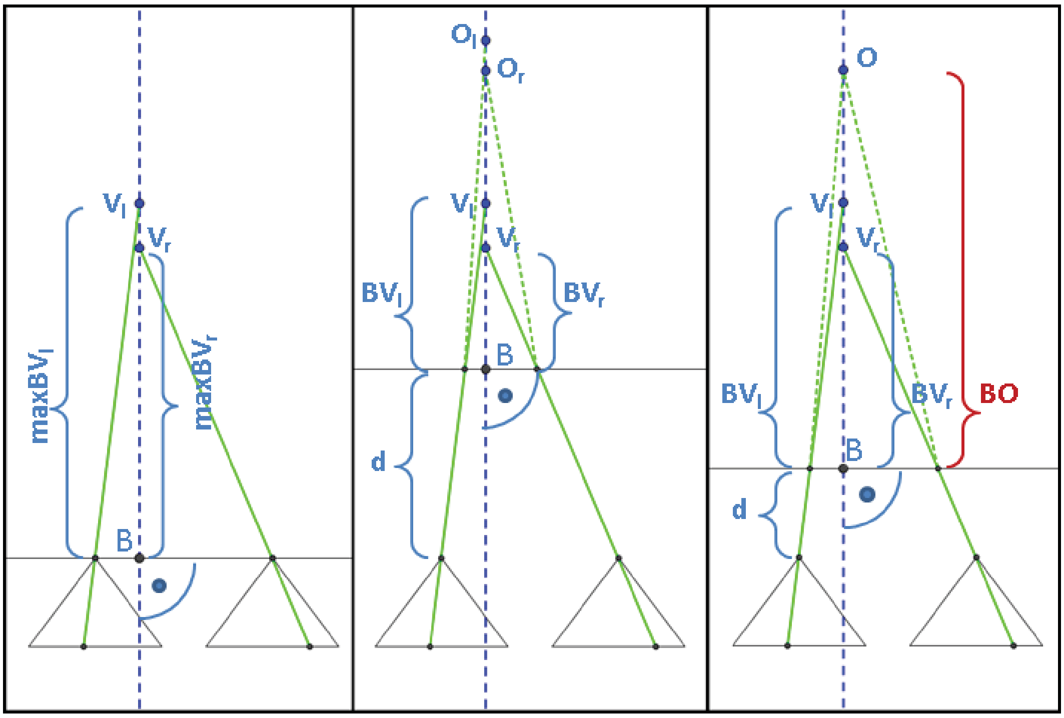


Fig. 3. 3D-reconstruction as a constraint for refractive calibration. Left: Initial conditions for a given axis (dashed blue line). Middle: Wrong placement of the interface at distance d. Right: Correct placement of the refractive interface results in a single real point.

5 Evaluation Concepts

After the calibration of the refractive parameters, a physically correct tracing of the rays can be done resulting in a 3D-reconstruction with explicit consideration of refraction. The calibration as well as the 3D-reconstruction should be evaluated. In-air algorithms use the reprojection error for this purpose. The reconstructed 3D points are projected perspectively onto the image and are compared to the corresponding pixels. The distance between projected and detected pixels forms the error metric. Since perspective projection is invalid underwater due to refraction, the reprojection error has to be modified.

5.1 Refractive Reprojection Error

The projection of underwater 3D points onto the image requires solving a polynomial of 4th degree for a single refraction and of 12th degree for two refractions [1] (water-glass-air). The proposed concept is to determine a virtual object point V - non-ambiguously related to the 3D point O - that can be projected perspectively (see Fig. 4). The relation between the location of the virtual and the real object point was described in previous works [3]. Utilizing this relation, one can determine V by simple bisection. The following perspective projection is straightforward. This refractive reprojection error is an efficient means for evaluation of reconstructed 3D points.

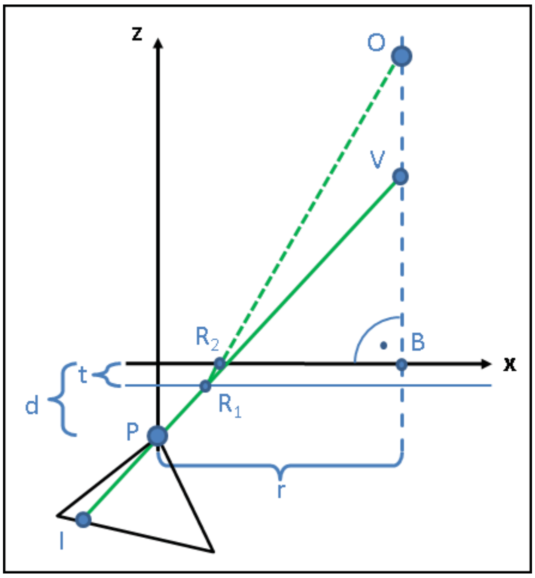


Fig. 4. Refractive reprojection for a calibrated system. 3D point O is related to a virtual object point V that can be projected perspectively onto the image.

5.2 Computation of Correspondence Lines

Another means for evaluation of reconstructed 3D points are correspondence lines. The computation of correspondence lines in underwater computer vision is supposed to correspond to the computation of epipolar lines with the aid of epipolar geometry in air. Since perspective projection is invalid underwater, epipolar geometry is as well. Epipolar geometry is used to reduce the search space for a corresponding pixel in the second view to a single straight line. These lines are curves underwater due to refractive effects. The correspondence lines can be computed in a similar way as the refractive reconstruction error in the last section. Therefore, a ray in water that belongs to the pixel for which the correspondence is searched for, is sampled into a specific number of 3D points. The points start at the water-sided interface border and end at a user-defined distance.

Besides the application for reduction of the search space for correspondences, the so computed correspondence lines can be used as a visual cue for evaluation of the refractive calibration. An example can be seen in Fig. 5. If the calibration is correct, the correspondence line for a chosen pixel should hit the same feature point in the second view.

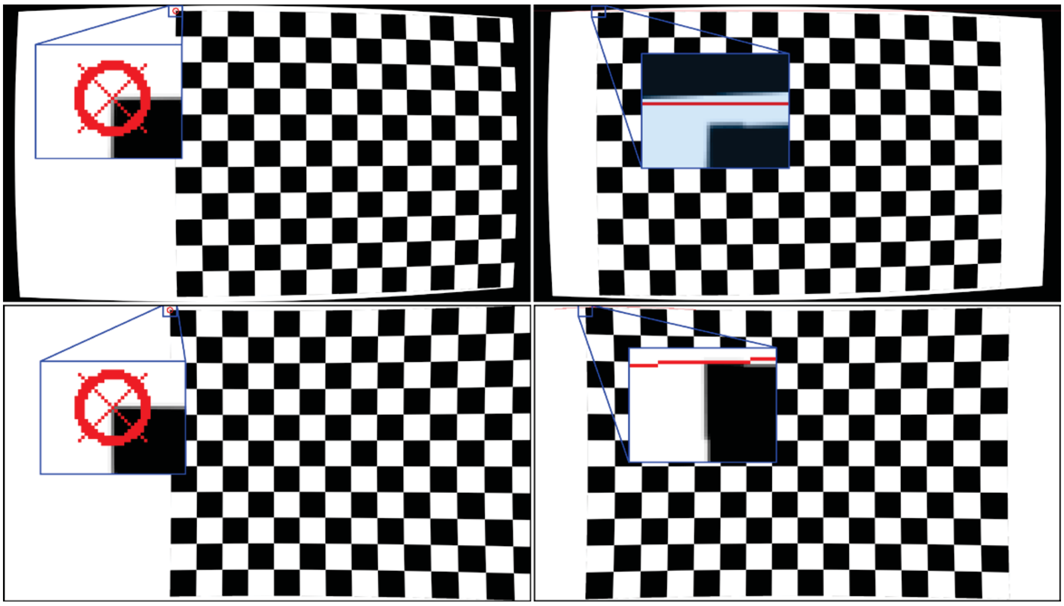


Fig. 5. Comparison of epipolar line computation after image rectification (top row) and computation of a correspondence line (bottom row) for a selected pixel for simulated image data. As can be seen, the epipolar line wold clearly miss by several pixels.

5.3 Generation of Ground Truth Data

The last proposal for evaluation is simply the generation of ground truth data. As can be seen in Fig. 6, a solid frame was built around a fish tank. Profile rails were used to fix a checker pattern target and a stereo camera system rigidly. Two GoPro Hero 3 Black cameras were used. The intrinsic and relative extrinsic parameters were calibrated in air. The whole frame can be lowered into the water. Hence, a 3D point cloud can be computed in air with conventional reconstruction algorithms. This point cloud in the stereo camera's coordinate system serves as ground truth data. It can be directly compared with the reconstructed 3D points underwater.

6 Conclusion and Future Work

The proposed concepts a preliminary works that are currently tested and improved. They mostly build upon basic findings from previous works of the author [3]. Handling refractive effects correctly in underwater computer vision tasks like system calibration and 3D-reconstruction is a major challenge. The concepts are supposed to lead to a solution for calibration of a stereo camera system with a flat refractive interface in underwater usage. The refractive calibration can be done without the need of a known calibration target. Simultaneously, the capability of 3D-reconstruction was presented. Combining the independence of *axis determination* of refractive calibration with the proposed 3D-reconstruction constraint seems to be a promising concept for calibration.

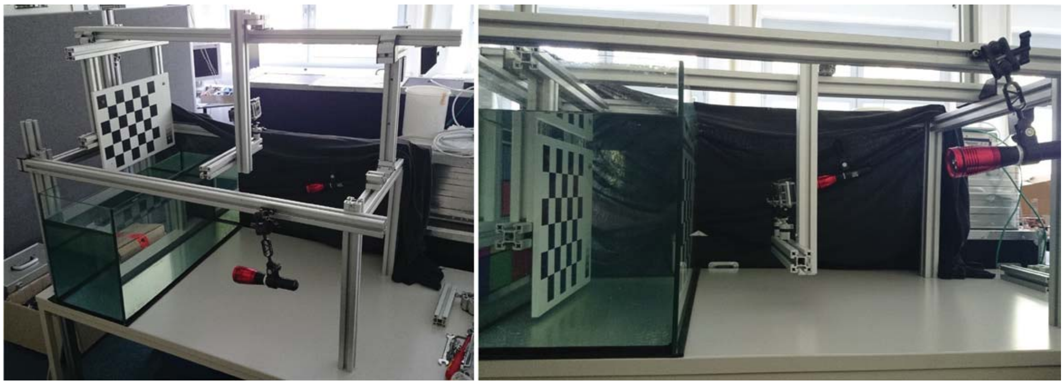


Fig. 6. Generation of ground truth data. A checker target that is fixed to the stereo camera system can be lowered into water.

Evaluation of the refractive calibration is naturally a difficult task, since most of the times it can be measured physically. The presented evaluation concepts like generation of ground truth data, refractive reprojection error or computation of correspondence lines are means to check the quality of the computations, both visually and computationally.

Acknowledgment

This research has been supported by the German Federal State of Mecklenburg-Western Pomerania and the European Social Fund under grant ESF/IV-BM-B35-0006/12

References

1. A. Agrawal, S. Ramalingam, Y. Taguchi, and V. Chari, "A theory of multi-layer flat refractive geometry," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3346 – 3353.
2. T. Dolereit and A. Kuijper, "Converting underwater imaging into imaging in air," in *VISAPP 2014 - Proceedings of the 9th International Conference on Computer Vision Theory and Applications, Volume 1, Lisbon, Portugal, 5-8 January, 2014*, S. Battiato and J. Braz, Eds. SciTePress, 2014, pp. 96–103.
3. T. Dolereit, U. Freiherr von Lukas, and A. Kuijper, "Underwater Stereo Calibration Utilizing Virtual Object Points," in *OCEANS 2015*, 2015, pp. 1–7.
4. R. Ferreira, J. P. Costeira, and J. A. Santos, "Stereo reconstruction of a submerged scene," in *Proceedings of the Second Iberian conference on Pattern Recognition and Image Analysis - Volume Part I*, ser. IbPRIA'05, 2005, pp. 102–109.
5. N. Gracias and J. Santos-Victor, "Underwater video mosaics as visual navigation maps," *Computer Vision and Image Understanding*, vol. 79, pp. 66 –91, 2000.
6. M. Johnson-Roberson, O. Pizarro, S. B. Williams, and I. Mahon, "Generation and visualization of large-scale three-dimensional reconstructions from underwater robotic surveys," *Journal of Field Robotics*, vol. 27, no. 1, pp. 21–51, 2010.

7. A. Jordt-Sedlazeck and R. Koch, “Refractive calibration of underwater cameras,” in *Proceedings of the 12th European conference on Computer Vision - Volume Part V*, ser. ECCV’12, 2012, pp. 846–859.
8. C. Kunz and H. Singh, “Stereo self-calibration for seafloor mapping using AUVs,” in *Autonomous Underwater Vehicles (AUV), 2010 IEEE/OES*, 2010, pp. 1–7.
9. J. M. Lavest, G. Rives, and J. T. Lapreste, “Dry camera calibration for underwater applications,” *Mach. Vision Appl.*, vol. 13, no. 5-6, pp. 245 – 253, 2003.
10. A. Meline, J. Triboulet, and B. Jouvencel, “A camcorder for 3D underwater reconstruction of archeological objects,” in *OCEANS 2010*, 2010, pp. 1–9.
11. A. Sedlazeck, K. Koser, and R. Koch, “3D reconstruction based on underwater video from ROV kiel 6000 considering underwater imaging conditions,” in *OCEANS 2009 - EUROPE*, 2009, pp. 1–10.
12. A. Sedlazeck and R. Koch, “Perspective and non-perspective camera models in underwater imaging - overview and error analysis,” in *Outdoor and Large-Scale Real-World Scene Analysis*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7474.
13. M. R. Shortis and E. S. Harvey, “Design and calibration of an underwater stereo-video system for the monitoring of marine fauna populations,” *International Archives Photogrammetry and Remote Sensing*, vol. 32, no. 5, pp. 792–799, 1998.
14. A. P. Silvatti, F. A. Salve Dias, P. Cerveri, and R. M. Barros, “Comparison of different camera calibration approaches for underwater applications,” *Journal of Biomechanics*, vol. 45, no. 6, pp. 1112–1116, 2012.
15. T. Treibitz, Y. Y. Schechner, and H. Singh, “Flat refractive geometry,” in *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.* IEEE, Jun. 2008, pp. 1–8.
16. T. Yau, M. Gong, and Y.-H. Yang, “Underwater camera calibration using wavelength triangulation,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2499–2506.

Underwater Image Restoration: Effect of Different Dictionaries

Fahimeh Farhadifard

University of Rostock, Rostock 18051, Germany,
Fahimeh.Farhadifard@uni-rostock.de

Abstract. Ocean engineering has a strong need for clear and high quality underwater images. Capturing a clear scene underwater is not a trivial task since color cast and scattering caused by light attenuation and absorption are common. The poor quality hinders the automatic segmentation or analysis of the images. In this work, an image restoration based on compressive sensing is reported which tackles with blurring caused by light scattering and provides better structural details. Furthermore, the effects of different dictionaries on the quality of restoration is studied. The aim is to use a single degraded underwater image and improve the image quality without any prior knowledge about the scene such as depth, camera-scene distance or water quality.

Key words: Underwater image restoration, Compressive sensing

1 Introduction

Digital imaging and image processing have been established in a wide range of challenging topics, in surveillance tasks, industrial quality assurance, inspection applications and exploration. Autonomous Underwater Vehicles (AUV) and Remotely Operated Vehicles (ROV) are usually employed to explore the deep sea. These robots can reach to the depths where divers cannot operate safely and effectively.

Imaging systems underwater, give poor visibility results. This is due to the light interaction with water and its inherent particles. Light attenuation is an exponential function as it travels in water, and results in a poor contrasted and hazy scene. Visibility in this medium is limited by the light attenuation at distance about twenty meters in clear water and five meters or less in turbid water. The overall performance of underwater imaging system is influenced by the absorption and scattering which are the reasons of light attenuation. Forward scattering is the light which is deviated from its way from the object to the camera and generally leads to blur of the image features. On the other hand, backscattering which is a fraction of light that is reflected back to the camera by water or floating particles before it even reaches to the object, and generally limits the contrast of the images, generating a characteristic veil that superimposes itself on the image and hides the scene.

Furthermore, the amount of light reduces by traveling deeper into water, and colors drop off depending on their wavelengths. According to the selective absorption of water, colors with longer wavelength are much easier to be absorbed, so red light will be absorbed before colors with shorter wavelengths such as the blue and green. On the other hand, based on Rayleigh scattering theory, scattering intensity is inversely proportional to the fourth power of wavelength, so that shorter wavelengths of violet and blue light will scatter much more than the longer wavelengths of yellow and especially red light. As a conclusion, water absorbs the longer wavelength of red and scatters the blue and violet when visible light disseminates in it. Absorption and scattering effects are not only due to the water itself but also due to the components such as a dissolved organic matter so-called marine snow. Although, the visibility range can be increased with artificial illumination of light on the object, but it produces non-uniform of light on the surface of the object and producing a bright spot in the center of the image with poorly illuminated area surrounding it. So, underwater images suffer from limited range of visibility, low contrast, non-uniform lighting, blurring, bright artifacts, color diminished and noise. (The reader is referred to [1] for more on scattering and absorption).

A possible approach to deal with mentioned challenges is to consider the image transmission in water as a linear system [2] and recover the image using some priority knowledge. Image restoration aims to recover the original image $X(i,j)$ from the degraded image $Y(i,j)$ using a model of the degradation and of the original image formation; it is essentially an inverse problem.

Several techniques have been proposed to handle the restoration of underwater images from different perspectives. E. Truco et.al [3], considered the forward scattered component of light as the main reason of degradation and proposed a self tuning restoration filter using a simplified version of the well-known Jaffe-McGlamery image formation [4] [1] by eliminating the backward scattered component. X. Wu and H. Li [5], followed the same fundamental but considered that water made of lots of layers which scatter the same amount of light to take into consideration the relationship between the components of the light that enters into the camera. On the other hand, some methods considered the backward scattered component as the main problem and using statistical approaches, attempted to recover the clear image from hazy version[6][7][8]. For this, using dark channel prior, the depth map of each image is estimated. Once the depth map is calculated, the foreground and background is segmented, then the presence of artificial light is determined and finally, the haze phenomenon will be corrected by considering the effect of artificial light.

In this work, we use a learning based algorithm based on compressive sensing which is proposed [9], in order to recover the blur free underwater image using the simplified image formation of Jaffe-McGlamery presented at [5]. The main goal is to do a study about the results of this method using dictionaries learned by two different image data sets and also two different degradation models in order to find the best combination. We used two image sets, including underwater images and in air images and as degradation model we applied Gaussian blur and blur model proposed at [5].

The rest of this report is structured as follows, first some preliminaries are given in two subsection: compressive sensing and underwater degradation model. In section 3, the approach is explained in details. Simulation results and the study discussion are reported at section 4 and then at last conclusion is placed.

2 preliminaries

Before further proceeding, we will give some preliminaries which are needed to proceed to the algorithm.

2.1 Compressive Sensing

Compressive sensing (CS) has received considerable attention in different fields such as computer science, electrical engineering and mathematics. It suggests that it is possible to surpass the traditional limits of sampling theory. The fundamental of CS is built based on sparse representation of a signal which says a signal can be represented with only a few non-zero coefficients in a suitable domain if the signal is sparse in that domain. Using nonlinear optimization, the recovery of such a signal from very few measurement is possible.

CS enables a potentially large reduction in the sampling and computation costs for sensing signals that have a sparse or compressible representation. While the Nyquist-Shannon sampling theorem states that a certain minimum number of sampling is required in order to perfectly capture an arbitrary bandlimited signal, when the signal is sparse in a known basis we can vastly reduce the number of measurements that need to be stored. The basic idea of CS is that rather than first sampling at a high rate and then compressing the sampled data, we would like to find ways to directly sense the data in a compressed form, (at a lower sampling rate).

This field grew out of the work of Candes, Remberg and Tao and of Donoho who showed that a finite dimensional signal having a sparse or compressible representation can be recovered from a small set of linear, nonadaptive measurements [10][11][12][13].

2.2 Underwater degradation Model

The Jaffe-McGlamery [4][1] underwater imaging model, shows a complete and comprehensive statement of light interaction with water and the corresponding components of light which enters into the camera. McGlamery stated that the total irradiance enters into the camera is the line superposition of the three components, direct, forward scattered and backward scattered.

$$E(i,j) = E_d(i,j) + E_{fs}(i,j) + E_{bs}(i,j) \tag{1}$$

Since the aim is to recover the blurry images and it caused by forward scattered component, therefore, the backward scattered component which has no information about the scene can be ignored and only the relation between the direct component and forward scattered one is considered.

$$E(i,j) = E_d(i,j) + E_{fs}(i,j) \tag{2}$$

where

$$E_{fs}(i,j) = E_d(i,j) * g(i,j/R,G,c,B) \tag{3}$$

and

$$E_d(i,j) = e^{-cR}E_d(i,j,0) \tag{4}$$

where G and B are empirical coefficients, c is the attenuation coefficient and R is the scene-camera distance. Thus, the formation of $g(i,j/R,G,c,B)$ is the key to restore the clear underwater image.

In frequency domain we have:

$$E(f) = S(f)E_d(f) \tag{5}$$

Because of sea water complexity, it is not trivial to come up with an accurate model. The most authoritative model is proposed by Jaffe [4] as the following:

$$S(f) = (e^{-GR} - e^{-cR})e^{-Bzf} + e^{-cR} \tag{6}$$

After some simplifications and optimization proposed at [5], the final degradation model which represent the blurriness caused by forward scattered component is obtained as:

$$E(f) = (1 + k\frac{1-e^{-bf}}{f})E_d(f) \tag{7}$$

where $E(f)$ represents the spatial form of blurry image, and $E_d(f)$ states for spatial form of direct component which represent the clear underwater image.

3 Image Restoration

In the problem of restoration, we are given a degraded image Y and asked to recover the original image X using a prior knowledge such as degradation model. In this report, we are going to make a study about the recovery of blurry underwater images using a learning based algorithm proposed at [9], which is based on compressive sensing theory. The fundamental of the algorithm will be explained first and then the results of study will be discussed.

The method has two stages, first a pair of dictionaries is learned. Dictionaries are linked to each other by the degradation function and have corresponding atoms. Then using the dictionaries together with sparse representation theory, the sparse coefficients are calculated and afterwards, the clear image is recovered using the same sparse coefficients and the dictionary of clear images.

To be more precise, consider degraded underwater image Y which is blurred version of desired clear image . And assume that there is an over-complete dictionary $D_h \in R^{n\times k}$ of k bases which is a large matrix learned using high quality and clear image patches. Then the vectorized patches of image X, $x \in R^n$ can be sparsely represented over dictionary D_h. So the high quality and clear patch x can be represented as $x = D_h\alpha_0$ where $\alpha_0 \in R^n$ is a vector with very few nonzero elements ($\ll k$). The relationship between a high quality and clear image patch x and its degraded counterpart y can be expressed as:

$$y = Lx = LD_h\alpha_0 \tag{8}$$

Note that L represents the blurring model. Substituting the representation for the high quality and clear patch x into (Eq. 8) and noting that $D_l = LD_h$, one gets:

$$y = LD_h\alpha_0 = D_l\alpha_0 \tag{9}$$

Equation (9) implies that the degraded image patch y will also have the same sparse representation coefficients α_0. Now given the degraded image patches, one can obtain the representation coefficients using a vector selection such as OMP[14]. After obtaining the sparse coefficients, one can reconstruct the high resolution patch x.

$$x = D_h\alpha_0 \tag{10}$$

The sparse representation problem (vector selection) has the formulation as an optimization problem which results in finding the sparse coefficient α using dictionary D_l. For obtaining the sparse representation coefficients for the degraded image patch y, one solves the following optimization problem:

$$\min_{\alpha_0} \|y - D_l\alpha_0\|_2 \quad s.t. \quad \|\alpha_0\|_0 < T \tag{11}$$

where T is a threshold which is used to control the sparseness of the representation. The l_0 norm is used to identify the number of nonzero elements of the vector α_0. The level of sparsity can vary depending on the complexity of test signal, higher sparsity can give more accurate representation. An error based formulation of the vector selection problem can also be employed. In order to represent the signal of interest, a suitable dictionary and a sparse linear combination of the dictionary atoms is needed. The sparse representation problem subject to find the most proper selection of those linear combination vectors from an over-complete dictionary D_l. To find such a representation different pursuit algorithms can be used such as OMP[14] and the over-complete dictionary can be formed using K-SVD[15].

3.1 K-SVD approach

As it was mentioned previously, an over complete dictionary together with the sparse coefficients are needed to represent a signal. The joint dictionary learning and sparse representation of a signal can be defined by the following optimization problem:

$$\min_{D,\alpha} \|X - DQ\|_F^2 \quad s.t. \quad \forall i, \|\alpha_i\|_0 < T \tag{12}$$

Consider a set of over-complete basis vector f, and an initial dictionary which is formed by choosing its elements from the set randomly, D. In order to find the sparse coefficients of such a set over the dictionary, once the dictionary is assumed to be fixed and then the sparse coefficients are calculated using OMP by solving following optimization problem for each and every input signal

$$\|y_i - D\alpha_i\|_2^2 \quad s.t. \quad \min_{\alpha_i} \|\alpha_i\|_0 \quad i = 1, 2, ..., N \tag{13}$$

Since the K-SVD algorithm attempts to update dictionary by replacing one atom at a time to reduce the error in representation, thus in every iteration, the dictionary and effective sparse coefficient vectors are considered to be fixed and just one atom in the dictionary is questioned to be replaced and the corresponding sparse coefficient is calculated.

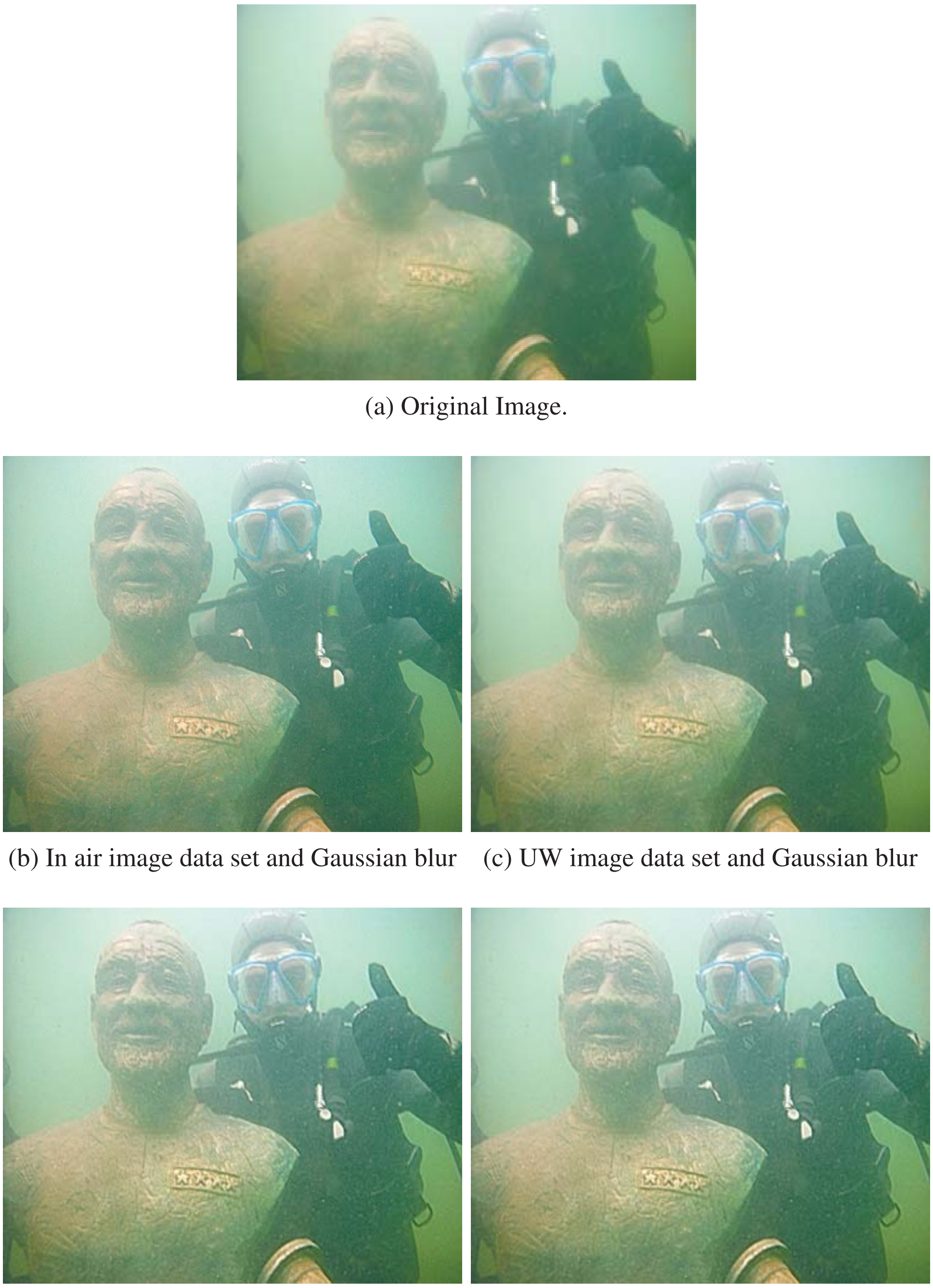
3.2 Orthogonal Matching Pursuit (OMP)

As it was mentioned before, finding an exact sparse representation of a signal is not easily achievable. As the result, many researchers have aimed to find the best approximate solution. Among all the methods Orthogonal Matching Pursuit (OMP) has been the main choice. OMP is a simple method which, enjoys fast running time. Given a dictionary, OMP as a greedy algorithm, aims to find sparse representation of the signals of interest over that dictionary. It is an iteratively algorithm which updates the basis vector in every iteration and as the result reduces the error in the representation. According to this scheme, the dictionary atoms with the largest absolute projection on the error vector are selected. This results into selection of atoms which contain maximum information and consequently reduce the error in the reconstruction.

4 Simulation Results

In order to evaluate the method, we used several underwater images which, were taken in different seas and unknown depths and try to recover the blur free images. Some of the test images are the same as those used in [9]. For this purpose, we learned four different pairs of dictionaries. We used two training sets, one contains in air images and the second one, underwater (UW) images. Then using two possible blur models, Gaussian blur and UW blur model explained above, four possible pair of dictionaries are trained. Once we have all dictionaries, the quality of restored images are studied over some test images which are independent from training data sets. Qualitative comparison is provided to evaluate the results. We did not provide any quantitative comparison such as SNR, since the ground truth data are not available.

As it can be seen in both Fig. 1 and Fig. 2, the reconstructed images using the both training sets where Gaussian blur is used as the degradation model, show almost the same quality (almost identical). Both provide good quality at recovering blur free and detail enhanced images. Experimental results show that, we can achieve better reconstruction if the degradation model is designed specially for underwater situation. Training dictionaries using the UW blur give results with sharper edges and better contrast while the details are more enhanced (Fig.3). But this does not apply for results of the dictionaries with in air training set. Fig. 4, shows that clearly by the artifacts, and overcompensation which gives an unnatural appearance to the image.



(a) Original Image.

(b) In air image data set and Gaussian blur

(c) UW image data set and Gaussian blur

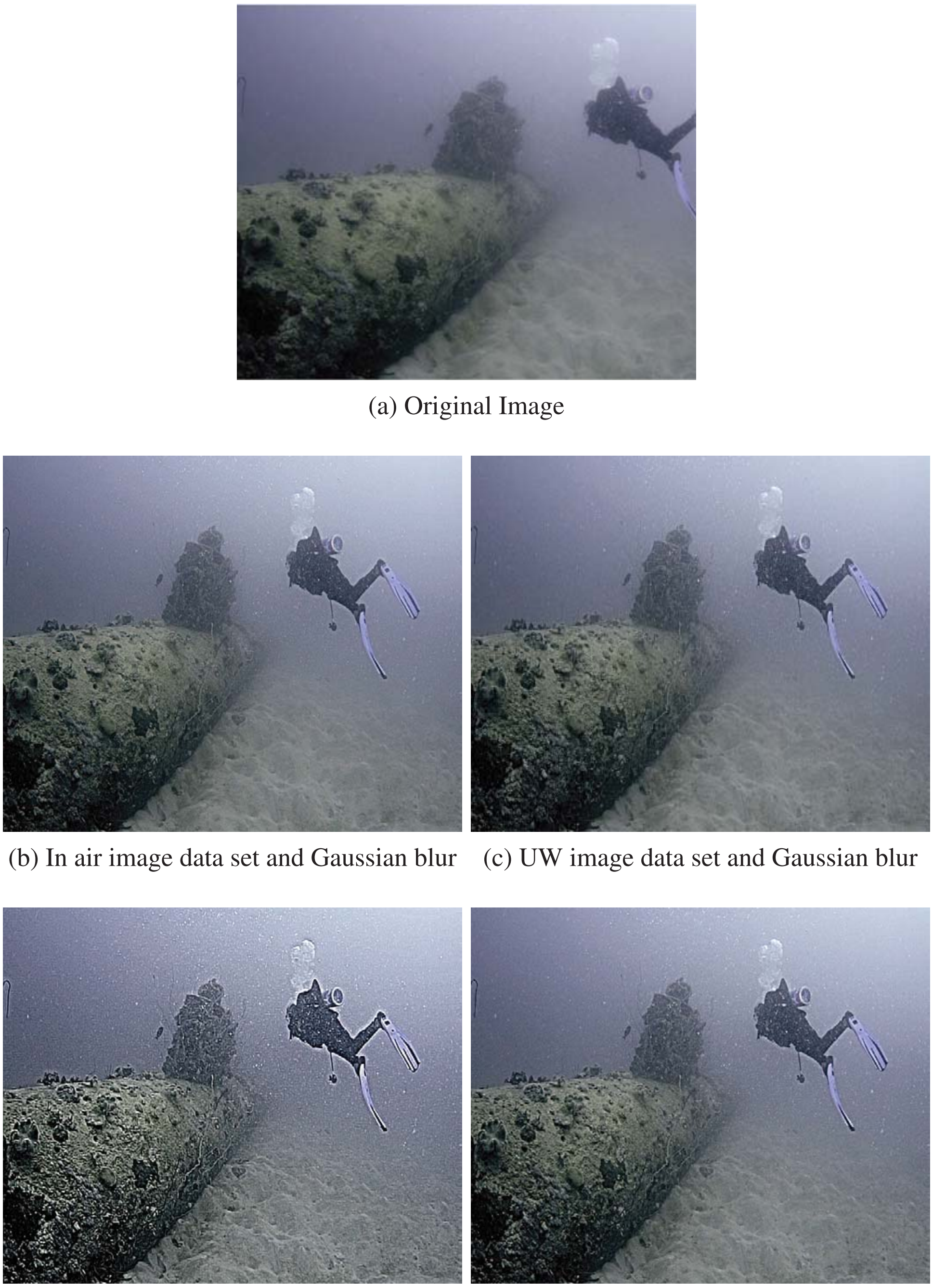
(d) In air image data set and UW blur

(e) UW image data set and UW blur

Fig. 1: Comparison of the results using different dictionaries and degradation models. Original image is Courtesy of C. Ancuti et al. [16]

5 Conclusion

In this paper, we reported an underwater image restoration method based on compressive sensing and further, studied the effects of different dictionaries in the final result.



(a) Original Image

(b) In air image data set and Gaussian blur

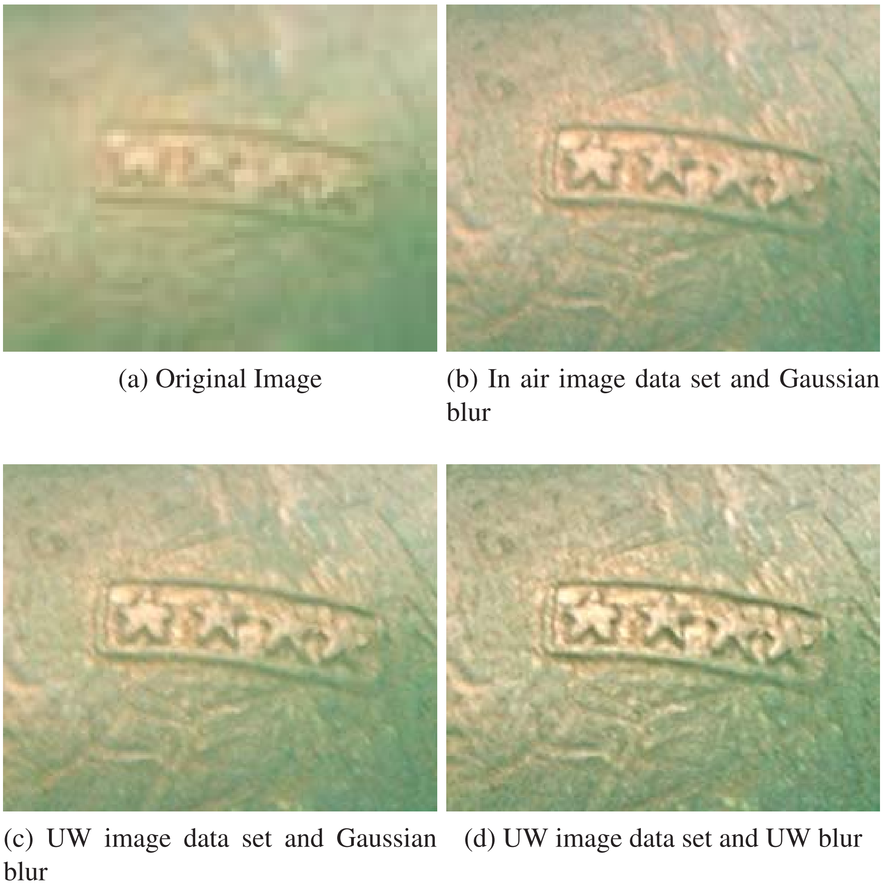
(c) UW image data set and Gaussian blur

(d) In air image data set and UW blur

(e) UW image data set and UW blur

Fig. 2: Comparison of the results using different dictionaries and degradation models.

Based on experimental results, we can recover a blurry underwater image without any knowledge about the depth or camera-scene distance while the local contrast is enhanced and better structural details are provided. The study illustrates that, when we learn dictionaries using UW image data set which have similar statistical nature as in-



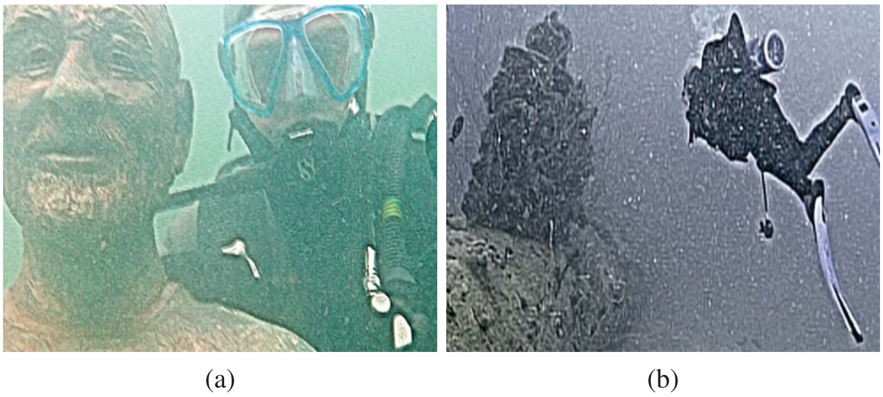
(a) Original Image

(b) In air image data set and Gaussian blur

(c) UW image data set and Gaussian blur

(d) UW image data set and UW blur

Fig. 3: Zoom-in view of the results



(a)

(b)

Fig. 4: Zoom-in view of results by in air data set and UW blur model.

put and specific blur model caused by forward scattered light, the best reconstruction can be achieved while UW images keep their natural appearance.

6 Acknowledgement

This research has been supported by the German Federal State of Mecklenburg-Western Pomerania and the European Social Fund under grants ESF/IV-BM-B35-0006/12 and V630-S-179-2013/238.

References

1. B. L. McGlamery. A computer model for underwater camera systems. volume 0208, pages 221–231, 1980.
2. L. E. Mertens and F. S. Replogle. Use of point spread and beam spread functions for analysis of imaging systems in water. *the Optical Society of America*, 67:1105–1117, 1977.
3. E. Trucco and A. T. Olmos-Antillon. Self-tuning underwater image restoration. *Oceanic Engineering, IEEE Journal*, 31(2):511–519, 2006.
4. J.S. Jaffe. Computer modeling and the design of optimal underwater imaging systems. *Oceanic Engineering, IEEE Journal*, 15(2):101–111, Apr 1990.
5. Xiaojun Wu and Hongsheng Li. A simple and comprehensive model for underwater image restoration. In *Information and Automation (ICIA), 2013 IEEE International Conference on*, pages 699–704, Aug 2013.
6. Haocheng Wen, Yonghong Tian, Tiejun Huang, and Wen Gao. Single underwater image enhancement with a new optical model. In *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*, pages 753–756, May 2013.
7. H. Y Yang, P. Y Chen, C. C Huang, Y. Z Zhuang, and Y. H Shiau. Low complexity underwater image enhancement based on dark channel prior. In *Innovations in Bio-inspired Computing and Applications (IBICA), 2011 Second International Conference*, pages 17–20, Dec 2011.
8. J. Y. Chiang and Y. C Chen. Underwater image enhancement by wavelength compensation and dehazing. *Image Processing, IEEE Transactions*, 21(4):1756–1769, April 2012.
9. Fahimeh Farhadifard, Zhiliang Zhou, and Uwe Freiherr von Lukas. Learning-based underwater image enhancement with adaptive color mapping. In *Image and Signal Processing and Analysis, 9th International Conference on*, pages 50–55, September 2015.
10. Emmanuel J Candès et al. Compressive sampling. In *Proceedings of the international congress of mathematicians*, volume 3, pages 1433–1452. Madrid, Spain, 2006.
11. Emmanuel J Candes and Justin Romberg. Quantitative robust uncertainty principles and optimally sparse decompositions. *Foundations of Computational Mathematics*, 6(2):227–254, 2006.
12. David L Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.
13. Richard G Baraniuk. Compressive sensing. *IEEE signal processing magazine*, 24(4), 2007.
14. Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference*, pages 40–44 vol.1, Nov 1993.
15. M. Aharon, M. Elad, and A. Bruckstein. K -svd: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions*, 54(11):4311–4322, 2006.
16. C. Ancuti, C.O. Ancuti, T. Haber, and P. Bekaert. Enhancing underwater images and videos by fusion. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 81–88, June 2012.

Comparison of Spatial Models for Foreground-Background Segmentation in Underwater Videos

Martin Radolko

University of Rostock, Rostock 18051, Germany,
Martin.Radolko@uni-rostock.de

Abstract. The low-level task of foreground-background segregation is an important foundation for many high-level computer vision tasks and has been intensively researched in the past. Nonetheless, unregulated environments usually impose challenging problems, especially the difficult and often neglected underwater environment. There, among others, the edges are blurred, the contrast is impaired and the colors attenuated. Our approach to this problem uses an efficient Background Subtraction algorithm and evaluates it in combination with different spatial models.

Key words: Background Subtraction, Gaussian Switch Model, Markov Random Fields, Belief Propagation, NCut

1 Introduction

Nowadays Computer Vision Systems are used in various fields of applications such as automation, surveillance, human assistance or inspection. Background Subtraction has been used for many years for Computer Vision problems but is still a very valuable source for low level information. It can recognize almost arbitrary objects in any scene, as long as they are in motion. This information can later be reprocessed in different high-level vision tasks.

In order to gather information about the objects of interest in a specific scene, the background of this scene has to be modeled. The task of creating and sustaining an adequate background model is not trivial and associated with many difficulties like changes in the lightning conditions, slightly moving background objects or shadows. A large number of different approaches have been developed to tackle these requirements and create an adequate background model even under harsh conditions. Some use Subspace Learning Models like LDA [1], INMF [2] or PCA [3] to model the background. Other renowned methods adopted techniques like Kalman Filters [4], SVMs [5] or Histograms [6] to background modeling in the hope that they could better cope with these problems.

However, the most promising and common method at the moment is a statistical approach where each background pixel is modeled as a Gaussian Distribution. This approach is justified by the fact that the intensity of a pixel in a completely static scene will vary over time according to a Normal distritution $\mathcal{N}(\mu,\sigma^2)$ due to the inevitable measurement errors inherent in every camera system. A threshold per pixel and channel can be easily derived from the mean and variances of these distributions to distinguish between foreground and background.

There exist some approaches which use just one Normal Distribution per pixel [7], algorithms which use a Mixture of Gaussians [8, 9] or Gaussian-Kernel based methods [10] to model the background. Methods which use a Mixture of Gaussians (MoG) produce in most cases better results than the Single Gaussian (SG) algorithms, because they can model difficult situations (like a swaying tree) better if they are correctly adjusted. Nevertheless, they have a higher memory usage and more parameters which need to be carefully tuned.

A common disadvantage of all Background Subtraction approaches is the missing incorporation of spatial information about the scenes in the model. Natural images are assumed to be very smooth because they usually depict real objects like trees,animals or buildings. This assumption can be used to improve the segmentation which was derived from the Background Subtraction. An example of this is [11], where a simple method is used which erases all connected areas containing less than a certain amount of foreground (or background) pixels. A more sophisticated approach is applied in [12] where a Conditional Random Field models the neighbourhood relations of the pixels. Graph Cuts are used in [13] and in [14] the spatial information is represented in a tensor to whom a Subspace Learning algorithm is applied. The usage of a tensor ensures that all dimensions are treated equally.

We implemented two different spatial models to test both on underwater videos. The first method is based on the popular Normalized Cut (NCut) approach which has been used intensively for single image segmentation [15, 16]. Nonetheless, it has never been applied on videos because of some inherent characteristics which make the NCut unsuitable for videos (see section 2.2) and the high computational costs which forbid any real time application. The first problem can be adressed with a reformulation of the NCut, which adopts it to some video specific requirements. The second problem can be solved by the usage of a simple and fast local optimization algorithm which lowers the computational cost significantly.

The second approach uses Markov Random Fields to represent the spatial relation in the image. This model consists of an undirected graph which is underlaid with a probability map. The probabilites for each pixel are deduced from the Background Subtraction. Nonetheless, the most important model parameter is the neighborhood system, which is a generalized Moore Neighborhood in our case. These large neighborhood systems can model the natural smoothness in images better than the simple 4-connected Neighborhoods normally used.

All of these approaches have been optimized for air images and have not been evaluated on the more difficult underwater images[17, 18, 8, 16]. In the results section we used different self-made underwater videos to compare the two approaches for spatial modeling. Although Markov Random Fields fall behind in accuracy on air images, the same method wins clearly on the difficult underwater images. This result suggest that more special analysis should be made for underwater images and that maybe special algorithms are required for the same task there.

2 Our Approach

In the first part of this section we will explain the Background modeling with the Gaussian Switch Model (GSM) and Background Subtraction with a voting algorithm[19]. The second part describes the N^{2}Cut [20] as a new spatial model for video segmentation and in the last segment a Markov Random Field combined with a Belief Propagation algorithm is introduced as another spatial model. Both of them will be evaluated on underwater images in the results section.

2.1 The Gaussian Switch Model

As justified previously, Gaussian distributions are used to model the colour values of each pixel. The most obvious approach would be a batch method where the n last pictures are saved and then for each pixel and channel the best fitting normal distribution is calculated for the given data. However, this is extremely resource demanding, wherefore we use running gaussians instead. This method just updates the old Gaussians with the values from the newest frame and does not compute the distributions over the n last data points from scratch every time. Thereby, the algorithm gives the new pixel values automatically a higher weight than old values and thus even improves the results in comparison to the batch method because the newer samples usually carry more information about the current background. To be exact: for every Gaussian, the mean μ and variance σ^2 have to be computed. The mean is initiated with the pixel value taken from the first frame of the video stream and the variance is set to a predefined value. Afterwards, they are updated in the following way

$$\mu^{t+1} = \alpha\, \mu^t + (1-\alpha)\, v^t, \tag{1}$$

$$(\sigma^{t+1})^2 = \alpha\, \sigma^t + (1-\alpha)(\mu^t - v^t)^2. \tag{2}$$

The variable α is the update rate and v^t is the pixel value taken from the t-th frame. With these formulas, the Gaussian distribution of a background pixel can be approximated very efficiently.

Nevertheless, one problem is that the model becomes erroneously when a foreground object is visible because it starts to model the foreground object and not the background. To cope with this problem another Gaussian is introduced, the Background Gaussian $\mathcal{N}(\mu^{bg}, (\sigma^{bg})^2)$, which is updated after the segmentation process and only if the corresponding pixel is classified as background. This results in a more stable background model as the corruptions from foreground objects are minimized. Nonetheless, there is an inherent problem with this because the model now only accepts values which agree with the current model and acts like a self fullfilling prophecy. One issue are foreground objects already visible in the first frame. At the beginning the model will assume them as background and afterwards will never include the real background into the model because the real background will be classified as foreground. Foreground objects that become background, e.g. a car that parks, will also never get included into the background model for the same reasons. To eliminate these errors a second Gaussian, the Overall Gaussian $\mathcal{N}(\mu^{og}, (\sigma^{og})^2)$, is introduced, which will be updated with every new frame.

If a foreground object was visible but immoble for a long period of time, it should be included into the background. Such events result in an Overall Gaussian with a small variance and a mean which is different from the Background Gaussian mean. If such an incident is detected, the Background Gaussian is set to the values of the Overall Gaussian, so that the Object gets included into the background model. This model is applied to every pixel and every channel seperately and later a voting algorithm is used to unify the results and get a definite label for each pixel.

To make the best use of the color information, a special color space is used, which normalises the different intensities in respect to the illumination [14]. Let R, G and B be the given values for a single pixel in the standard RGB color space, then these will be transformed into the three new image channels

$$\begin{aligned} I &= R + B + G, \\ \tilde{R} &= R/I, \\ \tilde{B} &= B/I. \end{aligned}$$

Afterwards the intensity I is scaled to the range $[0,1]$. The color information stored in $\tilde{R}$ and $\tilde{B}$ are normalised with the intensity and will thus not be altered by small or medium changes in the lightning conditions. This can be used to prevent the detection of shadows as foreground. However, if the shadow is very strong the color information may be completely lost in the image and this approach will fail.

At the end a thresholding is applied at each channel seperately. The statistical approach allows to get an adaptive threshold for each pixel and channel. If the variance in the statistical background model is low (high) the noise level at the corresponding pixel and videoframe can also be expected to be low (high). Hence, the variance can be used as an threshold. If p_R is the new value for the red-channel of a pixel, the thresholding inequality is given by:

$$(p_R - \mu_R^{bg})^2 < \max(\beta \cdot (\sigma_R^{bg})^2, 0.001). \tag{3}$$

The maximum is used because the variance could approach near zero values, especially since only matching values are included into the Background Gaussian. The parameter β can control the range of values which are still classified as “matching the model”.

To derive a decision for a pixel as a whole a voting procedure is chosen. If equation (3) is satisfied for at least two of the three channels, the pixel is marked as background, otherwise as foreground. Thereby, the color information can overrule the brightness information and hence shadows should not be detected as foreground. At the end of this process a pixelwise foreground-background segregation is derived only from the temporal information of the video.

2.2 N^{2}Cut

To incorporate spatial information into this segmentation we evaluated two different methods. The first is based on the NCut. In this approach the image is transformed into a graph with a von Neumann Neighborhood to evaluate the best cut. To create an adequate spatial model the weights of the edges in this graph have to be chosen very carefully. We defined the weight of the edge between the nodes i and j (depicted as w_{ij}) by the Manhattan distance of the corresponding color values.

$$w_{ij} = |r_i - r_j| + |g_i - g_j| + |b_i - b_j| \tag{4}$$

The use of this quite simple metric reduces the computational complexity of building the model. Nonetheless, the weights are accurate enough to build reliable spatial models which produce good segmentation results.

Graphs like these have been used many times in segmentation algorithms [21]. In most cases, an energy function is defined on the graph to evaluate a specific segmentation. This transforms the image segregation problem into a well-known minimization task. In the literature, there are different approaches for this, one example is [22], who use the cut-value as an energy function. A more elaborated energy function is NCut. It maximizes the association in the different regions while minimizing the cut between them [18, 23].

Approaches using NCut usually provide better results but finding the optimal solution is an NP-hard problem [23] which makes approximative methods necessary for the optimization step (e.g. spectral graph theory). Given a weighted graph $G = (V, E, w)$ and a partition $A \dot{\cup} B = V$ the NCut for that partition (segmentation) is defined as follows:

$$Ncut(A,B) = \frac{Cut(A,B)}{Assoc(A)} + \frac{Cut(A,B)}{Assoc(B)} \tag{5}$$

with the standard $Cut(A,B)$ and $Assoc(A)$ terminologies.

$$Assoc(A) = \sum_{i \in A, j \in V} w_{ij} \tag{6}$$

$$Cut(A,B) = \sum_{i \in A, j \in B} w_{ij} \tag{7}$$

This energy function is well suited for the evaluation of segmentations in single images but not for videos. There it can occur that the scene is completely free of foreground objects. These cases cannot be mapped by NCut as an 100% background segmentation would result in a division by zero. Therefore, this energy function inherently works with the false assumption that there are always foreground objects visible. Furthermore, NCut also favors segmentations with roughly equal amounts of fore- and background. If there is only a small amount of foreground, the corresponding association will attain a very small value and hence one of the summands in equation 5 will become very large. This prevents segmentations with only minor foreground or background areas.

These problems can be adressed with a modified normalized cut [20] which has no bias for any specific amount of foreground:

$$N^2cut(A,B) = \frac{Cut(A,B)}{nAssoc(A)} + \frac{Cut(A,B)}{nAssoc(B)}, \tag{8}$$

$$nAssoc(A) = \frac{Assoc(A) + 1}{\sum_{i \in A, j \in V, \exists e_{ij}} 1 + 1}. \tag{9}$$

In equation 9 the association is normalized by dividing by the number of edges contributing to the association. Consequently, the new association is the average edge value which is not dependent on the size of the set. The addition of one to the denominator and numerator of the fraction in equation 9 prevents the divisions by zero for empty sets. An obvious extension to this seems to be the normalisation of Cut(A) in the same way, but this is not reasonable. It would remove the favor of cuts that are short and would hence result in very long cuts zigzagging through the image. This would not reflect the smoothness of natural images.

The N^{2}Cut is based on the spatial information in one single image only. To get meaningful segmentations the temporal information derived form the GSM Background Subtraction have to be added. This is done by taking the GSM segmentation as a starting point for the N^{2}Cut optimization. Based on this a local optimization process is run and produces the final segmentation. It is important that the optimization algorithm is only local and may get stuck in local minima because this ensures that the basic structure of the segmentation is derived from the temporal information (GSM) and the N^{2}Cut optimization only makes it spatially coherent.

2.3 Markov Random Fields

Another way to add spatial information to the GSM results are Markov Random Fields (MRF). To achieve this the MRF described in [19] is used. It models the spatial relations between single pixels and hence forces the segmentation to be locally coherent.

The most important part of a MRF is the neighborhood system. We use a generalized Moore Neighborhood because it assures the homogeneity of the MRF and also can easily be changed in size. In the generalized Moore Neighborhood, the neighborhood for a pixel is defined by a square which is centered at that pixel and which can vary in size. The number of different combinations of neighbouring pixels (cliques) will increase radically with the size of the square. The input data, probabilities of being background or foreground for each pixel, is derived from the GSM Background Subtraction.

After constructing the MRF model of the spatial relations of the image, the most likely state (segmentation) of that model has to be computed. This maximum a posteriori (MAP) is very difficult to compute and can only be approximated for a problem of reasonable size. First a cost function is needed which can evaluate the different segmentations based on the MRF model. This function consists of two parts, one part measures how good the segmentation matches the GSM result. Basically, the smaller w_i, the higher is the penalty for labeling the pixel i as foreground. The second part of that function evaluates the spatial coherence of the segmentation. As our assumption is that natural images are smooth, neighbouring pixels should have the same label. If this is violated, there will be a penalty to the cost function.

This cost function is then converted to a factor graph and optimized with a loopy max-product Belief Propagation algorithm. Although this will only approximate the MAP, it can still take a long time and requires a lot of memory to do so. This is due to the fact that the amount of cliques increases so drastically with the size of the neighborhood. To reduce this effect we decided to simplify the model and only take one clique size (the largest cliques) into account. Also, the spatial component of the energy function was kept as simple as possible to further reduce the computational load. It returns zero if all neighbours of the pixel have the same label and one if at least one neighbour has a different label. These simplifications allowed us to build and optimize the MRF model on an 1920×1080 image in around one minute. Without them it would have been infeasible to do so in less than a week.

3 Results

To evaluate these algorithms we tested them first on the popular but old wallflower dataset. The results can be seen in Table 1 and show that our methods perform quite well in air and that N^{2}Cut clearly outperforms the Markov Random Fields there. Additionally to the accuracy increase, the optimization of the N^{2}Cut is also 2 orders of magnitude faster and can be done in real time.

For the evaluation in underwater environments no data sets are freely available at the moment. Hence, we took some underwater videos ourselves with a Go Pro Hero 3 and manually created some ground truth data for them. Two frames of these videos and the corresponding segmentations can be seen in Fig. 1. To measure the accuracy of these segmentations we use the F1-Score and Matthews Correlation Coefficient [24]. They are a better indicator of the quality of segmentations than the simple amount of wrongly classified pixels (which is the standard measure for the Wallflower dataset and was also used here for comparison reasons), especially when the amount of foreground is very small. The reason for this is that, the weight of foreground and background pixels changes according to the amount of foregound visible in the image.

In both pictures the N^{2}Cut performs substantially worse than the MRF algorithm (see Table 2). In the right image even the GSM Background Subtraction without any spatial model is better. This behaviour is quite constant in all the underwater videos we took, although not as strong as in these two selected examples. The reason for this is that the MRF approach smoothes the segmenation just based on the background subtraction result as opposed to the N^{2}Cut which alligns the segmentation to the nearest edges in the image. However, this allignment fails in underwater images because the blurring impedes any clear edges. This behavior is enhanced by the often low color disparity between fishes and the background, which enables them to hide from enemies. In the end, instead of aligning the segmentation to the edges the N^{2}Cut often degenerates foreground objects to simple rectangles because there are no clear edges to which the object can be alligned to. All in all, MRF is better suited as a spatial model in underwater situation if real-time capability is not an issue.

4 Future Work

In the future, we want to use some underwater image enhancement algorithms (mainly deblurring and color correction methods) on the images before the segmentation process starts. We hope that these will allow the N^{2}Cut to perform better and will give the same accuracy and speed advantages in underwater circumstances as it does achieve for air images.

Algorithm	**Errors**
Single Gaussian [7]	35133
Mixture of Gaussian (MoG) [8]	27053
Kernel Density Estimation [10]	26450
MoG with PSO [25]	13916
MoG in improved HLS Color Space [9]	9739
MoG with MRF [17]	3808
Gaussian Switch Model (GSM) [this paper]	9718
GSM with MRF [this paper]	7169
GSM with N^{2}Cut [this paper]	5064

Table 1. The results of different algorithms on the Wallflower [11] data set.

		GSM	GSM + MRF	GSM + N^{2}cut
Left Image	F1-Score:	0.990687	0.991428	0.982705
	MCC:	0.852013	0.879739	0.796699
Right Image	F1-Score:	0.995831	0.996647	0.996094
	MCC:	0.424601	0.540039	0.43656

Table 2. The F1-Score and Matthews Correlation Coefficient for the different segmentations in Fig. 1.

Acknowledgements

This work has been funded by the federal state of Mecklenburg-Vorpommern and EFRE within the project Basic and Applied Research in Interactive Document Engineering and Maritime Graphics.

References

1. Tae-Kyun Kim, Kwan-Yee Kenneth Wong, B. Stenger, J. Kittler, and R. Cipolla. Incremental linear discriminant analysis using sufficient spanning set approximations. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007.
2. S.S. Bucak, B. Gunsel, and O. Guersoy. Incremental nonnegative matrix factorization for background modeling in surveillance video. In *Signal Processing and Communications Applications, 2007. SIU 2007. IEEE 15th*, pages 1–4, June 2007.
3. Marghes, Bouwmans T., and Vasiu R. Background modeling and foreground detection via a reconstructive and discriminative subspace learning approach. In *Proceedings of the 2012 International Conferecne on Image Processing, Computer Vision and Patternrecognition*, pages 106–113, 2012.
4. G.T. Cinar and J.C. Principe. Adaptive background estimation using an information theoretic cost for hidden state estimation. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 489–494, July 2011.
5. Horng-Horng Lin, Tyng-Luh Liu, and Jen-Hui Chuang. A probabilistic svm approach for background scene initialization. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 3, pages 893–896 vol.3, June 2002.

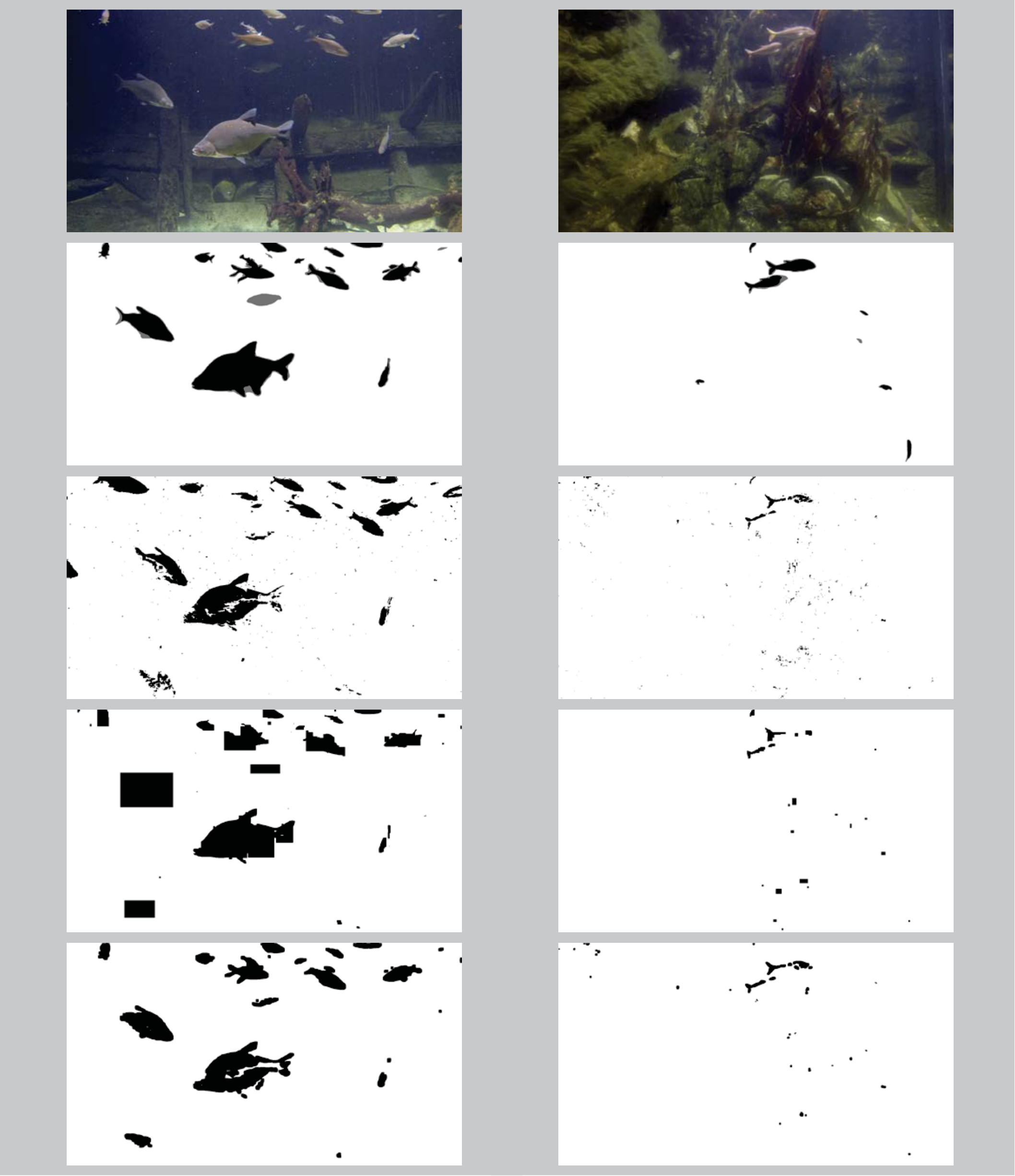


Fig. 1. Two exampls frames of the underwater images we took. The first row shows the original frame, the second row the hand segmented ground truth data, then the result after the background subtraction (GSM), the next row shows the segmentation after combining the N^{2}Cut with GSM and the last row shows the combination of GSM and MRF.

6. Shengping Zhang, Hongxun Yao, and Shaohui Liu. Dynamic background subtraction based on local dependency histogram. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(07):1397–1419, 2009.
7. Christopher Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:780–785, 1997.

8. Chris Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Vol. Two*, pages 246–252. IEEE Computer Society Press, June 1999.
9. N. Setiawan, S. Hong, J. Kim, and C. Lee. Gaussian mixture model in improved ihls color space for human silhouette extraction. In *16th Int Conf on Artificial Reality and Telexistence*, pages 732–741, 2006.
10. Ahmed M. Elgammal, David Harwood, and Larry S. Davis. Non-parametric model for background subtraction. In *Proceedings of the 6th European Conference on Computer Vision-Part II*, ECCV '00, pages 751–767, London, UK, UK, 2000. Springer-Verlag.
11. Kentaro Toyama, John Krumm, Barry Brumitt, and Brian Meyers. Wallflower: Principles and practice of background maintenance. In *Seventh International Conference on Computer Vision*, pages 255–261. IEEE Computer Society Press, Septempber 1999.
12. K.-F. Loe Y. Wang and J.-K. Wu. A dynamic conditional random field model for foreground and shadow segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, pages 279–289, 2006.
13. Yuri Boykov and Gareth Funka-Lea. Graph cuts and efficient n-d image segmentation. *International Journal of Computer Vision*, 70:109–131, November 2006.
14. Xi Li, Weiming Hu, Zhongfei Zhang, and Xiaoqin Zhang. Robust foreground segmentation based on two effective background models. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, MIR '08, pages 223–228, 2008.
15. P. Pouladzadeh, S. Shirmohammadi, and A. Yassine. Using graph cut segmentation for food calorie measurement. In *Medical Measurements and Applications (MeMeA), 2014 IEEE International Symposium on*, pages 1–6, June 2014.
16. R. Hansch, O. Hellwich, and Xi Wang. Graph-cut segmentation of polarimetric sar images. In *Geoscience and Remote Sensing Symposium, 2014*, pages 1733–1736, 2014.
17. Konrad Schindler and Hanzi Wang. Smooth foreground-background segmentation for video processing. In *Proceedings of the 7th Asian Conference on Computer Vision - Volume Part II*, ACCV'06, pages 581–590, 2006.
18. M.A.G. de Carvalho, A.L. da Costa, A.C.B. Ferreira, and R. Marcondes Cesar Junior. Image segmentation using component tree and normalized cut. In *Graphics, Patterns and Images (SIBGRAPI), 2010 23rd SIBGRAPI Conference on*, pages 317–322, Aug 2010.
19. Martin Radolko and Enrico Gutzeit. Video segmentation via a gaussian switch background-model and higher order markov random fields. In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications Volume 1*, pages 537–544, 2015.
20. Martin Radolko, Fahimeh Farhadifard, Enrico Gutzeit, and Uwe Freiherr von Lukas. Real time video segmentation optimisation with a modified normalized cut. In *Image and Signal Processing and Analysis, 9th International Conference on*.
21. Faliu Yi and Inkyu Moon. Image segmentation: A survey of graph-cut methods. In *Systems and Informatics (ICSAI), 2012 International Conference on*, pages 1936–1941, May 2012.
22. Yanmin Peng and Rong Liu. Object segmentation based on watershed and graph cut. In *Image and Signal Processing (CISP), 2010 3rd International Congress on*, volume 3, pages 1431–1435, Oct 2010.
23. Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, Aug 2000.
24. B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, 405:442–451, 1975.
25. B. White and M. Shah. Automatically tuning background subtraction parameters using particle swarm optimization. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 1826–1829, July 2007.

Survey: Hidden Markov Model Based Approaches for Hand Gesture Recognition

Amin Dadgar

Technical University of Chemnitz, 09126 Chemnitz, Germany
WWW home page: https://www.tu-chemnitz.de/informatik/GDV/
seyed-amin.dadgar@informatik.tu-chemnitz.de,

Abstract. In this paper, different technologies and approaches to design hand gesture recognition (HGR) systems are discussed. Then an important technique to address the main phase of such systems (e.g. hidden Markov model for training phase) together with the advantages and limitations of that are illustrated. Furthermore, some variants and solutions of the model that researchers proposed to overcome these limitations is stated. Finally, two different methodologies that are currently under design and implementation, to overcome some other disadvantages that the standard hidden Markov model is facing are discussed briefly.

Key words: Hand Gesture Recognition, Vision-Based Technology, Model-Based (Generative) Approach, Hidden Markov Model, Language-based Pose-Gesture Space Relation, Non-Uniform State Transition Topology

1 Introduction

Many science fiction books and movies show the controlling and commanding of objects by a user while moving and rotating his/her hand in an empty space. Although those media report that this takes place at the mid or end of 21 century, the touchless technology, both as hardware and software, may arrive many decades earlier [9]. To that end, intelligent vision-based hand gesture recognition (HGR) system plays a crucial role and thus has gained much attention of the research community in recent years. Such a system does not only enhance the scientific and industrial applications, but also changes the way of life.

As an example, if one looks at the evolution of human-computer interaction (HCI): text-based interfaces, keyboard, 2D graphical-based interface, mouse, pen, touch screen multimedia-supported interface, fledged multi-participant virtual environment; a 3D mid-air application [1] in which one can select, move, copy and in general interact with objects on the screen simply by moving and rotating one's hand (all without touching any input device) seems to have high potential in the future market [9].

We can categorize hand gesture recognition (HGR) technologies into two broad classes. One class of technology to recognize hand gesture is the marker-based approach in which the subject wears data gloves or puts on optical or mechanical sensors. These devices are for digitizing the hand and finger motions into multi-parametric data. Although it is accurate, its high cost and the difficulty of wearing its tools limit its applicability in many real life scenarios to a great extent [9]. The other class of technology is vision-based approach which provides a noninvasive, easy and natural environment. This class of solutions, however, is still far from generic utility and many technical issues (mathematical, software and hardware) are still to be overcome [9]. Therefore, researchers have systematically considered two broad approaches to overcome vision-based problems: model-based approaches and appearance-based approaches.

The appearance-based approaches which can also be considered as "discriminative design approaches", use the parameters directly derived from images or videos using a template database. In other words, as much information as possible (e.g. color, texture and motion) are extracted directly from the input (e.g. image or the sequence of images), step-by-step. These pieces of information are then combined into one single feature vector and compared with the parameters of the training data (e.g. another sequence of hand gestures). They have the advantage of being real time [9]. However, their disadvantages are: they are very sensitive to lighting conditions and cluttered background [9], their performance is influenced a lot by camera movements and specific user variances [7], they have difficulties dealing with ill-poses or non-singular problems [7].

The model-based approaches which can also be considered as "generative design approach", use the 3D information of the key elements such as palm position and joint angles to "model" the hand skeleton (see Fig. 1). Based on this 3D kinematic hand model, one compares the input images to the possible 2D appearance, projected from the 3D hand model, to estimate the hand parameters. These approaches are ideal for realistic interactions in virtual environments. However their disadvantages are: 1. The initial parameters for each frame have to be close to the solution to prevent the system from being stuck at local minima, 2. They are sensitive to noises in the imaging process, 3. Inevitable self-occlusions of the hand cannot be handled, 4. Very large image databases are required to cover articulated deformation and different views, 5. They are computationally expensive [7].

This paper addresses a vision-based hand gesture recognition design approach. It focuses on model-based approaches within a dynamic Bayesian network (DBN) framework by acquiring hidden Markov model (HMM) techniques [19]. Therefore, the following phases need to be considered.

Hand Gesture Recognition Pipeline

Phase 1. 3D model of the hand: The focus of this phase is to model a hand (skeleton, mesh or both) in order to store its finger joints positional and angular relations in the three-dimensional world and represent its different posture and gestures configurations. Moreover, for the system to be able to compare this 3D model with 2D image cues (e.g. input), calibration and projection data and their correspondence matrices should be taken into account. In that way we can easily restore the information loss when relating the 3D and 2D worlds.

Phase 2. Hand features extraction and tracking: This phase is aimed to detect, track and localize the palm/wrist center at the scene while it is moving in the 2D image plane. In other words, here we only consider the 2D features as input and output, and therefore many techniques that use appearance-based methods (because they acquire parameters directly from the images) can be utilized to enhance the performance.
Phase 3. Training the spatial-temporal a-priori: This phase focuses on training the system with a set of particular gestures and their parameters obtained by motion capture (MoCap) data. In other words, we provide the system with the solutions (the training set) and teach the system, how the corresponding gestures should look like. A well trained system with a sufficient number of similar gestures will then be able to recognize these gestures whenever it receives them as the input. Therefore, having a robust training method and a big and diverse data set is of importance.
Phase 4. Dimensionality reduction: Since the state space of an articulated object is high-dimensional, highly nonlinear with a large number of different degrees of freedom (DOF), this phase focuses on extracting the similarities between deviation of each different dimensions of the data and representing the state space in a low dimension of space to reduce time and computational complexities.
Phase 5. Inference of the posterior probability of the gesture: Due to the loss of information in the imaging process (e.g. image acquisition from 3D to 2D), hand gesture recognition is mathematically classified within the family of ill-pose problems. Therefore the solution can only be estimated within stochastic frameworks. The idea here is to use a data set (e.g. a test set) as input and on this basis the most probable gesture would be returned by increasing the posterior probability.

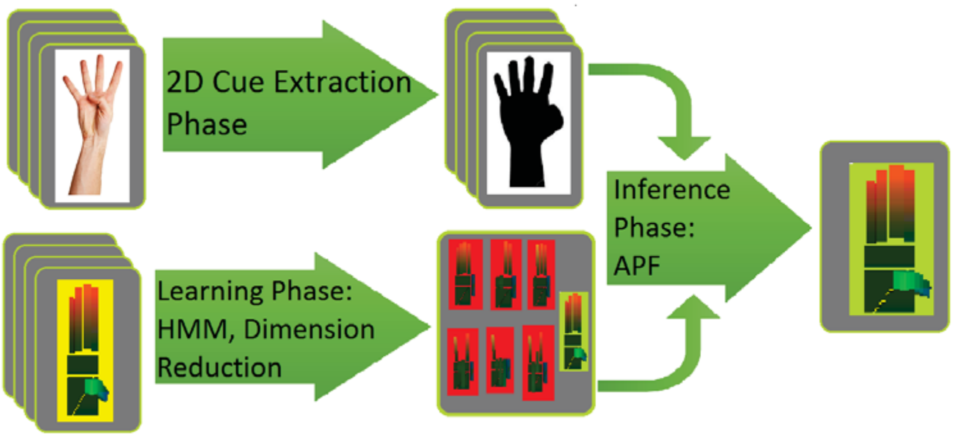


Fig. 1. Model-based framework

Amongst these phases, the focus of the paper lies on the training techniques (e.g. phase 3). In that context, the most frequently used framework is the so-called hidden Markov model (HMM). Its advantages and limitations together with some of the solutions proposed in the scientific community are investigated.

2 Related Work

Since Rabiner [16] applied HMM in a speech processing application for the first time (e.g. as a form of sequential data processing), arguably, it becomes the most suitable framework to automatically process any spatial-temporal data (e.g. hand gestures) [19, 15, 22, 13]. In that direction, Nag et al [14] and He et al [10] who applied the HMM techniques to vision, Yamato et al [25] who applied the HHM to whole human action recognition and Starner et al [21] who applied the HMM to hand gesture recognition, were among the first in their disciplines.

The main reason for this wide application of the model is its suitability to quantize a real world configuration space into a finite number of discrete states (see Fig. 2). In general, HMM needs an index of the current state of the system [16]. Additionally, the state changes, which approximate the dynamics of the system, should be described in a separate table with transitional probabilities represented as matrices. Moreover, this representation should fit the Markov condition which states that any information about the history of the process needed for future inferences, must be reflected in the current state [16].

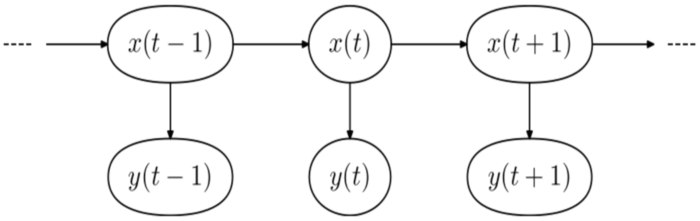


Fig. 2. Hidden Markov Model standard framework

HMM has the following advantages: a) the ability to efficiently model any kind of data that contains spatial-temporal relations, b) having clear Bayesian semantics, c) its three main basic problems (evaluation, decoding, training), can be addressed by the well-known algorithms such as the forward-backward algorithm [16], the Viterbi algorithm [23], the Baum-Welch algorithm [16], respectively, d) it can easily be extended to spatial classifiers such as support vector machines (SVM) and Gaussian mixture models (GMM), e) the straightforward state space and output space relation (i.e. any changes in the hidden states, cause systematic changes in the output distribution), g) it is applicable to many optimization algorithms such as the annealed particle filter (APF) and genetic algorithm (GA), f) it is extensible to other classes of classification and learning such as the artificial neural network (ANN), g) it has high potential to apply the structural changes to get a better result or to adopt to different application domains and/or scenarios.

However, HMM has many limitations too. Therefore modifications need to be done in order to efficiently parametrize the gestures. In the following subsections, some of these limitations and the proposed solutions are stated.

2.1 GMM-HMM: Continuous State Space

One important issue when considering HMM is the marginal state issue, and that is when the real observation happens outside of the defined state (as a result of the HMM-discretizing of real world configuration).

This problem can be addressed by acquiring Gaussian mixture models (GMM) within a HMM framework. Such a configuration introduces continuity to the observation space and enhances the HMM state space (and also enlarges it) due to the growing knowledge of the system. [11]

2.2 Linked-HMM and Coupled-HMM: Interactive Gestures

The use of HMM is limited to a simple state space with "one" discrete hidden variable at a time [4, 22] (e.g. a small number of states with limited state memories). This means representing two or more independent hidden variables (e.g. two interaction hands) with standard HMM will result in a large and complex state space (e.g. as big as the product of the sizes of each state space individually). This exponential growth in state (search) space affects the computational complexity in training, model likelihood estimation in inference and robustness of initialization to a great extent.

The idea here is to have two or more parallel hidden states which are internally linked to each other, and which correspond to their own observations and outputs. In that context, linked-HMM and coupled-HMM are introduced [22] as other extensions to HMM, which are tailored to represent the interaction of several independent processes (e.g. hands).

To compare these techniques it is worthy to indicate that, a standard HMM is quite sensitive to the initial values of the parameters. A linked-HMM (LHMM) is generally more robust, depending on the structure of the gestures. A coupled-HMM (CHMM) is least sensitive to the initial conditions and produces the highest likelihood [4]. Note that a LHMM is a simplification of a CHMM with symmetric, non-causal joint probabilities between chains.

2.3 Semi-HMM: Continuous Sequence of Gestures

An HMM is a kind of first-order Markov chain with the assumption that the transition to the next states at time $t + 1$, depends only on the state at time t. This implies that the probability of an observation for a certain interval of time declines exponentially with the length of the interval of the sequence [17]. This is especially an issue with visual events (contrary to speech events), including hand gestures, where the direction of a sub-event in the same event may vary from being very short to very long (e.g. the time a person shakes their friend's hand to say goodbye before waving).

The semi hidden Markov model (SMM) was proposed by [17] to address this issue. A SMM in general obtains the partial solutions independently and then recombining them in an efficient way. In other words, in a standard HMM, there is a state transition for every input symbol, whereas a semi Markov process, remains in a certain state for a number of time steps before transitioning into a new state is allowed.

This feature is very important, because during this segment of time, the system behavior is allowed to be even non Markovian, which enhances the model to be of variable duration (longer gestures) or to be nonlinear (sequences of continuous gestures instead of isolated atomic gestures).

To that extend, [17] introduced three types of features which are encoded in SMM variables. These features are: 1. Relation to the boundary of each segment, 2. Content characteristics about segments and 3. Interactions between neighboring segments. Moreover the paper suggests a combination of SMM with the support vector machine (SVM), to enhance the performance of the system and to test two-hand interactions.

2.4 CHnMM: Gestures with Different Speed

Another main issue with HMM is that, it is an over simplified version of real physical systems. This is because first, an HMM does not incorporate explicit timing or duration information (the transition from one state to another cannot be expressed as a duration like two or three seconds), because an HMM implicitly models the state's duration with geometric distributions [5] (e.g. "space"-time relation). Second, the information on when a symbol has been created is not considered in an HMM, (periodic symbols cannot be modeled correctly). Third, only the Markovian systems can be modeled exactly, although these systems rarely occur in practice.

As an elaboration, this is particularly a problem when we want to model similar gestures with different speeds. In this case another extension of HMM, called Conversive Hidden non-Markovian Model (CHnMM) [5], should be used where the definition of an observation is extended to also contain the time of the symbol emission. To that end, a $tuple[Symbol, t]$ is defined, where t is the point on time-axis when the symbol is observed.

2.5 VLMM: Higher Order Temporal Dependencies

Another limitation in a standard HMM is its difficulty of encoding higher order temporal dependencies. A variable length Markov model (VLMM) is a mathematical framework which can be used for modeling interactive behaviors based on their ability to capture behavioral dependencies at variable temporal scales [8]. In their approach VLMM has been applied to recognize whole human "body" actions (e.g. not hand). First, a posture template selection algorithm is developed based on a modified shape context technique. Next, the selected posture templates constitute a codebook, which is used to convert the input posture sequences into discrete symbol sequences for subsequent processing. Finally, the VLMM technique is applied to learn the symbol sequences that correspond to atomic actions.

2.6 IOHMM: Frame Rate Independence

An extension of HMM is input-output-HMM (IOHMM) [12] which is based on a non-homogenous Markov chain emission and its transition probabilities depend on Inputs. In the contrary, the HMM is based on a homogenous Markov chain since the dynamics of the system are determined only by the transition probabilities, which are time dependent.

Compare to an HMM which tries to "best" model the observations of a given gesture class, the IOHMM learns to map the input sequences (the observations), to the output sequences (the gesture class), for all observations of all gesture classes, using a supervised discriminant learning (function of input) [12]. Therefore, the model is capable of revealing and encoding the relevant information of the data when it is trained or to be inferred by a system having camera(s) with frame rates.

To end the section, we give quick review on different phases and their techniques. Various techniques and their potential enhancements such as the annealed particle filter (APF) [3], the Kalman filter [24], and the Monte Carlo Markov chain (MCMC) [2] can be considered in the inference phase (phase 5). Moreover, the principle components analysis, PCA [20], (e.g. as linear dimensionality reduction) or the Gaussian process latent variable model, GPLVM [6], (e.g. as non- linear dimensionality reduction) can be a suitable candidate for the dimensionality reduction phase (phase 4). Furthermore, many feature extraction (image processing) techniques can be utilized to extract more informative features from 2D input images/videos to address the issues of the second phase.

3 Methodology

In this section, two methodologies that are currently under design and implementation are briefly discussed.

3.1 Methodology A: Natural Language-Based Pose-Gesture Space Relationship

One of the main limitation of the HMM is its imperfect modeling of multiple isolated gestures in a meaningful relation to each other. In other words, the standard approach for recognizing a set of gestures is to train one HMM for each gesture. However, this is not computationally efficient since, as the number of gestures to be recognized increases, the number of the models and therefore the complexity (e.g. state-space complexity and time complexity) will increase. This is due to the fact that the relation between the number of gestures and the space-time complexity is linear. To reduce this complexity (which causes great technical challenges for a big gesture set), we propose a framework which is inspired by natural language processing [19].

This framework aims to model the gesture-posture space to grammatically relate the postures (e.g. static hand pose without any movement involved [9]) and the gestures (e.g. dynamic movement of a sequence of hand postures over a short time span which are connected by continuous motion patterns [9]) to each other (see Fig. 3). More specifically, first, we consider a dictionary of hand postures and a set of gestures. Then, by defining each posture as a word and each gesture as a sentence, we train a comprehensive grammar which reveals a meaningful relation between them.

Grammar of a language, models repetition of a word at different positions of various sentences. Similarly, by training and acquiring a grammar for the postures and gestures, we introduce a comprehensive model for repetition of one posture at different time-step of a gesture and in various gestures. This model reduces the number of duplicated states (e.g. duplicated postures) drastically (see Fig. 3).

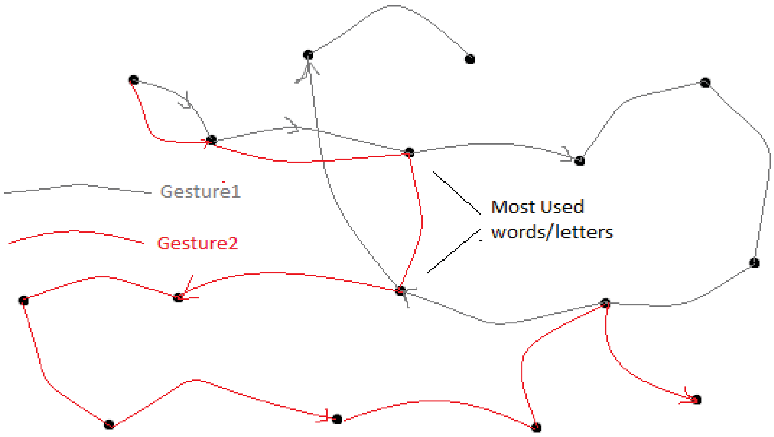


Fig. 3. Methodology A: Pose gesture space relation

Therefore, within this framework, it is expected that many different gestures can be modeled more compactly (e.g. we will have a set of words which can be combined to each other in different ways). Moreover, it must be extensible to different scenarios, and different number of hands in a more efficient way. Furthermore, the spotting time which is very much dependent on the camera frame rate, different gesture speeds, personalities and situations, must be handled more accurately (since it can deviate more from Markovian process and converge to a non-Markovian sequence). Additionally, geometrical corrections between poses and similar gestures, using the geometrical correction approach proposed by Zhang [26], would be achieved more systematically.

3.2 Methodology B: Non-Uniform State Transition Topology

Another limitations of an HMM is the lack of systematic way to determine the topology of the transitional matrix which defines the correct transitions between the states. Inspired by [18], where the uniform transitions are drawn in the form of rectangular and hexagonal networks (topology) (see Fig. 4.a and 4.b), this methodology concerns on defining of a non-uniform network (see Fig. 4.c).

To that end, we consider each pose in a gesture as a node in a graph, and the transition between two postures is modeled by an arc (edge), therefore, we will have a planar graph (the length of edges changes according the value of the transitional probability between states).

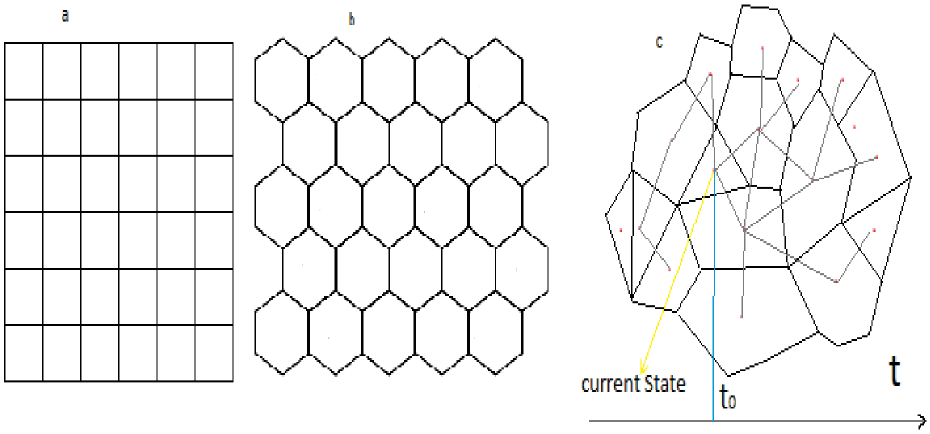


Fig. 4. a), b) uniform state transition, c) non-uniform state transition topology

Using this approach we will have a greater number of reachable previous and next states at any current point comparing to the uniform transitional topology in a standard HMM. Furthermore, the number of previous and next states for each node is not necessarily constant anymore (according to Euler rule for planar graph each vertex on average can have six surrounding triangles or poses in this terminology), which allows more flexible behavioral modeling.

4 Conclusion

In this paper, an introduction to hand gesture recognition systems was described. The standard HMM technique, its advantages and some of its disadvantages, as well as different solutions researchers have proposed to overcome these weaknesses has been investigated. Two main methodologies that are currently under investigation and implementation, are discussed in brief. With some hand gesture data, both from MOCAP systems and RGB (or RGB-D) cameras, the success of these two methodologies can be evaluated quantitatively.

References

1. Roland Aigner, Daniel Wigdor, Hrvoje Benko, Michael Haller, David Lindbauer, Alexandra Ion, Shengdong Zhao, and Jeffrey Tzu Kwan Valino Koh. Understanding mid-air hand gestures: A study of human preferences in usage of gesture types for hci. Technical Report MSR-TR-2012-111, November 2012.
2. Christophe Andrieu. An introduction to mcmc for machine learning, 2003.
3. M. Sanjeev Arulampalam, Simon Maskell, and Neil Gordon. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, 50:174–188, 2002.
4. M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, CVPR '97, pages 994–, Washington, DC, USA, 1997. IEEE Computer Society.
5. Tim Dittmar, Claudia Krull, and Graham Horton. A new approach for touch gesture recognition: Conversive hidden non-markovian models. *Journal of Computational Science*, 10:66 – 76, 2015.
6. Carl Henrik Ek, Philip H. S. Torr, and Neil D. Lawrence. Gaussian process latent variable models for human pose estimation. In *Machine Learning for Multimodal Interaction , 4th International Workshop, MLMI 2007, Brno, Czech Republic, June 28-30, 2007, Revised Selected Papers*, pages 132–143, 2007.
7. Yikai Fang, Kongqiao Wang, Jian Cheng, and Hanqing Lu. A real-time hand gesture recognition method. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 995–998, July 2007.
8. Aphrodite Galata, Anthony Cohn, Derek Magee, and David Hogg. Modeling interaction using learnt qualitative spatio-temporal relations and variable length markov models. In *In Proceedings of the European Conference on Artificial Intelligence*, pages 741–745, 2002.
9. Pragati Garg, Naveen Aggarwal, and Sanjeev Sofat. Vision based hand gesture recognition. *World Academy of Science, Engineering and Technology*, pages 972–977, 2009.
10. Y. He and A. Kundu. 2-d shape classification using hidden markov model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(11):1172–1184, 1991.
11. M.A.T. Ho, Y. Yamada, and Y. Umetani. An adaptive visual attentive tracker for human communicational behaviors using hmm-based td learning with new state distinction capability. *Robotics, IEEE Transactions on*, 21(3):497–504, June 2005.
12. Agnès Just and Sébastien Marcel. A comparative study of two state-of-the-art sequence processing techniques for hand gesture recognition. *Comput. Vis. Image Underst.*, 113(4):532–543, April 2009.
13. John McCormick, Kim Vincs, Saeid Nahavandi, Douglas C. Creighton, and Steph Hutchison. Teaching a digital performing agent: Artificial neural network and hidden markov model for recognising and performing dance movement. In *International Workshop on Movement and Computing, MOCO '14, Paris, France, June 16-17, 2014*, page 70, 2014.
14. R. Nag, K. Wong, and F. Fallside. Script recognition using hidden markov models. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.*, volume 11, pages 2071–2074, Apr 1986.
15. Chang-Beom Park and Seong-Whan Lee. Real-time 3d pointing gesture recognition for mobile robots with cascade hmm and particle filter. *Image Vision Comput.*, 29(1):51–63, January 2011.

16. Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
17. Gunnar Ratsch and Soren Sonnenburg. Large scale hidden Semi-Markov SVMs. In *Neural Information Processing Systems*, 2006.
18. Sam Roweis. Constrained hidden markov models. In *In Solla et al. (2000*, pages 782–788. MIT Press, 1999.
19. Feng sheng Chen, Chih ming Fu, and Chung lin Huang. y huang, c.: Hand gesture recognition using a real-time tracking method and hidden markov models. *Image and Video Computing*, pages 745–758, 2003.
20. Lindsay I. Smith. A Tutorial on Principal Component Analysis.
21. Thad E. Starner and Alex Pentland. Visual recognition of american sign language using hidden markov models, 1995.
22. Heung-Il Suk, Bong-Kee Sin, and Seong-Whan Lee. Hand gesture recognition based on dynamic bayesian network framework. *Pattern Recogn.*, 43(9):3059–3072, September 2010.
23. A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theor.*, 13(2):260–269, September 2006.
24. Greg Welch and Gary Bishop. An introduction to the kalman filter. Technical report, Chapel Hill, NC, USA, 1995.
25. J. Yamato, Jun Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference on*, pages 379–385, Jun 1992.
26. Liang Zhang and Guido Brunnett. Combining inverse blending and jacobian-based inverse kinematics to improve accuracy in human motion generation. *Computer Animation and Virtual Worlds*, pages n/a–n/a, 2014.

Part III

Image Usage: Perception, Cognition, Interaction, and Annotation

Combined use of eye-tracking and EEG to understand visual information processing

Nina Flad[1,2], Heinrich H. Buelthoff[1,3], and Lewis L. Chuang[1]

[1] Department of Perception, Cognition and Action,
Max Planck Institute for Biological Cybernetics, Tuebingen
lewis.chuang@tuebingen.mpg.de
[2] IMPRS for Cognitive and Systems Neuroscience
University of Tuebingen
[3] Department of Cognitive and Brain Engineering, Korea University

Abstract. Eye-tracking and electroencephalography (EEG) are popular methods to respectively evaluate visual information sampling and processing behavior in humans. It has been shown that the properties of visual stimuli and their mode of presentation can influence sampling behavior. However, it is less clear how information is processed after it has been sampled during natural eye-movement behavior. This is because EEG and eye-tracking tend to be performed separately given that eye- movements cause artifacts in the EEG. This paper provides an overview on recent developments that allow EEG recordings to be performed even in the presence of eye-movements. Modern algorithms can remove ocular artifacts in the EEG, enabling the use of EEG data recorded during naturalistic viewing conditions. In combination with EEG, electrooculography (EOG) recordings can be a viable alternative to conventional eye-trackers for measuring fixations, because EOG is recorded together with EEG and is therefore already synchronized.

Keywords: EEG, EOG, eye-tracker, visual sampling, information processing

1 Introduction

Many eye-movement studies have investigated where people are looking in a visual scene, given their motivations. The most famous study was performed by Yarbus, who recorded where participants looked in a famous painting, 'TheUnexpected Visitor', when required to answer different questions [1]. He noticed that the question influenced the pattern of fixations on the painting. For example,when the participants were instructed to determine the wealth of the family inthe picture, they looked more frequently at the objects in the room, such aspictures and furniture. In comparison, when he asked for the ages of the peoplein the painting, participants fixated on the faces in the painting instead. While itcan be inferred, eye-movement recordings do not directly allow the researcher to judge the relative importance of each fixated object with regards to the questionasked. For example, were the pictures on the wall more important than the dining table in determining affluence? Were the two objects equally processed by the brain?

Comparably few studies have analyzed eye-movements in conjunction with cortical activity. Therefore, not much is known about cortical processes that are elicited in conditions that allow for natural eye-movement behavior. In recent years, modern technologies have enabled researchers to study cortical processes together with the eye-movements that they accompany. This presents the researchers with the opportunity to gain more insight into the way we seek outand process visual information.

This paper aims to give an overview about recent developments on visual information processing research in the presence of eye-movements. We will particularly focus on research that addresses how visual sampling behavior and cortical processing of acquired information are influenced by physical properties of the visual scene. Henceforth, we will use information processing as an umbrella term for target detection, object recognition, stimuli discrimination andother similar processes that are supported by the human visual system.

First, we will give examples of stimulus properties that can influence an observer's behavior and - possibly - subsequent processing. Eye-movements (i.e. overt attention shifts) can be observed by eye-tracker recordings. Nonetheless, eye-tracking data does not directly measure the extent to which fixated stimuliare processed. It is well-established that attention can be covertly shifted to anobject other than the one that is currently fixated [2]. Second, we will presenthow EEG and event-related potentials (ERP) could allow visual informationprocessing in the cortex to be measured. Third, we will discuss technical issues that arise when EEG is recorded in the presence of voluntary eye-movements, which are inevitable during natural viewing behavior. Fourth, we review recentsolutions that could allow EEG/ERP to be unaffected by eye-movement inducedartifacts. Finally, we propose an alternative method to record eye-movements inthe presence of EEG instead of a conventional video-based eye-tracker.

2 Stimuli-driven eye-movements

We move our eyes to fixate and access information from different locations of the visual scene. Eye-movements direct the region on the retina with the highest visual acuity (i.e. fovea) towards the object or region in the visual field that holds interest for the observer. Thus, human observers can be expected to constantly move their eyes to sample different regions in a dynamic changing visual scene, in order to access task-relevant information. Each region (e.g., cockpit instrument) can be considered as a channel of visual information with definable physical properties (e.g., updating frequency). This as well as the physical properties of the visual stimuli themselves (e.g., contrast luminance) are known to influenceeye-movement (or information sampling) behavior.

2.1 Information-driven eye-movements

In a seminal study, Senders [3] demonstrated that trained observers moved their eyes across different instruments at a rate that corresponded to the instruments' update frequencies. More interesting, the exhibited sampling rate was nearly optimal as defined by information-theoretic principles [4]. This means, trained observers fixated visual instruments at a rate that approximated their Nyquist rates. The Nyquist rate describes the lower bound of sampling rates that allow accurate reconstruction of the signal after sampling. Thus, recorded eye- movements showed that trained observers executed roughly the right number of fixations that was necessary to extract the available information, but no additional, irrelevant ones.

The sampling strategies employed by the observer can also vary in accordance to the utility of the information presented. An ability to fluidly switch between strategies in order to maximize information sampling utility could be an indication of observer expertise. For example, Bellenkes et al.[5] demonstrated that expert pilots were able to switch between different scanning patterns in accordance to non-stationary situational demands. They selectively sampled only those instruments that were crucial for performing the current flight task. In other words, they were efficient and did not fixate task-irrelevant instruments. In contrast, novice pilots who were yet to require this flexibility in eye-movement planning sampled all instruments with a uniform pattern. This increased the risk of missing critical events, which impaired their flight control performance.

2.2 Covert vs overt attention shifts

Overt attention shifts can be measured with eye-trackers. Conventional eye-trackers use cameras and image-recognition algorithms to determine the orientation of the eye relative to the head. Comparing the orientation with reference data from a calibration procedure, the object or area on a display or world environment that is fixated can be determined.

It is often assumed that fixated information is processed, since eye-movements are always preceded by covert attention shifts [6]. However, covert shifts can occur in the absence of eye-movements. It is possible to fixate a target, but attend to and process another target. Thus, it is possible that information is fixated but unattended, because the attention has covertly shifted to another target.

3 EEG applications

Cognitive processing and covert attention shifts can be observed via EEG recordings. EEG employs scalp electrodes to measure electrical activity, which can be partially attributed to brain activity. The recorded signals can be decomposed into different frequency bands. Activity in these bands at certain locations has been associated with different cognitive processes.

For example, frontal delta (1-4 Hz) activity reflects attention to internal processing [7]. Changes in the gamma (> 30 Hz) band at parietooccipital sites can be induced by changes in visual-spatial attention [8]. Frontal theta (4-8 Hz) and occipital alpha (8-12 Hz) oscillations have been shown to reflect information encoding and visual processing, respectively [9]. Similar to alpha, mu rhythms (8-13 Hz) are related to sensorimotor processing. Even though mu and alpha oscillations share the same frequency ranges, the two rhythms can be distinguished by their locations and associated cognitive processes. Mu can be measured mostly in frontoparietal areas, whereas alpha can be measured in posterior areas [10].

3.1 Event-related potentials

Event-related potentials, and especially the P300, are of special interest to visual information processing researchers. The P300 is a comparatively large, and hence easily detectable, positive potential that occurs approximately 350 ms after a target event has been detected. It has been shown that the P300 reflects attention allocation to the appearance of discrete events [11] and cognitive processes [12].

During visual scanning, the P300 can serve as a tool to study covert shifts of attention to different regions of interest. This is due to the following reasons. The P300's amplitude is inversely proportional to the probability of the stimulus occurrence [13] and increases with increasing stimulus intensity. In addition, increasing the physical intensity of the stimulus can lead to a reduced latency of its generated P300 [14]. This means, rare events and task-relevant events can be expected to elicit an earlier and larger potential than common andirrelevant events. In addition, the same difference occurs between detected and missed events, which could be respectively treated as those that the participant paid covert attention to and otherwise. However, EEG is not commonly employed during visual scanning for the technical reasons that follow.

4 Ocular artifacts in EEG

The presence of eye-movements can present a problem for EEG measurements. This is because there is an electrical potential between the positively charged cornea and negatively charged retina in the eye, termed the corneo-retinal dipole, that contributes to the sum of measured activity in EEG electrodes, especially in frontal electrodes that are near the eyes [for a review, see 15].

Strong artifacts can be caused by saccadic eye-movements and blinks, which can occlude comparatively weak cortical activity that reflects information processing. More specifically, eye-movements re-orient the corneo-retinal dipole relative to the measurement electrodes, hereby distorting the measured signal. In addition, blinks short-circuit the cornea to the skin of the eyelids, which temporarily alter the strength of dipole potential.

Besides these, smaller artifacts have also been reported, which are related to eye-movement planning and not information processing per se [16]. They can distort measured EEG signals around the onsets of saccades.

4.1 Methods to avoid ocular artifacts

Given the above, EEG studies are typically designed to minimize eye-movements. A common method in EEG studies is rapid serial visual presentation. In an example, the words of a sentence are presented one after the other, at the point where the participant is fixating, to simulate how words are sequentially processed during reading. This paradigm allows researchers to employ EEG to investigate the semantic processing of words in a sentence without requiring voluntary eye-movements [17]. Alternatively, participants could be instructed to inhibit eye-movements by shifting their attention to peripheral stimuli for theirdetection, whilst fixating a central cross [18].

Whilst effective in generating EEG data that are uncorrupted by eye-movement artifacts, both methods constrain natural information sampling behavior. More recently, methods have been introduced (see Section 5) to remove eye-movement induced artifacts from the EEG data without distorting the EEG data itself.

5 EEG artifact-correction methods

Traditionally, EEG data segments that demonstrated artifacts were simply removed from further analysis. This can result in a considerable loss of data. This has, in turn, motivated the development of signal processing algorithms that allow artifacts to be isolated and separated from the EEG signal, leaving unadulterated cortically-induced signals.

5.1 Correction with subtraction

The most straightforward method is to directly subtract eye-movement related artifacts from the EEG signal. This approach assumes that the artifacts are only influenced by the direction and traveled distance of the eye-movement. In their study, Marton et al. [19] had participants perform equal number of saccades towards visual targets, in opposite directions, and averaged out the antagonistic eye-movement related artifacts. Dias et al. [20] collected multiple matching eye-movements for those that were accompanied by the cortical activity of interest and averaged over the signals to remove random noise. This noise-free 'mean saccade' was then subtracted from the saccades in the data to remove the effects of eye-movements.

Unfortunately, this subtraction method introduces a systematic error into the data, particularly when its main assumption does not hold. Eye-movement properties (e.g., latency) are influenced by the observer's purpose for generating the eye-movement [21]. This means that eye-movements that were performed in a given context, such as during calibration, might not generalize to match experimentally generated eye-movements.

5.2 Correction with a regression model

Another method is to create a regression model [e.g., 22]. First, a regression model is built, based on training data with controlled (i.e. predetermined by the experimenter) eye-movements. This model assumes that the measured EEG signal is a combination of the cortically-induced signals and overlying noise. Since noise is assumed to be mostly due to eye-movements, it is modeled as a mixture of the horizontal, vertical and radial dipole changes in the eyes. Changes in the corneo-retinal dipole can be measured by electrodes that are placed next to the eyes (i.e., EOG). Subsequently, the regression model tries to fit the measured dipole activity to the activity measured by the EEG electrodes. Once the mathematical model of the interaction between EEG and ocular dipole has been established, it can be inversed and the effects of the measured dipole shifts can be removed during the actual EEG recording.

The main assumption of the regression method is the independence of EEG and noise as well as the independence of the three spatial dimensions of the eye dipole, which is mostly correct. Schlögl et al. [22] was able to show that this method reduces eye-movement related artifacts by 80 percent. The maindrawback of this method is that the training data needs to be free of any corticalactivity to enable the model to perfectly fit the influence of the eyes on the EEG signal.

5.3 Correction with ICA

Many EEG studies increasingly utilize independent component analysis (ICA) for artifact correction. ICA can be used to decompose a multivariate signal (like the EEG recorded at multiple electrodes) into independent non-Gaussian subcomponents. Similar to the regression model, ICA assumes that artifactual components such as blinks, saccades and muscle noise are independent from sources of neural activity in the brain. Therefore, it is possible to use ICA to decompose the corrupted EEG into its contributing sources and to selectively remove undesired components from the signal [23].

The application of ICA does not require training data with special properties. Nonetheless, ICA works best with a EEG system with at least 64 electrodes. This is because ICA decomposes the data into as many underlying source components as there are electrodes. Thus, EEG that is recorded from only a few electrodes cannot reliably discriminate between sources of non-cortical and cortical activity.

ICA was employed in a recent study that compared the old/new effect of visual word recognition for conditions with and without eye-movements [24]. A video-based eye-tracker recorded eye-movements of the reading observer and treated the fixation onset of the target word as the onset event of an ERP. Ocular artifacts were removed using ICA. This study found comparable old/new effect regardless of eye-movement activity. In other words, ICA was able to remove eye- movement related artifacts, while retaining the EEG features of interest [25].

In contrast to previously mentioned methods, ICA can be utilized to remove all types of artifacts (e.g. line noise, muscular activity, etc.) from the EEG data.

However, this requires experienced researchers to manually inspect the temporaland spectral properties of the derived components and to label them accordingly.

5.4 Other approaches

Other approaches include mixtures of the previously mentioned methods. For example, the surrogate multiple source eye correction (MSEC) [26, 27] combines several aspects of the previously mentioned methods. It treats controlled eye-movements from each individual participant as a reference signal. It also employs a type of component analysis to decompose the data and, in doing so, creates a dipole model. Dimigen and Sommer [28] used MSEC to correct for the artifacts in their reading study and found typical EEG features. However, it was shown that MSEC has problems with removing the artifacts completely since the cortical signals are only approximated by a brain model [27].

6 EOG-based eye-tracking

All of the above mentioned studies utilized an eye-tracker to record eye-movements. However, it is also possible to collect information about fixations by means of EOG recordings [29]. For EOG recordings, electrodes are placed around the eyes to measure changes in the electrical dipole created by each eye. Since eye-movements result in a shift of these dipoles, this induces a measurable change in the EOG signal. It is possible to validate fixations on a target via the EOG signals, for example, by means of a regression model [29]. From a technical perspective, both eye-tracking with a conventional eye-tracker and with EOG come with their own advantages and disadvantages.

EOG can be recorded alongside the EEG with the same amplifier. This means that it is synchronized with the EEG and has a high sampling frequency (EEG is usually recorded at 250 to 1000 Hz), which can be changed easily depending on the researcher’s requirements. This gives EOG a better temporal resolution than many eye-trackers, which commonly sample at 60 Hz. Higher temporal resolution allows for more accurate detection of the on- and offsets of fixations. For this reason, EOG is often employed to monitor the fixations ofparticipants that are not allowed to move their eyes during an EEG study (e.g. [18, 30]).

EOG has been shown to achieve a spatial resolution of 2° [31]. To achieve this resolution, a linear mapping between EOG amplitude and orientation of the eye is created during a calibration procedure which is similar to the calibration procedure of optical eye-trackers [32]. Unlike video-based eye-trackers, EOG does not only record signals that are due to eye-movements. It also picks up muscular activity, line noise, cortical activity and blinks. As such, the EOG is inherently noisy and needs to be processed to improve the signal to noise ratio, e.g. by filtering and component analyses, as it is already common with EEG data. Since EOG is not widely used at present, no software is available for easy application. Therefore, its full potential in the combination with EEG remains unclear to date.

7 Conclusion

Studying information sampling and processing together, as they occur in everyday situations, requires the combined use of two different techniques. First, eye-tracking can be used to record overt shifts of attention, i.e. changes in eye-movement behavior, to sample information in the visual scene. Second, EEG recordings can reflect covert shifts of attention, i.e. whether the fixated stimulusis actually cortically processed. The combination of these techniques calls for advanced methods to correct for the artifacts that will arise due to eye-movements. The recent development of such methods enables researchers to improve our understanding on information retrieval and processing.

In the future, EOG could replace optical eye-trackers, especially for EEG recordings. Until now, linear mappings are established between EOG amplitude and eye orientation. These mappings have to be calibrated elaborately and repeatedly. However, unsupervised machine learning algorithms could remove these issues and allow for the easy application of EOG. These algorithms provide easy and simple fixation estimation in exchange for lower spatial resolution. Unsupervised clustering on the EOG data reliably and automatically extracts the participants' regions-of-interest and classifies fixations. Further combination of clustering with modern signal processing algorithms could improve spatial resolution while keeping the speed of unsupervised calibration procedures.

8 Acknowledgements

The authors thank the German Research Foundation (DFG) for financial support within project C03 of SFB/Transregio 161.

References

[1] Yarbus, A.: Eye Movements During Perception of Complex Objects. In: Eye Movements and Vision, pp. 171–211, Springer US (1967).

[2] Carrasco, M., McElree, B.: Covert attention accelerates the rate of visual information processing. Proceedings of the National Academy of Sciences, vol. 98(9), pp. 5363–5367 (2001).

[3] Senders, J.: The Human Operator as a Monitor and Controller of Multidegree of Freedom Systems. IEEE Transactions on Human Factors in Electronics, vol. HFE-5(1), pp. 2–5 (1964).

[4] Shannon, C.: A Mathematical Theory of Communication. The Bell System Technical Journal, vol. 27(July 1928), pp. 379–423 (1948).

[5] Bellenkes, A., Wickens, C., Kramer, A.: Visual scanning and pilot expertise: the role of attentional flexibility and mental model development. Aviation, Space, and Environmental Medicine, vol. 68(7), pp. 569–579 (1997).

[6] Peterson, M.S., Kramer, A.F., Irwin, D.E.: Covert shifts of attention precede involuntary eye movements. Perception & Psychophysics, vol. 66(3), pp. 398–405 (2004).

[7] Harmony, T., Femgndez, T., Silva, J., Bemal, J., Diaz-comas, L., Reyes, A., Marosi, E., Rodriguez, M., Rodriguez, M.: EEG delta activity: an indica- tor of attention to internal processing during performance of mental tasks. International Journal of Psychophysiology, vol. 24(1), pp. 161–171 (1996).

[8] Gruber, T., Müller, M.M., Keil, A., Elbert, T.: Selective visual-spatial attention alters induced gamma band responses in the human EEG. Clinical Neurophysiology, vol. 110(12), pp. 2074–2085 (1999).

[9] Klimesch, W.: EEG alpha and theta oscillations reflect cognitive and mem- ory performance: a review and analysis. Brain Research Reviews, vol. 29(2), pp. 169–195 (1999).

[10] Pineda, J.A.: The functional significance of mu rhythms: translating seeing and hearing into doing. Brain Research Reviews, vol. 50(1), pp. 57–68 (2005).

[11] Polich, J., Kok, A.: Cognitive and biological determinants of P300: an integrative review. Biological Psychology, vol. 41(2), pp. 103–146 (1995).

[12] Kok, A.: Event-related-potential (ERP) reflections of mental resources: a review and synthesis. Biological Psychology, vol. 45(1), pp. 19–56 (1997).

[13] Duncan-Johnson, C.C., Donchin, E.: On quantifying surprise: The variationof event-related potentials with subjective probability. Psychophysiology, vol. 14(5), pp. 456–467 (1977).

[14] Polich, J., Ellerson, P.C., Cohen, J.: P300, stimulus intensity, modality, and probability. International Journal of Psychophysiology, vol. 23(1), pp. 55–62 (1996).

[15] Plöchl, M., Ossando´n, J.P., König, P.: Combining EEG and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. Frontiers in Human Neuroscience, vol. 6, pp. 278 (2012).

[16] Jagla, F., Jergelov´a, M., Riecansky´, I.: Saccadic eye movement related potentials. Physiological Research, vol. 56(6), pp. 707–713 (2007).

[17] Barber, H.A., Ben-Zvi, S., Bentin, S., Kutas, M.: Parafoveal perception during sentence reading? An ERP paradigm using rapid serial visual presentation (RSVP) with flankers. Psychophysiology, vol. 48(4), pp. 523–31 (2011).

[18] Thut, G., Nietzel, A., Brandt, S.a., Pascual-Leone, A.: Alpha-band electroencephalographic activity over occipital cortex indexes visuospatial attention bias and predicts visual target detection. The Journal of Neuro- science, vol. 26(37), pp. 9494–502 (2006).

[19] Marton, M., Szirtes, J., Donauer, N., Breuer, P.: Saccade-related brain potentials in semantic categorization tasks. Biological Psychology, vol. 20(3), pp. 163–184 (1985).

[20] Dias, J., Sajda, P., Dmochowski, J., Parra, L.: EEG precursors of detectedand missed targets during free-viewing search. Journal of Vision, vol. 13, pp. 1–19 (2013).

[21] Bieg, H.J., Bresciani, J.P., Bülthoff, H.H., Chuang, L.L.: Looking for discriminating is different from looking for looking's sake. PLoS ONE, vol. 7(9), pp. e45445 (2012).

[22] Schlögl, A., Keinrath, C., Zimmermann, D., Scherer, R., Leeb, R., Pfurtscheller, G.: A fully automated correction method of EOG artifacts in EEG recordings. Clinical Neurophysiology, vol. 118(1), pp. 98–104 (2007).

[23] Jung, T.P., Makeig, S., Humphries, C., Lee, T.W., McKeown, M.J., Iragui,V., Sejnowski, T.J.: Removing electroencephalographic artifacts by blind source separation. Psychophysiology, vol. 37(2), pp. 163–78 (2000).

[24] Hutzler, F., Braun, M., V˜o, M.L.H., Engl, V., Hofmann, M., Dambacher, M., Leder, H., Jacobs, A.M.: Welcome to the real world: validating fixation- related brain potentials for ecologically valid settings. Brain Research, vol. 1172, pp. 124–129 (2007).

[25] Iriarte, J., Urrestarazu, E., Valencia, M., Alegre, M., Malanda, A., Viteri,C., Artieda, J.: Independent component analysis as a tool to eliminate artifacts in EEG: a quantitative study. Journal of Clinical Neurophysiology, vol. 20(4), pp. 249–57 (2003).

[26] Berg, P., Scherg, M.: A multiple source approach to the correction of eye artifacts. Electroencephalography and Clinical Neurophysiology, vol. 90(3), pp. 229–41 (1994).

[27] Ille, N., Berg, P., Scherg, M.: Artifact correction of the ongoing EEG using spatial filters based on artifact and brain signal topographies. Journal of Clinical Neurophysiology, vol. 19(2), pp. 113–24 (2002).

[28] Dimigen, O., Sommer, W.: Coregistration of eye movements and EEG in natural reading: analyses and review. Journal of Experimental Psychology: General, vol. 140(4), pp. 552 (2011).

[29] Kelly, S., Lalor, E.: Visual spatial attention tracking using high-density SSVEP data for independent brain-computer communication. IEEE Trans- actions on Neural Systems and Rehabilitation Engineering, vol. 13(2), pp. 172–178 (2005).

[30] Ding, J., Sperling, G., Srinivasan, R.: Attentional modulation of SSVEP power depends on the network tagged by the flicker frequency. Cerebral Cortex, vol. 16(7), pp. 1016–1029 (2006).

[31] Stern, R.M., Ray, W.J., Quigley, K.S.: Psychophysiological recording. Ox- ford University Press (2001).

[32] Finocchio, D.V., Preston, K.L., Fuchs, A.F.: Obtaining a quantitative mea-sure of eye movements in human infants: A method of calibrating the electrooculogram. Vision Research, vol. 30(8), pp. 1119–1128 (1990).

The Digital Health Companion: Personalized Health Support on Smartwatches via Recognition of Activity- and Vital-Data

John Trimpop[1], Marian Haescher[1], Gerald Bieber[1], Denys J.C. Matthies[1],
Friedrich Lämmel[1], and Paul Burggraf[1]

[1] Fraunhofer IGD, Joachim-Jungius-Strasse 11, 18059 Rostock, Germany
{john.trimpop;marian.haescher;gerald.bieber;denys.matthies;
friedrich.laemmel;paul.burggraf}@igd-r.fraunhofer.de

Abstract. It has been shown that in various fields of social life, people tend to seek opportunities to measure their daily activities, bodily behaviors, and health related parameters. These kinds of activity tracking should be accomplished comfortably, unobtrusively and implicitly. Tracking behavior can be important for certain user groups, such as the growing population of elderlies. These people have a substantially higher risk of falling down, as they often live alone and thus have a greater need for other supporting services, as emergencies quickly occur. We would like to support these people, while providing a comfortable emergency detection and a monitoring of physical activities. Moreover, we believe such tracking applications to be beneficial for any user group, since we can perceive the trend of quantified self: knowing about one's own body characteristics, which is expressed in body movement. Simultaneously, we also perceive that a strong desire for a comprehensive monitoring of vital and health data is emerging. In this paper we describe the concept and implementation of the Digital Health Companion, a smart health support system that combines research developments of activity, vital data, and anomaly recognition with the functionality of contemporary smartwatches. The system's health monitoring includes an emergency detection and allows for the prevention of health risks in the short and long term through the recognition of body movement patterns.

Keywords: Health Monitoring, Activity Recognition, Emergency Support, Smartwatch, Inertial Sensors, Quantified Self

1 Introduction

Latest customary smart devices are already capable to serve as a personal assistant while permanently collecting body parameters in an unobtrusive way. Moreover, a quantified-self-movement emerged during the last years; people desire to analyse and evaluate themselves, their movements, their sportive activities, and their sleeping behavior. To date, companies already provide extensive platforms to collect and save personal activity data [14].

While the process of collecting abstract activity data in everyday situations is continuously developing and improving, it is still a challenge to extract useful information in a format beneficial to end-users (e.g. doctors, patients, caretakers). Especially people in need of care such as elderlies, who live with a higher risk of requiring emergency support, could exceedingly profit from a system that is not just able to track activity data, but that is also able to interpret data and thus prevent emergencies and lower risks while automatically monitoring body functions and user activities.

In this paper we present a concept and a prototypical implementation of a smart health support system that is based on customary smartwatches, which we call the **Digital Health Companion (DHC)**. We combine activity, vital data, and anomaly recognition technologies with default functions of current smartwatches, such as location tracking, push messaging, or phone calls, to allow for the usage of a stigma-free automatic emergency assistant. We thus contribute improved activity recognition algorithms that can be implemented energy efficiently and user independently on customary smartwatches, while keeping the desired complexity and recognition accuracy. Furthermore, we address an important issue - permanent vital data extraction with smartwatches - that can be exploited as a feature and that allows for the detection of health risks. Moreover, we try to find new ways regarding interaction and usability with small screens that can be controlled by old and/or handicapped people. The DHC not only offers a smart health support to consumers, but also a solution for house emergency services and other aid organizations in order to efficiently support customers. Our prototypical implementation runs on customary smartwatch models (*see in Fig. 1*).

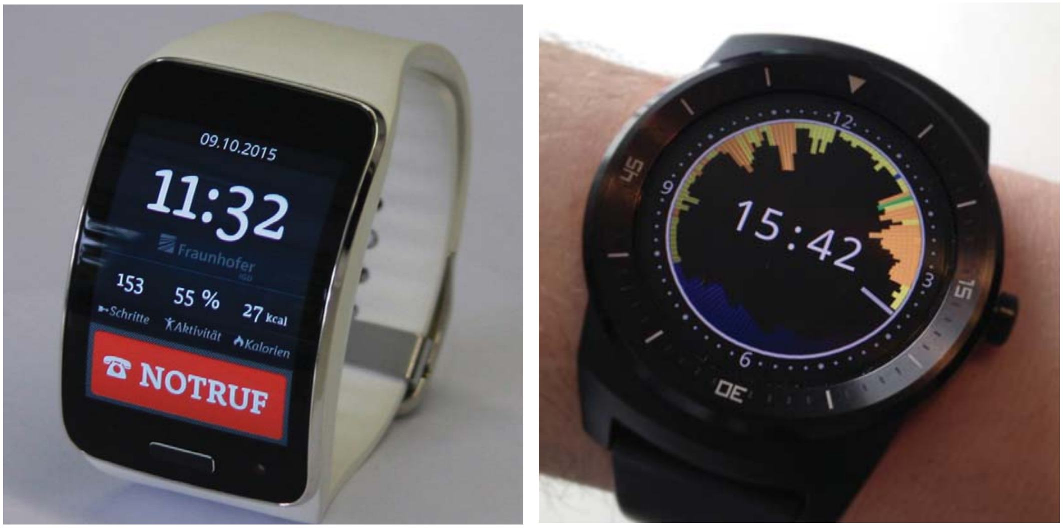


Fig. 1. Prototype of the DHC system implemented on a Samsung Gear S Smartwatch (left) and Activity watchface implemented on an Android Wear LG G Round (right)

2 Related Work

While physical activity recognition has become an attractive field of research over the past years [1, 6], different body positions and a variety of sensor types have been evaluated for many fields of applications, including mobile scenarios. To position our paper with respect to the mobile scenario, we here provide a brief overview of activity recognition approaches via wrist-mounted sensors and smartwatches as we present general smart health support systems.

2.1 Activity Recognition with Wrist-Mounted Sensors

While the hip can be seen as the classic body position for activity recognition with wearables, lately the wrist has increasingly been used in research works as well. Scientists developed watch-based prototypes while making use of built-in sensors (primarily accelerometers, or directly attached sensor units at the wrist) that are able to store or stream movement or activity data. Those prototypes are demonstrated to be used for activity recognition tasks or similar approaches as a single sensor setup or being embedded into a sensor networks – see also Bao and Intille [2] or Maurer et al. [17]. Surveys of different activity recognition systems are presented by Avci et al. [1] or Lara and Labrador [15].

Nowadays, especially smartwatches gain much more popularity since new consumer devices are being developed, which allow for a broad range of different applications, such as activity recognition, fall detection, sleep detection, or applications in industrial environments [4]. Besides smart bands, smartwatches can also provide a tracking of personal activity, which usually synchronize their data with another third party device or which directly broadcast the gathered data on the internet.

2.2 Smart Health Support Systems

The research area for smart health support is broad and diversified, due to many applications available for inter alia handicapped people or elderly care. In the scope of *Ambient Assisted Living* (AAL), many ideas, concepts, and also implementations have been presented [9, 10].

Various wearable health support systems can already provide an emergency detection, such as solutions for intelligent and automatic fall detection systems - Chen et al. [8] or Salomon et al. [20]. A review of different fall detection approaches has been published by Mubashir et al. [18]. Furthermore, Bieber et al. [3] describe a concept for an activity recognition-based anomaly detection system with smartwatches, which is intended to work for elderly users and their family members. Lutze and Waldhör [16] depict the possibility of a smartwatch-based house emergency service system and highlight the capability of current smartwatch models. Following their statements, smartwatches already incorporate all necessary functional units, such as communication services (e.g. microphone, 3G, GPS, WiFi) and enough relevant sensors (e.g. accelerometer, gyroscope, altitude sensor, and also vital sensors such as a PPG). All authors agree on the high potential of smartwatches in general, outline interesting concepts, and cite application experiences, but also current challenges.

In conclusion, the authors clearly demonstrate that smartwatches are capable of relieving classic house emergency services, while being unobtrusive, not stigmatizing and while providing a great variety of new functions [16].

Still, due to current hardware constrains, a reliable health support system for elderlies based on smartwatches is not yet available. In order to circumvent these issues and to still create a smartwatch-based health support system, we developed new improved activity recognition and anomaly detection algorithms with adaptive sampling rates.

3 Concept and Implementation

In this section we introduce the main concept of the DHC system and how the different technological parts are being implemented and how they interact with each other. Firstly, the single system components are being described, which mainly consist of the smartwatch (client side) and the server implementation. Subsequently, we outline the idea of our automatic emergency and long-term anomaly detection. To allow for a quick overview of the envisioned system architecture, the figure 2 illustrates the interaction between all system's components.

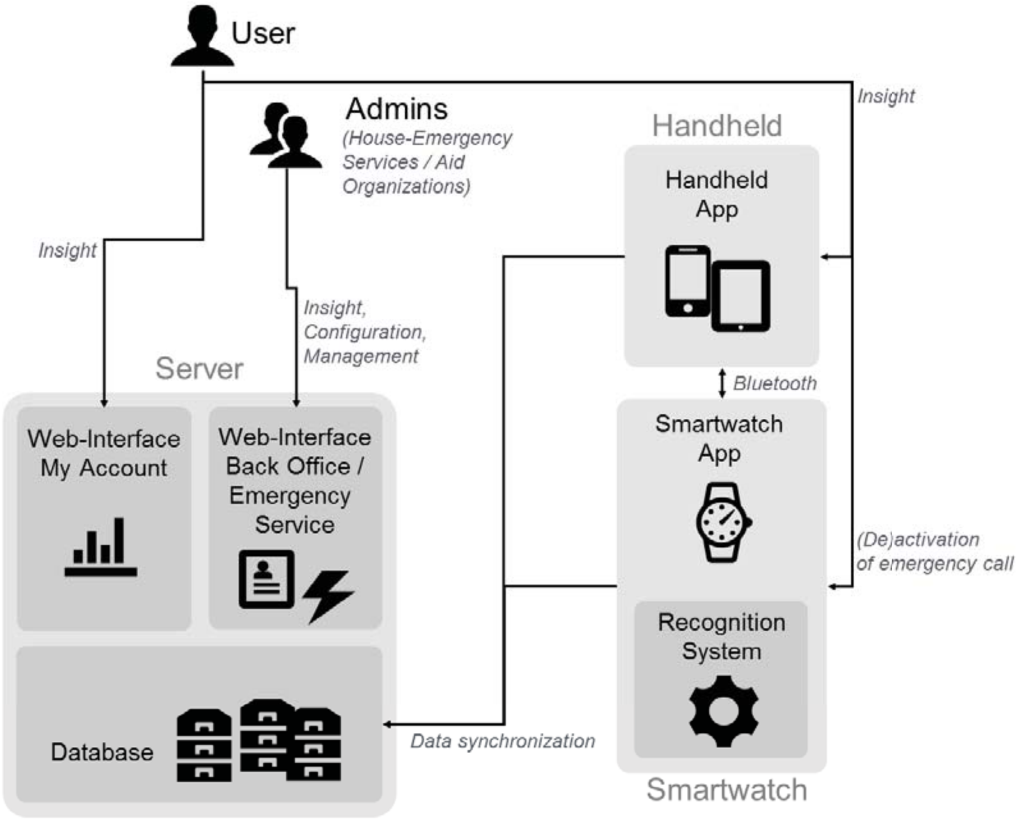

User
Admins
(House-Emergency Services / Aid Organizations)
Handheld
Handheld App
Insight
Insight
Insight, Configuration, Management
Server
Bluetooth
Web-Interface My Account
Web-Interface Back Office / Emergency Service
Smartwatch App
(De)activation of emergency call
Recognition System
Database
Data synchronization
Smartwatch

Fig. 2. Overview of the Digital Health Companion System

3.1 Smartwatch

The commercially available smartwatch can be considered as the key component of our system. All recognition technologies (*modules*) are implemented in a client app on the device, which also handles all incoming and outgoing connections and communication processes, respectively. All recognition algorithms were developed over the last years at Fraunhofer IGD Rostock and are now integrated into our system.

While there is a huge amount of customary smartwatches available, most watches dependent on a secondary hub device (e.g. *Android Wear* smartwatches or the *Apple Watch*) for communication services, but which are not as suitable for our autarkic solution. It has been shown that especially handicapped people have problems using an additional device at the same time. An autarkic smartwatch is usually independent from a secondary handheld (e.g. the *Simvalley AW-420.RX*, the *Samsung Gear S* or the upcoming *Samsung Gear S 2*) and can utilize all necessary communication possibilities, such as WiFi, GSM, and GPS without requiring a third party device to run.

In order to decrease energy consumption, we established algorithms that rely on adaptive sample rates, as well as on dedicated features and classifier models. We also implemented routines to disable and enable connection settings on demand. Instead of using new sensor units, such as the PPG heart rate sensor, we created a new approach to extract vital data from a conventional inertial sensor; the accelerometer [11, 12, 13]. This information can be obtained in resting situations of the user, for example during sleep periods. Furthermore, we created new visualization schemes, which for instance allow for an activity visualization as an integrated watchface to highlight activity and sleep patterns directly in the background. Moreover, our app provides a clear and intuitive user experience, which is especially designed for elderly or handicapped people.

All recognition modules are not visualized, but automatically running hidden in background. Moreover, we also developed single user screens with oversimplified one-button user input possibilities, to enable an easily accomplished emergency call that likewise can be cancelled comfortably (Fig. 1). In addition to that, we are planning to improve those user interfaces with the results of field studies involving many users. We believe this to provide a good basis for new design decisions to improve usability and user experience. Additionally, in the future, we also plan to perform field studies with sleep laboratories to evaluate our collected smartwatch vital data against vital data from validated laboratory devices.

The following table lists and categorizes all modules that are implemented in the app. Citations of the Fraunhofer technologies are also mentioned.

Table 1. Different modules of the smartwatch component of the DHC system

Recognition Modules	Communication Modules	Additional Modules
Accelerometer based activity recognition [6]	Manual emergency call	Text to speech
Sleep pattern recognition [4]	Push notifications	Energy efficient recognition routines [5]
Vital data recognition [12]	Remote phone calls	Big-Data analysis
Doffed detection [11]	Location services	
Microvibration recognition [11]		
Fall detection [20]		
Automatic emergency call		
Anomaly detection		

3.2 Server Implementation and Handheld

To provide a system with the desired functionality, a scalable and well-tested server side is being required. As seen in Fig. 2, the server consists of a database to store all relevant user data, which yields application interfaces to the smartwatch and handheld for upload and synchronizing purposes.

The server also incorporates two web interfaces. The *back office* serves as the main configuration, maintenance and insight service for administrators: the emergency services or aid organizations. The *back office* also has a connection to the emergency call service if needed. The second web interface, *My Account*, is intended to serve the user or his family members in order to provides insights in his personal user data and visualizations of activity patterns. This web interface is also accessible from a smartphone or tablet, while these devices of course can also server as a hub device for connection capabilities if required.

This server design allows specifically adapted options regarding user profile management and configuration, corresponding to the user groups' task profile. As the user's individual activity, vital, and health information constitute very sensitive data, best state of the art security standards are implemented on the server frontend and backend (e.g. OWASP).

3.3 Automatic Emergency and Anomaly Detection

The research domain of *Anomaly Detection (AD)* is widespread in multiple application domains. These domains include topics such as: credit card fraud detection, network intrusion detection, as well as the detection of health related anomalies. As a subcategory of pattern recognition, AD is closely related to data mining and machine learning approaches. The scarce occurrence of anomaly patterns in the data available is the biggest challenge researchers face in this area. As a result of this, initial anomaly patterns are hard to train or to learn. The anomaly detection, incorporated in the DHC system, is based on activity data, as well as vital parameters, which are gathered by a smartwatch, which allows for an easier, permanent and more extensive gathering of (anomaly) data. Due to the fact that sensor signals can differ in signal quality, an evaluation of the vital parameters captured is crucial in a medical or health context. Moreover, a possible loss of sensor data has to be taken into account. The AD concept of the DHC is based on a signal quality evaluation unit, that rates the quality of sensor inputs to weight the anomaly class detected based on the sensor input given. We envision detection scenarios such as: cardiac anomalies, falls, and unconsciousness, epileptic seizures, or sleep anomalies, such as sleep apnea.

4 Application and Distribution Potential

Fig. 3. Motorola Moto 360 smartwatch with DHC watchface

While the benefits of the DHC system as a whole are not limited to medical support or emergency services that supervise handicapped people, we also envision ways to directly commercialize it in the consumer market. In regard to this, we imagine DHC to have the potential to be used for several purposes such as:

- general health monitoring of body functions (for anyone at any age, who is interested)
- long term activity recognition and risk prevention (e.g. adults or active elderlies, who are interested in disease prevention and want to reveal risky physical and mental symptoms)
- special guardianship purposes (e.g. by basically healthy elderlies, who perform longer outside activities or young adults, who want to prevent severe injuries while conducting extreme sportive activities).

We already acquired potential customers and medical partners, with whom we are evaluating all ranges of functions of DHC and validate its capabilities. In contrast to contemporary activity tracking systems, we expect that our technologies have the ability to be efficiently used as a reliable support and medical tool by the wide public.

5 Conclusion and Future Work

Smartwatches are seen as the next big trend in relation to the development of mobile devices [16]. When using newly developed algorithms, they clearly offer the possibility to not only serve as rudimentary activity trackers, but also to reliable recognize activities and vital data to enable a trustworthy health data recognition systems. Interpreting and analysing this is the basis for offering proper health support and risk prevention based on smartwatches

Next to the big trend itself, a strong demand for new medical support technologies can be perceived. Europe experiences a demographic change; people simply live longer. Emergency service companies aim to prolong their average user subscription duration and lower the average user age by means of more comfortable and non-stigmatic solutions. Furthermore, researchers forecast the global market of mobile health services (which was at 6 billion US dollar in 2014) to rise to 26 billion US dollar in 2017 [19].

A prototype of the DHC system has already been implemented and will be further improved and rolled out as a product in the near future. In this respect, the core system is improved, existing functionality is validated for health care usage with renowned medical partners and additional features are added through profound research.

Building on that, the team dedicated to the deployment of the DHC system is about to launch a spin-off company; to ensure a continuous product improvement together with our initial partners and to realize the full commercialization potential of this next-generation health monitoring system

6 Acknowledgements

This work has been funded by the federal state of Mecklenburg-Vorpommern and EFRE within the project “Basic and Applied Research in Interactive Document Engineering and Maritime Graphics” and by the European Social Fund under grant ESF/IV-BM-B35-0006/12.

References

1. Avci, A., Bosch, S., Marin-Perianu, M., Marin-Perianu, R., & Havinga, P. (2010, February). Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. In *Architecture of computing systems (ARCS),* (pp. 1-10). VDE.
2. Bao, L., & Intille, S. S. (2004). Activity recognition from user-annotated acceleration data. In *Pervasive computing* (pp. 1-17). Springer Berlin Heidelberg.

3. Bieber, G., Fernholz, N., & Gaerber, M. (2013). Anomalienerkennung durch Analyse der körperlichen Aktivität. *Fachkongress Social Business*. Tagungsband, Rostock-Warnemünde.
4. Bieber, G., Haescher, M., and Vahl, M.(2013). Sensor requirements for activity recognition on smart watches. In *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments* (p. 67). ACM.
5. Bieber, G., Kirste, T., & Gaede, M. (2014). Low sampling rate for physical activity recognition. In *Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments* (p. 15). ACM.
6. Bieber, G., Voskamp, J., & Urban, B. (2009). Activity recognition for everyday life on mobile phones. In *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments* (pp. 289-296). Springer Berlin Heidelberg.
7. Bulling, A., Blanke, U., & Schiele, B. (2014). A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys* (CSUR), 46(3), 33.
8. Chen, J., Kwong, K., Chang, D., Luk, J., & Bajcsy, R. (2006, January). Wearable sensors for reliable fall detection. In *Engineering in Medicine and Biology* Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the (pp. 3551-3554). IEEE.
9. Costa, R., Carneiro, D., Novais, P., Lima, L., Machado, J., Marques, A., & Neves, J. (2009, January). Ambient assisted living. In 3rd Symposium of Ubiquitous Computing and Ambient Intelligence 2008 (pp. 86-94). Springer Berlin Heidelberg. ISO 690
10. Georgieff, P. (2008). Ambient Assisted Living. Marktpotenziale IT-unterstützter Pflege für ein selbstbestimmtes Altern, *FAZIT Forschungsbericht*, 17, 9-10.
11. Haescher, M., Bieber, G., Trimpop, J., Urban, B., Kirste, T., & Salomon, R. (2014). Recognition of Low Amplitude Body Vibrations via Inertial Sensors for Wearable Computing. In *Proc. of Conference on IoT Technologies for HealthCare* (HealthyIoT).
12. Haescher, M., Matthies, D. J.C., Trimpop, J., & Urban, B. (2015). A study on measuring heart- and respiration-rate via wrist-worn accelerometer-based seismocardiography (SCG) in comparison to commonly applied technologies. In *Proceedings of the 2nd international Workshop on Sensor-based Activity Recognition and Interaction* (p. 2). ACM.
13. Haescher, M., Matthies, D. J.C., & Urban, B. (2015). Anomaly Detection with Smartwatches as an Opportunity for Implicit Interaction. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (pp. 955-958). ACM.
14. Harvey, M. J., & Harvey, M. G. (2014). Privacy and security issues for mobile health platforms. *Journal of the Association for Information Science and Technology*, 65(7), 1305-1318.
15. Lara, O. D., & Labrador, M. A. (2013). A survey on human activity recognition using wearable sensors. *Communications Surveys & Tutorials*, IEEE, 15(3), 1192-1209.
16. Lutze, R., & Waldhör, K. (2015, January). SmartWatches als Hausnotrufsysteme der nächsten Generation. In *AAL-Kongress 2015*. VDE VERLAG GmbH.
17. Maurer, U., Rowe, A., Smailagic, A., & Siewiorek, D. P. (2006). eWatch: a wearable sensor and notification platform. In *Wearable and Implantable Body Sensor Networks*. BSN 2006. International Workshop on (pp. 113-116). IEEE.
18. Mubashir, M., Shao, L., & Seed, L. (2013). A survey on fall detection: Principles and approaches. *Neurocomputing*, 100, 144-152.
19. Research2Guidance (2014). mHealth App Developer Economics 2014. *mHealthEconomocs.com*, 05-06-2014.
20. Salomon, R., Luder, M., & Bieber, G. (2010). iFall-a new embedded system for the detection of unexpected falls. In *Pervasive Computing and Communications Workshops* (PERCOM Workshops), (pp. 286-291). IEEE.

A brief Survey on Understanding the Interaction between Human and Technology at the Task of Pedestrian Navigation

Anita Meier[1], Denys J.C. Matthies[2], Frank Heidmann[1]

[1] University of Applied Sciences Potsdam (FHP), Germany
mail@aboutiam.de, heidmann@fh-potsdam.de
[2] Fraunhofer IGD Rostock, Germany
denys.matthies@igd-r.fraunhofer.de

Abstract. In this paper we present a brief summary of an online survey we conducted in 2014. 135 participants successfully completed this survey, whereby 46% of the subjects were females and 54% males. We found out, that nowadays many users fall back on using smartphones in order to orientate themselves in unknown environments. In terms of analog navigation methods, people still rely on street name signs and landmarks such as characteristic buildings. However, relying on current smartphone habits for navigation tasks, visual attention is usually heavily drawn, which can cause a reduced perception and potentially makes smartphone map navigation more dangerous.

Keywords: Pedestrian Navigation, Urban Complexity, Survey, Human-Computer Interaction, Smartphone Usage, User Habits.

1 Introduction

The goal of this study was to understand 1) the usage patterns of classical ‘analog’ and new ‘digital’ guidance as well as 2) the process of planning paths and routes. While landmarks can already provide sufficient information for orientation, many users tend to use smartphones for getting directions. Previous studies have already investigated the smartphone usage in certain locations (e.g. at home, at the office) as well as the awareness of location-based services among smartphone users and non-smartphone users [3,4]. It has been found that directions as well as nearby points of interests (POI), such as shops or restaurants, have high recognizability and thus a great potential usage by both user groups [4]. Moreover, especially when involved in traffic as a car driver, devices for navigation became very popular. In 2013 three-quarters of car drivers used navigation devices, whereby every fifth device was a smartphone [2]. However, this survey aims to provide an insight into the target group of the pedestrian to find out everyday scenarios and common issues while navigating in public space with classical approaches and technology-assisted approaches.

2 Survey

2.1 Preface

Survey Instruments. The frequency of use of classical concepts and orientation aids in comparison to the use of digital and mobile applications for pedestrian has not been investigated in any study yet. To quantify this information we designed a complex questionnaire based on three survey guidelines based on Wester et al. [9], Kirchhoff et al. [5] and Aschemann-Pilshofer [1].

Questionnaire Content. The survey included primarily dictated closed questions, which had to be rated on different rating scales. Additionally, the participants were also able to respond with qualitative feedback in corresponding text boxes. 16 carefully chosen questions cover the following areas:

- daily locomotion
- use of classical/analog guidance
- memorizing unknown routes
- use of digital devices and digital map applications, and route planners
- behavior in unknown scenarios.

Evaluation and Statistics. To carry out the study, we used SoSci from Leiner [6]. The survey was online and accessible for 25 days. Within this time, the survey has been successfully completed 135 times, whereby 46% of the participants were females and 54% males. 61% of all participants were younger than 30 years old, 30% between 30–49 years old and 9% had an age of 50 or above. 80% rated themselves as an intensive smartphone user. 39% of all participants stated to usually take the car, while 46% use public transportation and 15% use both transportations equally.

2.2 Usage of Map Applications and Route Planners

To determine the frequency of usage of digital map applications and route planners we asked the users how often and on which device they are using navigation services.

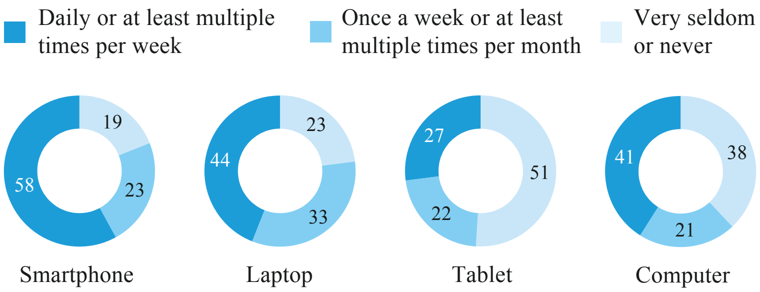


Figure 1. Demonstrating the frequency of use of directions for four devices: Smartphone, Laptop, Tablet, Computer.

While more than half of the respondents use their smartphone regularly for navigation/directional tasks, smartphones are also used noticeably more often than other devices *(see Figure 1).*

2.3 Memorizing Directions to Unknown Destinations

The survey participants were asked about their most likely behavior when planning a route from home to an unknown destination. We evaluated this question by the different user groups (gender, main transportation, usage of smartphone and age) to identify variances *(see Figure 2).*

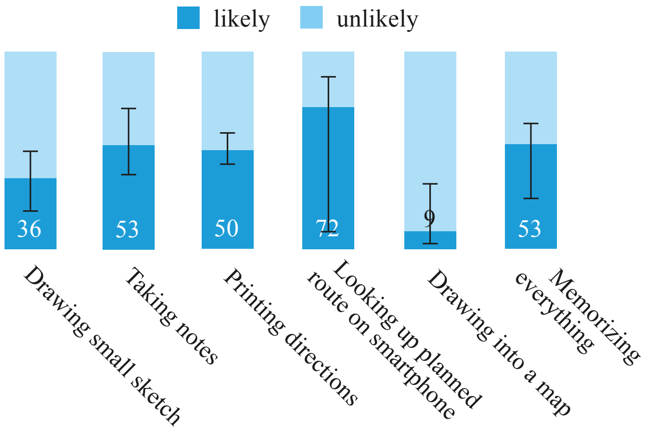


Figure 2. Preferred methods of memorizing when planning an unknown route from home.

Almost half of the respondents reported to possibly not *memorize the planned route* to the unknown destination. Tools such as *taking notes* and *printing directions* seem to be quite popular while *drawing into a map* is only reported to be used rather unlikely. 72% of the respondents reported *to look up directions on the smartphone.* Here, the variance seems to be very high because of certain user groups (intensive usage of smartphone: 87%, extensive usage: 8%).

2.4 Analog Navigation Methods

To gain an insight on how often alternative methods for orientation are being used *(see Figure 3)*, we asked the survey participants to rate the suggested alternative as either *often, sometimes, seldom* or *never* based on their subjective perception, since it is hardly measurable with numbers.

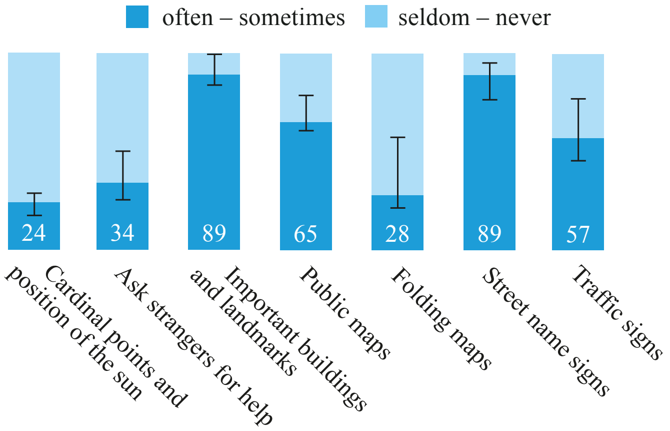


Figure 3. User tendency to alternative navigation aids.

The most significant navigation aids are *important buildings and landmarks* (89%), which was equale to *street name signs* (89%). Also striking is, that the user group above 50 years (58%) still relies on *folding maps*, which does not seem to be an option for younger users (<30 years: 21%, 30–49 years: 34%).

2.5 Smartphone Usage in Different User Groups

The survey included several scenarios, in which the participant had to rate the most appropriate answer on a 5-point Likert scale. In this case, we asked the respondent to imagine him-/herself being on the way as a pedestrian, while getting lost and searching the way to a place (such as the flat of a friend). The statement to be rated was: *"I will check the directions on my smartphone." (see Figure 4).*

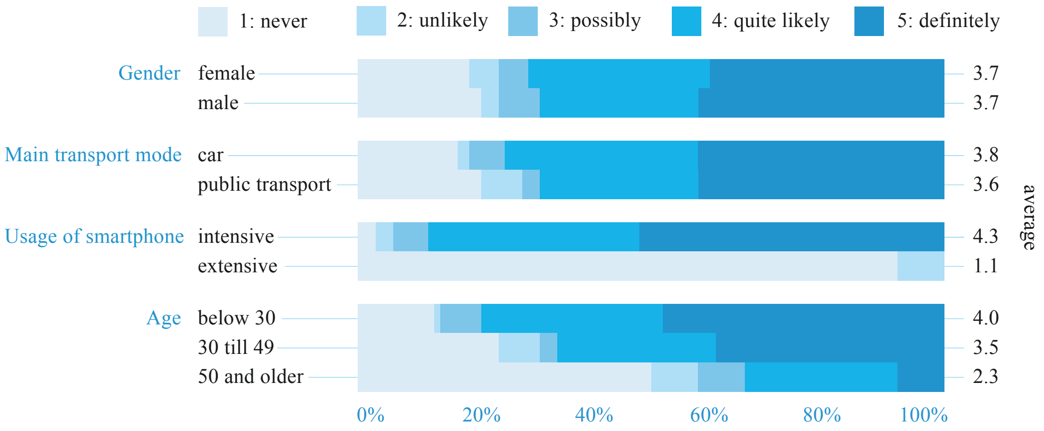


Figure 4. This figure shows the answer for different user groups.

As the result we can observe that users with the daily locomotion by *car* tend to use the smartphone in such situations more often. The discrepancy between genders does not yield any noticeable differences. Also clearly to see, young users most likely tend to use their smartphone and older people tend to use alternatives.

2.6 Qualitative Feedback (Excerpt)

Besides the quantitative rating, most questions were provided with text boxes for additional feedback. Especially this qualitative feedback turned out to provide us with a lot more valuable information, since people were already telling us about many problems or different solutions they find crucial when orientating and navigating in public space. In the following we will highlight some statements, which state problems already very clearly.

Heavy visual attention on the screen. *»I enter the address at home on my smartphone to leave on time and to follow the directions on the go.«*

High demand on memorization causes cognitive load. *»I look up the directions on the internet [at my workstation] and take the nearby station as a starting point. I always try to remember the route to my final destination.«*
»I often use Google Street View. There I can see striking points, which will help me to get my bearings on the go.«

Combining classical 'analog' and new 'digital' guidance. *»If I did not find the street, I maybe would have asked another pedestrian, with having Google Maps open on the smartphone, for help.«*

3 Conclusion

In this paper, we briefly presented insights in user habits when navigating in unknown environments. We can clearly see that smartphones not just caught up with common navigation devices – instead even emerged to the mainly used device for pedestrian navigation. However, smartphone navigation usually demands heavy visual attention, which makes it potentially dangerous to use. We believe that alternative navigation methods, such as vibrotactile feedback [8], can help here in order improve navigation experience for pedestrians and to make it safer. This paper only presents a brief summary with a small number of questions we asked the study participants. Further information about this survey can be found in the Master's Thesis of Anita Meier [7].

Acknowledgements. We would like to thank all those who co-operated in conducting this survey and for allowing themselves to be interviewed. Without their help, this survey could not have been carried out. This work has been funded by the federal state of Mecklenburg-Vorpommern and EFRE within the project "Basic and Applied Research in Interactive Document Engineering and Maritime Graphics" and by the European Social Fund under grant ESF/IV-BM-B35-0006/12.

References

1. Aschemann-Pilshofer, B. (2001) Wie erstelle ich einen Fragebogen. Ein Leitfaden für die Praxis. 2. Auflage. *Graz: Wissenschaftsladen Graz.* URL: *http://www.aschemann.at/Downloads/ Fragebogen.pdf* Retrieved: 2013/08/13.
2. BITKOM (2013a) Jeder dritte Smartphone-Nutzer teilt seinen Standort mit. *Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V.* URL: *http://www.bitkom.org/de/markt_ statistik/77793_77354.aspx* Retrieved: 2013/11/15.
3. BVDW/TNS (2013a): Studie zur Smartphone-Nutzung und ihren Einsatzorten. *Bundesverband Digitale Wirtschaft (BVDW) e.V. in Kooperation mit TNS Infratest.* URL: *http://www.bvdw.org/mybvdw/media/download/bvdw-tns-mobileclub-einsatzorte.pdf?file=2744* Retrieved: 2013/11/15.
4. BVDW/TNS (2013b) Studie zur Bekanntheit und Nutzung von Location-Based-Services (LBS) bei Besitzern und Nicht-Besitzern mobiler Devices. *Bundesverband Digitale Wirtschaft (BVDW) e.V. in Kooperation mit TNS Infratest.* URL: *http://www.bvdw.org/mybvdw/media/download/chartband-bvdw-mobile-daten-dienste.pdf?file=2615* Retrieved: 2013/11/15.
5. Kirchhoff, Sabine / Kuhnt, Sonja / Lipp, Peter / Schlawin, Siegfried (2001) Fragebogen: Datenbasis. Konstruktion. Auswertung. 2. überarbeitete Auflage. *Opladen: Leske + Budrich.*
6. Leiner, D.J. (2013) SoSci Survey (Version 2.3.05-i). URL: *https://www.soscisurvey.de/* Retrieved: 2013/11/24.
7. Meier, A. (2014). Orientieren mit allen Sinnen – Multisensorische Wahrnehmung und Orientierung am Beispiel vibro-taktiler Fußgängernavigation. *Master's Thesis. University of Applied Sciences Potsdam (FHP).*
8. Meier, A., Matthies, D.J.C., Urban, B., Wettach, R. (2015). Exploring Vibrotactile Feedback on the Body and Foot for the Purpose of Pedestrian Navigation. In *2nd international Workshop on Sensor-based Activity Recognition and Interaction (iWOAR2015)* in Rostock, Germany. ACM.
9. Wester, F., Soltau, A., Paradies, L. (2006) Hilfestellung zur Gestaltung eines Fragebogens. Landesinstitut für Schule, Bremen. URL: *http:// www.lis.bremen.de/sixcms/media.php/ 13/Skript%20Fragebogenerstellung.7024.pdf* Retrieved: 2013/08/13.

Towards integration and management of contextualized information in the manufacturing environment by digital annotations

Rebekka Alm[1,2] and Steffen Hadlak[1,2]

[1] Universität Rostock, 18051 Rostock, Germany
[2] Fraunhofer IGD, Joachim-Jungius-Str. 11, 18059 Rostock, Germany
{rebekka.alm;steffen.hadlak}@igd-r.fraunhofer.de

Abstract. Advanced manufacturing promises an evolution of industrial production processes by increasing flexibility and specialization of work tasks to deal with mass customization. To maintain a high quality and efficiency despite this increasing customization or even improve them, intelligent assistance systems are required supporting the workers.

This paper describes how to integrate digital information in a manufacturing environment, where workers use assistance systems to access task related information. To explain requirements and constraints of assistance systems, a survey was conducted. Based on the results of this survey, a conceptual approach is specified that focuses on quick and easy access to relevant information via a tablet. To provide manufacturing workers with relevant information, a method is presented to measure information relevance based on an ontology. A demonstrative scenario describes the application of the conceptual approach.

Key words: Ontology, digital annotations, context, manufacturing

1 Introduction

Widen-Wulff [29] emphasizes the importance of knowledge sharing in organizations because of their increasing complexity and growing scale of information activities. One example concerns the advances of industrial manufacturing processes. To cope with the increasing mass customization the former rigid product processes are substituted by more flexible yet also more specialized processes. To maintain an equally high quality and efficiency despite the increased flexibility "information and knowledge are the firm's strategically most important resources today" [29]. At the same time the intellectual resources are difficult to manage and require intelligent assistance systems that support the individuals such as workers.

In this regard, knowledge can be defined as "information processed by individuals including ideas, facts, expertise, and judgments relevant for individual, team, and organizational performance" [28]. Even as the importance of information sharing is widely accepted [20, 28], motivation and communication barriers are still a great obstacle to sharing knowledge [7]. To identify these obstacles and possible motivators multiple studies have been conducted [7, 12, 28] depending on individual characteristics and situational perceptions. An important result of these studies is that especially people who perceive significant time pressure are less likely to share knowledge while perceived competition was not directly related to knowledge sharing.

Traditional knowledge management tools are often provided stand-alone and rely on the user to explicitly search for additional information due to a demand regarding his current work task [18, 28]. That means a user who is already under time pressure has to switch between different systems and perform further tedious interactions to gain information. We aim to integrate the knowledge management fully into their daily work so the user does not have to switch between systems to gain additional information. We propose the usage of ontology-based annotations as an intuitive way to integrate work task related information into an intelligent assistance system, enabling workers: (1) to access contextually relevant information based on the task that they perform and (2) to easily create and share useful information with their co-workers. Contextually relevant content is automatically recommended to the user based on an ontology modelling the domain.

After investigating necessary aspects of such a system in Section 2 by a survey, we consider related work in Section 3 especially in the area of digital annotations. Afterwards, we explain our concept in detail in Section 4 using the example of supporting an assembly worker and discuss key points of a visual integration of the annotations into an assistance system in Section 5. A conclusion summarizes our results.

2 Survey: general demands for assistance systems

A survey was conducted to enquire the general demands for work task related information as well as factors influencing the willingness to use an information assistance system.

2.1 Test subjects

31 test subjects from Germany participated in an online survey questioning their opinion regarding work task related information and support by an assistance system. The age of the test subjects ranged from 22 to 51. Most of them did not have much experience with information assistance systems. The background of the test subjects covers a variety of professions from medicine, economics, law, education, engineering, IT consulting, administration, research and manual work.

2.2 Questions

The subjects were asked to answer questions to the following topics:

System demands: The subjects were asked to evaluate the usefulness of different kinds of work task related information, such as an overview of the work task, a detailed work step instruction, information regarding involved tools and materials, and tips for improving the task itself.
Furthermore, the subjects were asked to evaluate different features of an information assistance system in respect to their significance to support the worker in his work task, such as usability, information quantity and quality, and transparency.

Usage factors: The subjects were asked to assess what kind of information they would share in what extent and to evaluate different concerns in respect to working with an information assistance system, such as lack of time or motivation and the averseness to being monitored.

The subjects were asked to assess specific options of the topics with a four-level Likert scale. Furthermore, they were given the chance to give additional feedback as free text.

2.3 Results

In the following, the most important findings regarding the topics are summarized.

System demands. Figure 1 shows which content users want and expect of an assistance system. A process overview, a detailed explanation of the current work step, and an error detection were valued positively by 96.77% of the subjects. Tips for improvement and remarks by colleagues were valued by 87.1% each, and a technical discussion between colleagues was valued by 83.87%. Additional

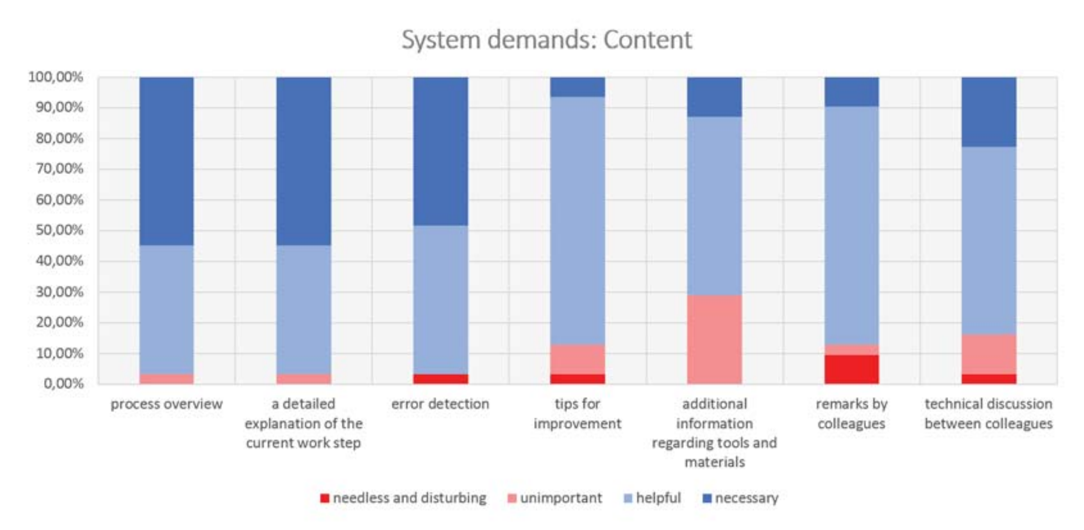


Fig. 1. Demands of the survey participants regarding the content of an assistance system.

information regarding tools and materials was valued the least with 70.96% of positive assessment.

When asked to evaluate specific kinds of additional information provided by experienced colleagues, remarks regarding the tool were estimated to be mostly "helpful" by 70.96%, "interesting" by 22.58% and "unimportant" by 6.45%. Nobody estimated this information to be "needless and disturbing". Remarks regarding the materials were estimated to be "helpful" (67.74%) and "interesting" (32.26%). Tips for improvement were estimated to be "helpful" by 45.16%, "interesting" by 48.39% and "unimportant" by 6.45%.

When asked about the importance of given system features (see Figure 2), the test subjects evaluated "providing relevant information of high quality" as most important (with the values of 83.87% "necessary" and 16.13% "important"), followed by an "easy and intuitive handling", "interactivity", "unobtrusiveness" and "sensible dealing with the data (anonymity)". "Quantity and multitude of information" was valued the least (12.90% "necessary", 58.06% "important", 16.13% "unimportant", and even 12.90% "needless and disturbing"). Additionally, the subjects named a reasonably small latency as very important.

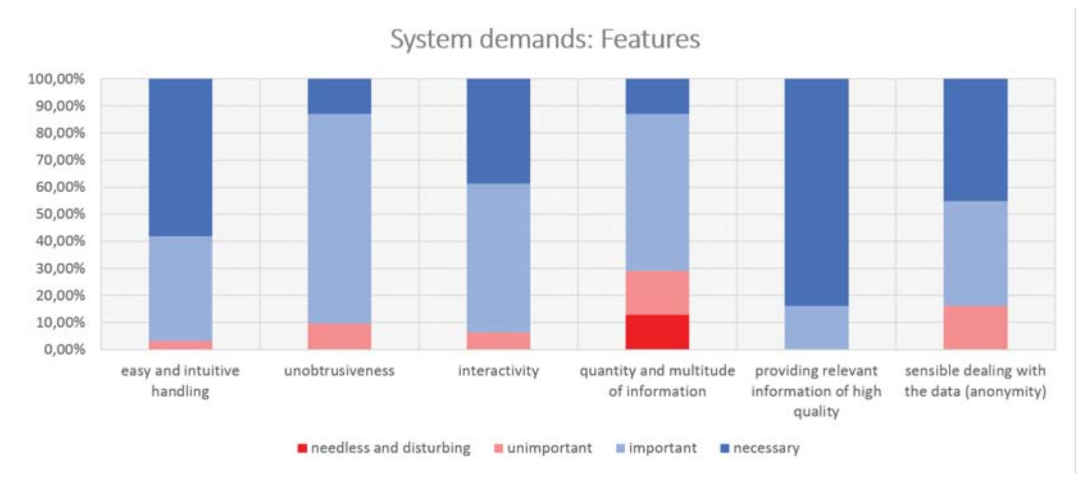


Fig. 2. Demand of the survey participants regarding the importance of features of an assistance system.

As further demands and nice-to-haves, the subjects named especially the possibility to give feedback (for correction or evaluation of the systems advice), integration of the support (by push-notifications, warnings), personal adjustment of the view and extent of support activities (more experienced users need less support), transparency (why did the system choose this; why should I follow its advice; what consequences are expected otherwise), and a display of remaining process time and how to save time (for a "smoke break").

Usage factors. When asked about their willingness to share their own experience with their colleagues, almost all of them (96.77%) were willing to share information about tools and materials in a digital way where everyone could access the information. They were a bit more reserved with information about found errors in the manual and tips for improving the work task. Here, 12.90% and 19.35% respectively would only share these information verbally.

29% of the subjects commented on their choice. All comments to this topic were like-minded: They would share their knowledge, because sharing of knowledge promotes the working atmosphere and helps everyone to improve themselves and the processes. They share as they want their colleagues to share their knowledge, too. The sharing is no problem if it does not result in further work. Sharing a more subjective opinion (such as improvements) only feels comfortably to them when performed verbally to prevent being negatively perceived as a "know-it-all".

When analyzing the evaluation of different concerns regarding the assistance system (See Figure 3), the greatest concerns are the fear of being surveilled (by the system 51.62%, by my supervisor 58.07%) and the concern of not having the time to maintain the system (51.62%). The other concerns scored significant less agreement: lack of willingness to maintain the system (25.8%), lack of motivation to deal with the system (19.35%), fear of being more distracted by the system than supported (19.35%), and fear of being replaceable, when they share their knowledge (19.35%).

Further concerns are that the system might not work correctly and that the supervisor is not really supporting the additional efforts of using the system. They fear requiring additional time and thus extra hours to maintain the system.

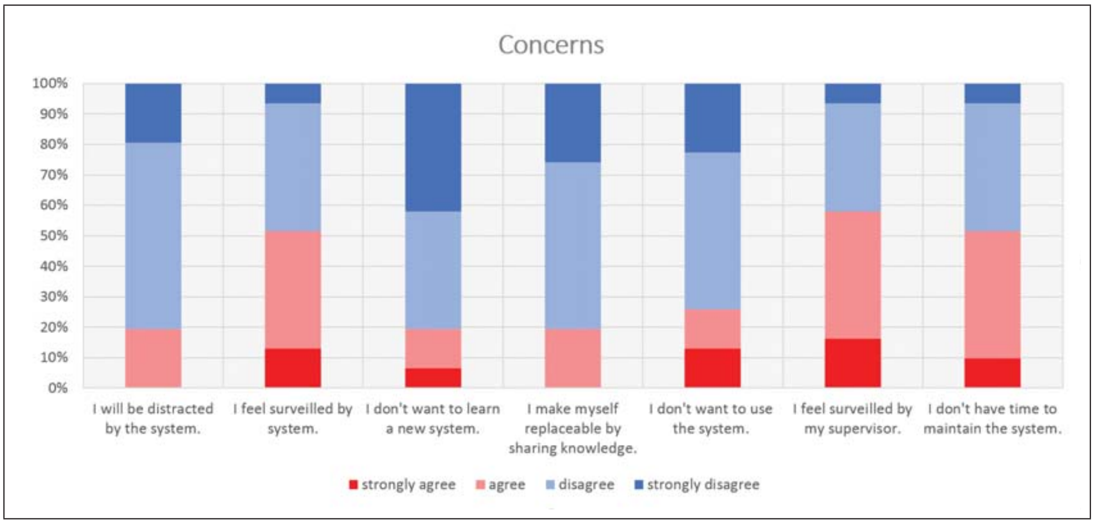


Fig. 3. Agreement of the survey participants to concerns regarding an assistance system supporting the work task.

2.4 Conclusion

Due to the rather small number of participants and their different qualifications, we cannot claim that the survey is representative, but a few general insights are still gained. We found out that the participants were generally open to a system that provides them with information. If the benefit is recognized, they would also share information in order to promote the team. We want to particularly emphasize that the highest ranked system features were "providing relevant information of high quality" followed by features alleviating the usage of the system. Furthermore, the greatest concern besides the averseness against being monitored was not having the time to maintain the system.

In summary, it can be stated that it is of great importance that the perceived benefit is greater than the perceived disadvantage. Consequently, additional efforts and especially the time required for using and maintaining the system have to be minimized. This tendency seems to be universal since participants of various fields agreed, so that an approach to solve this problem could also be of use for instance regarding workers in a production environment.

For this reason, a simple and intuitive information communication means is desirable. Digital annotations make this possible as we will show below.

3 Related Work

The digitization and accessibility of contextual relevant information is a broad field where diverse approaches are applied. Traditional knowledge management systems rely on the user to explicitly search for additional information by himself due to a demand regarding his current work task [18, 28]. But as our survey has shown the users require a more easy and intuitive means that minimizes this additional effort to get to the information. Consequently, we are focusing on ways that more automatically provide relevant information as it is typically done by recommender systems.

Recommender systems are a subclass of the information filtering systems that attempt to predict a "rating" or "preference" that a user would assign to an item. Traditional recommender systems neglect/disregard the notion of "situated actions" [26], the fact that users interact with the system within a particular "context" and that preferences for items within one context may be different from those in another context [2]. They simply produce a list of recommendations by collaborative or content-based filtering. Hence, context-aware recommender systems [2] define a context in order to create more intelligent and useful recommendations. Contextual factors such as time, location, purchasing purpose are then also taken into account.

Yet, recommender systems are mainly used for providing relevant information. They rarely provide intuitive and simple ways to create and integrate new information. However, as the survey has shown such an intuitive and simple access is necessary to reduce the time a user requires to share his knowledge. Regarding this demand for a simpler means of communication, basic forms of knowledge sharing are already performed using annotations by diverse groups of people. Students are writing annotations in their textbooks or people are putting sticky notes at objects to annotate them; the activity of annotating is easily done and useful to support information sharing and processing by human beings [6]. Therefore, it is not surprising that annotations are also used in the digital environment to enrich digital content with additional information [3].

The concept of digital annotation is not uniformly defined, but can be summarized as follows: *An Annotation is an object that contains information about one or more related entities.* Digital annotations are usually used for classification, documentation or communication. They can be provided in various forms to integrate different media depending on the purpose, e.g. as text, image, audio, video, etc. In general, two different kinds of digital annotations can be differentiated: Annotations for machine interpretation (often referred to as "semantic annotation") and annotations for human communication.

3.1 Annotations for Machine Interpretation

Most often, annotations are used to semantically enrich digital documents to support computers in processing and interpreting the information context [16]. Here, annotations classify documents or document sections by a word or word group using a standardized vocabulary. In this way, they support activities like searching for information, structuring and shaping a document as well as enabling service interoperability. Accordingly, they are of importance in semantic information retrieval [6].

In order to include the semantic context, various approaches use an ontology formalizing domain knowledge. Kara et al. [14] present an ontology-based information extraction and retrieval system applied to the soccer domain, while Nakatsuji et al. [21] use ontologies to index and classify a user's blog entries to derive the user's interest. The benefits of an ontology are widely recognized, but its biggest weakness is the complexity and expense of its creation and maintenance. Hence, Euzenat [8] proposes an ontology-based annotation approach in which the ontology should be expandable on the fly. Recently, Zhao and Ichise [30] proposed a Framework for InTegrating Ontologies (FITON), a semi-automatic system to integrate large and heterogeneous ontologies.

These annotations are very limited regarding their ability to carry information and thus cannot be used as a means for communication. Yet, this kind of annotations may support the automatic identification of context-relevant information.

3.2 Annotations for Human Communication

But annotations can also be used as a tool to support collaborative information exchange [16]. Known applications are different document readers (such as Adobe PDF Reader), that allow the user to add comments or notes (annotations) to the documents, or the collaboration features of MS Word (or similar software products) that enable the tracking and discussion of changes. Lortal et al.[17] describe a cooperative annotation tool used by a mechanical engineering team to discuss design drafts. These approaches often lack in indexing and annotation recall [16]. For most applications it is also not advisable, as the annotations are mostly valid for a very specific subject (e.g. an exact passage in a text). This annotation might not be of great use in another context.

Regarding repetitive "situations" such as work tasks which can be integrated in a broader context, we see advantages in retrieving associated annotations. Here, abnormalities in one work task are most likely also of interest regarding related or similar tasks. The classification of the annotations for re-usability in other contexts is interesting, but in the research area of annotations largely untreated. In [4] we presented a first concept for the use of contextualized annotations that are semantically enriched human annotations for an integrated visualization of heterogeneous manufacturing data. We extend this concept by applying these contextualized annotations within an assistance system as a communication medium for situation relevant information.

4 Conceptual Approach

Our objective is to provide the user with a quick and easy way to access and capture additional information about their current task (e.g., a worker assembling a certain machine). So, the overall team can benefit from an exchange of their experiences. For this purpose, we combine several ideas from the field of digital annotation to benefit from their advantages. By relying on the freedom of "annotations for human communication", the user should be able to digitally capture all sorts of information in a fast and easy way. This simplicity will minimize distractions from his primary task and thus motivate him to document his thoughts and experiences.

Yet, to present only context-relevant parts of this information to a user requires further steps. First, the information has to be put in its right context to make it versatilely reusable. We therefore apply the indexing mechanisms especially the underlying ontologies used for "annotations for machine interpretation" to the human readable annotations. Second, from this ontology modeling real-world objects and their relationships only that information has to be extracted that is actually relevant for a user's current task (such as information about the machine or necessary tools). Hence, we utilize recommendation mechanisms working on the structure of the ontology for this context-relevant information extraction. In summary, the following three steps have to be performed:

- Formalize domain and context knowledge as an *ontology*.
- *Capture additional information* as annotations and match them with corresponding ontology concept(s).
- Provide *contextually relevant* information for a given situation.

In the following, these steps are explained in more detail using the example of an assembly worker.

4.1 Modelling the Ontology

All relevant contextual elements of the work domain (such as work tasks, tools, parts or materials) are modeled by semantically interrelated concepts in a domain ontology. An ontology describes a shared conceptualization which formally represents a set of concepts and their relationships [10]. Since we want to extend the assistance of a worker using the ontology, it makes sense to represent the context of the work environment, in particular the work tasks. We are guided by the classical definition of context by Abowd et al. [1]. Here, a work task is defined by its relations to people, places and things. Accordingly, our ontology consists of four interrelated sub-ontologies:

- The *people* subgraph of the ontology formalizes the roles with their corresponding responsibilities and skills (e.g. planning engineers assign workers).
- The *places* subgraph of the ontology summarizes the local arrangements of the work places (e.g. the production hall contains different working groups).
- The *things* subgraph of the ontology summarizes all production-related materials such as tools, parts or products, and also encodes their composition as specific interrelations.
- The *work tasks* subgraph formalizes the different work tasks, from general work tasks such as monitoring, planning, and assembly to more specialized work tasks such as sticking and soldering. Moreover, the concepts of the *work tasks* subgraph connect all subgraphs.

The assembly *work task* in particular relates to all other subgraphs as it is performed by a worker (*people*), at a work station (*places*), and consumes parts to produce a product (*things*). In this way, all subgraphs of the ontology are interrelated through diverse work tasks. Figure 4 illustrates an example ontology.

4.2 Capturing Additional Information

To best support the worker in capturing additional information while minimizing the perceived disturbance, we adapt the metaphor of a sticky note. These sticky notes give the worker the ability to intuitively capture information in various forms, such as text, photo, audio, etc. To make this information widely available, they have to be assigned to ontology concepts describing those real world objects that the information annotates. As we are extending an assistance system with work task related information, this assignment can be done semi-automatically. The assistance system already knows about the current work task and can thus select those concepts of the ontology that represent the current work task and its context which are most likely to be annotated by the worker. In this way, the worker does not need to choose the corresponding concepts from all concepts of the ontology, but the specific concepts relevant to the current work tasks

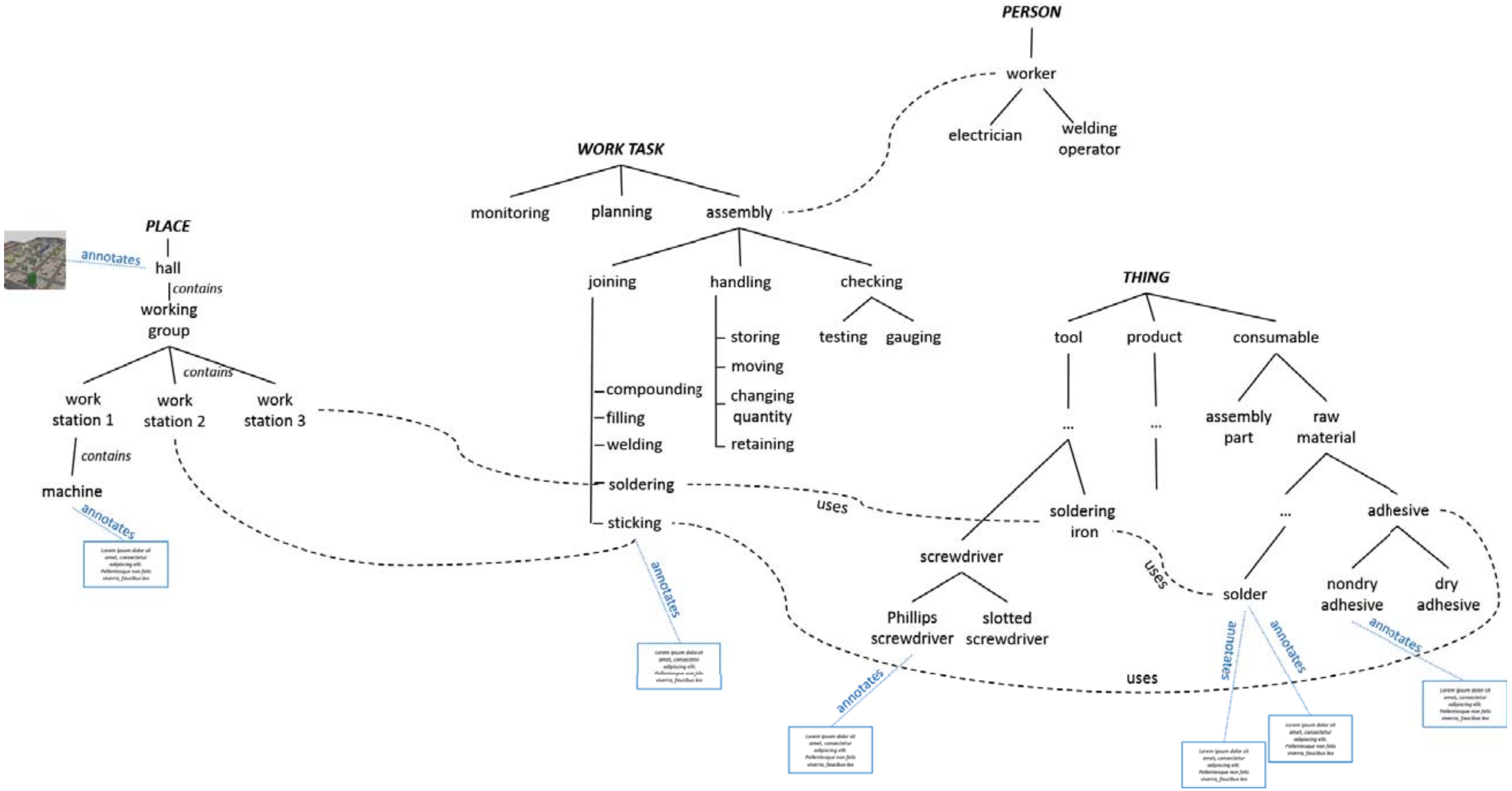


Fig. 4. Excerpt from a sample domain ontology. Straight lines represent hierarchical relationships, broken lines represent context-relationships, especially between the sub-ontologies (modelling the context between work tasks, people, places, and things). Annotations are displayed as blue boxes that are linked to at least one ontology concept.

are already available. From this preselection, the worker can then choose the concept(s) to annotate, for instance restricting it only to the machine and/or a used material.

4.3 Measuring the Information Relevance

For extracting context-relevant information to support a user in his current task but also for capturing additional information as described in the previous step, a measurement is necessary that indicates the relevance of an information to a specific "situation". In our case, such a situation describes a worker's current work task. As the ontology represents work tasks and their context, this situation can be identified as an ontology concept. Starting from the initial work task concept adjacent relationship paths (the edges of the ontology) can be followed to find related concepts annotated with additional information. Therefore, a relatedness measure can determine the degree to which a pair of concepts are related considering the whole set of semantic links among them [25]. We aim to adapt this notion to measure how relevant an annotation of a related concept is to the current work task.

Many publications cover special cases of semantic relatedness, the (semantic) similarity between objects [23]. They basically differ in which relationships they utilize for measuring relatedness in ontologies. Most often the hierarchical structure of the ontology is used to determine the semantic similarity between concepts [5, 27]. More recent research in the area of semantic relatedness consider different relationship types and thus also the non-hierarchical relationships in ontologies [19, 11, 22, 9]. That means starting from the initial work task concept, three types of relational directions can be identified that lead to information relevant to the situation and are shown in Figure 5:

1. **Upwards:** Path leads to more *general information* that annotates a parent concept. This information is more general and not restricted to our specific situation but still valid.
2. **Horizontal:** Path leads to *contextual information* that annotates concepts that are related (directly or indirectly) by contextual relationships (non-hierarchic relationships, especially cross-connections between subgraphs).
3. **Downwards:** Path leads to more *specific information* that annotates a child concept. This information is more specific, but under circumstances not relevant for the initial situation. It makes sense to choose the initial concept already as specific as possible (preferably a leaf concept) or to gradually specify the concept.

A common and very intuitive way to describe relatedness in a graph is based on the distance between two nodes which is basically the number of edges (relations) between them in the shortest path. In this sense, the shorter the path and thus the distance between two nodes, the more related they are. The problem with this approach is the assumption that the edges represent uniform distances within an ontology; i.e. the semantic connections are of equal weight [23]. Furthermore, the perception of similarity between concepts differs regarding the

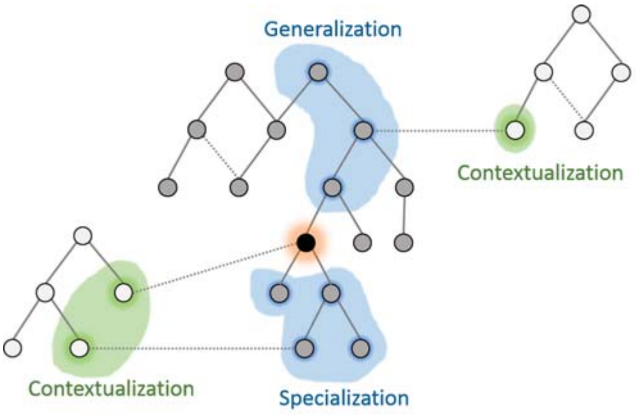


Fig. 5. Abstract ontology example highlighting which ontology concepts are of interest when looking for additional and relevant information.

relational directions as well as the types of relation. That means they have different influence on the measure and are often assigned weights to capture their importance.

Regarding hierarchical relations (upwards and downwards) especially path length and depth are of importance to get similarity results that compare to the human perception of similarity [15]. Since concepts located hierarchically deeper in the ontology are more specialized, the distance between those concepts is perceived as shorter than a distance of the same path length between concepts near the root. For instance, the path "Phillips screwdriver PH2 " - "Phillips screwdriver" - "screwdriver" is perceived shorter than the path "screwdriver" - "assembly tools for screws and nuts" - "tools" despite that both paths contain the same number of nodes and edges. Hence, the path depth should be included in the distance calculation. Jiang and Conrath [13] define an edge weighting that uses Resnik's [23] notion of a concept's information content implicitly containing path length and depth. We use this notion to weight upwards and downwards directed paths between two concepts c_1 and c_2:

$$W(path_H(c_1, c_2)) = |IC(c_1) - IC(c_2)| \tag{1}$$

The information content IC is defined as $IC(c) = -\log_2 p(c)$. The probability $p(c)$ of the occurrence of a concept is calculated by the count of concepts summarized by a parent concept as frequency $freq(c)$ and the count N of all concepts of the ontology with $p(c) = \frac{freq(c)}{N}$ [23]. The probability of a concept's occurrence decreases with hierarchical depth, while the information content increases.

As the information content is calculated according to the hierarchical structure of the ontology, a different weighting approach is needed for the horizontal non-hierarchical relations. Regarding this direction considering the relation type is of even higher importance compared to the upwards and downwards direction [19, 11, 22, 9]. For instance, the path "soldering"-*needs*-"soldering iron" should be ranked higher than the path “soldering”-*is located at*-“work station 3” as a worker knows where he actually is but could need additional information regarding specifics of the tool to use. We therefore adopt the formula by Mazuel and Sabouret [19]. They associate an individual weight TC_X to each relation type X to represent its semantic cost. The weight of a path between two concepts c_1 and c_2 is defined by:

$$W(path_X(c_1, c_2)) = TC_X \times \frac{|path_X(c_1, c_2)|}{|path_X(c_1, c_2)| + 1} \tag{2}$$

Both formulas only consider a single relationship type. Hence, to calculate the distance to any concept in the ontology the path which might contain mixed relationships has to be broken down into sub-paths with only a single relationship type. The weight of a mixed relationship path is then the sum of the weights of the sub-paths:

$$W(path(x, y)) = \sum\nolimits_{p \in path(x,y)} W(p) \tag{3}$$

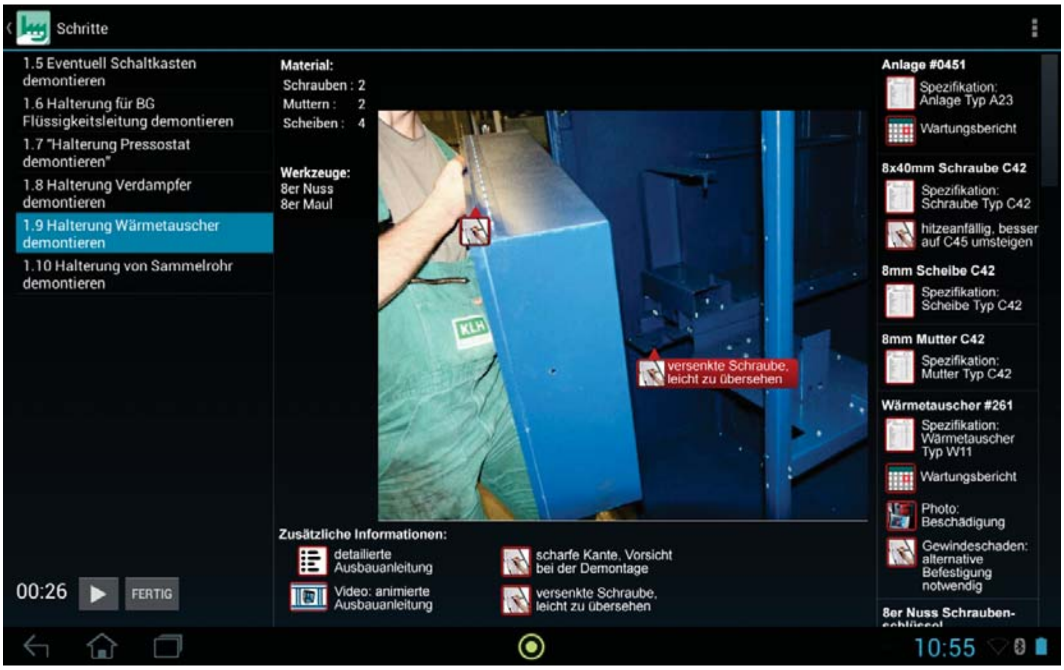
In this sense, concepts whose paths to the current work task score a lower path weight (are semantically closer) are assumed to be annotated with more relevant information than concepts scoring a higher path weight. Yet, some path constraints have to be considered to exclude non-relevant concepts that would possibly score a small path weight. While following upwards and downwards directed relationships (such as “is-A”, “include”) no change of direction should be performed. Other child concepts of the same parent concept (sibling concepts) are negligible for our retrieval even if they measure a short distance to the origin concept from an information theoretical point of view [23, 24]. As these concepts do not relate to the specific situation, considering them is expected to provide no meaningful improvements to our information retrieval. Hence, hierarchical paths are restricted only to direct ancestors or successors.

How this extracted context-relevant information can be represented to support a worker in fulfilling his current work task is discussed in the next section.

5 Integrated Presentation

Our application scenario is located within a networked factory. We envision that for assistance the worker or the assembly station is equipped with a tablet that provides him with information on the current assembly order and current process step. An example of such an assistance system is the assembly assistant of the Fraunhofer IGD[1]. We already emphasized the importance of integrating the communication of additional information directly into the work flow, as we found out that users prefer to not switch between different systems. Figure 6 shows a first design of how to integrate the annotations into the user interface of an assistance system.

[1] https://www.igd.fraunhofer.de/en/Institut/Abteilungen/IDE/Projekte/PlantHand-Assembly-Assistance-Production



1.5 Eventuell Schaltkasten demontieren
1.6 Halterung für BG Flüssigkeitsleitung demontieren
1.7 "Halterung Pressostat demontieren"
1.8 Halterung Verdampfer demontieren
1.9 Halterung Wärmetauscher demontieren
1.10 Halterung von Sammelrohr demontieren

Material:
Schrauben : 2
Muttern : 2
Scheiben : 4

Werkzeuge:
8er Nuss
8er Maul

versenkte Schraube, leicht zu übersehen

Zusätzliche Informationen:
detailierte Ausbauanleitung
Video: animierte Ausbauanleitung
scharfe Kante, Vorsicht bei der Demontage
versenkte Schraube, leicht zu übersehen

Anlage #0451
Spezifikation: Anlage Typ A23
Wartungsbericht

8x40mm Schraube C42
Spezifikation: Schraube Typ C42
hitzeanfällig, besser auf C45 umsteigen

8mm Scheibe C42
Spezifikation: Scheibe Typ C42

8mm Mutter C42
Spezifikation: Mutter Typ C42

Wärmetauscher #261
Spezifikation: Wärmetauscher Typ W11
Wartungsbericht
Photo: Beschädigung
Gewindeschaden: alternative Befestigung notwendig

8er Nuss Schrauben-

Fig. 6. An exemplary design of an assistance system showing multiple steps of an assembly work task on the left and a detailed description of the current step in the center. Annotations for the work task are embedded directly in the center view as sticky notes (see the two red annotations on top of the image) and summarized below. Additional relevant information regarding related concepts (such as reports to the machine and its parts, specifications of tools and materials) is integrated in a virtual clipboard on the right.

The main part of the interface is still dedicated for the current work tasks with the different steps on the left and a detailed representation of the current step in the center. Additional information regarding the task itself is directly embedded in the main view where applicable (see sticky notes on the image) and summarized below. Further information regarding related concepts are listed on the right and are ranked according to our explanation in Section 4.3.

Regarding the support of a worker to fulfill his current work task, horizontal relations to tools and materials are most likely of high interest to him. Yet, connections to person concepts are probably not so interesting for the worker himself (but may be for a planning engineer assigning the different workers). Hence, relationship types are assigned different weights according to their importance (the more important the smaller the weight). As these weights have to be determined relative to the overall information content of the ontology (as used to weight the hierarchical relations), and of the persona to be supported (worker, planning engineer ...) no general recommendations for these weights can be given at this point. Therefore, further investigations are necessary.

Based on these weights, annotations for concepts scoring the lowest path weights are retrieved. As annotations can be assigned to more than a single ontology concept, they can be retrieved more than once with different rankings (corresponding to the rankings of the ontology concepts). Our survey showed that users prefer quality of information over quantity, so recurring occurrences of one and the same annotation are filtered out.

To create new annotations the worker can simply select a concept by clicking in its area (in the center or on the right side) and choosing a type of annotation he wants to add (text, photo, audio ...) which opens a corresponding widget. He can either point in an empty part of a concept if he wants to annotate the concept in general, or he can select for instance a position in a text or image if he wants to annotate a specific part of its description.

6 Conclusions

As confirmed in our survey, one of the biggest barriers to knowledge sharing supported by a technical system is the fear that use and maintenance is too cumbersome and especially time-consuming. To address this problem, we presented a concept that integrates the knowledge sharing intuitively into the workflow. We have shown that annotations for human communication are an intuitive and simple means for this purpose and how to support the annotation's re-usability using a domain ontology. For a ranked retrieval of the annotations, we proposed a relatedness measure weighting paths within this ontology.

In future, we want to put our enhanced system in operation and conduct a user study with workers in production. Based on their feedback additional factors regarding the weighting of concepts can be evaluated. One aspect is the novelty of information, for instance the older the information the less important it might be. But also personal feedback may be an important aspect to steer which information will be shown. An assessment of the information rating relationship types in general but also individual annotations (from very helpful to wrong) could be integrated in the ranking of the information. Finally, it will be of importance to address the user's averseness of being monitored.

Acknowledgments. This research has been supported by the German Federal State of Mecklenburg-Western Pomerania and the European Social Fund under grant ESF/IV-BM-B35-0006/12.

References

1. G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles. Towards a better understanding of context and context-awareness. In *Proc. of the International Symposium on Handheld and Ubiquitous Computing*, pages 304–307, 1999.
2. G. Adomavicius and A. Tuzhilin. Context-aware recommender systems. *Recommender systems handbook*, pages 67–80, 2011.
3. M. Agosti and N. Ferro. A formal model of annotations of digital content. *ACM Transactions on Information Systems*, 26(1), Nov. 2007.

4. R. Alm, M. Aehnelt, S. Hadlak, and B. Urban. Annotated domain ontologies for the visualization of heterogeneous manufacturing data. In *Procs. of HCI International 2015*, Los Angeles, USA.
5. A. Bernstein, E. Kaufmann, C. Bürki, and M. Klein. How Similar Is It? Towards Personalized Similarity Measures in Ontologies. In O. K. Ferstl, E. J. Sinz, S. Eckert, and T. Isselhorst, editors, *Wirtschaftsinformatik 2005*, pages 1347–1366. Physica-Verlag HD, 2005.
6. G. Cabanac, M. Chevalier, C. Chrisment, and C. Julien. Collective annotation: Perspectives for information retrieval improvement. *Large Scale Semantic Access to Content*, pages 529–548, 2007.
7. C. E. Connelly, D. P. Ford, O. Turel, B. Gallupe, and D. Zweig. 'I'm busy (and competitive)!' Antecedents of knowledge sharing under pressure. *Knowledge Management Research & Practice*, 12(1):74–85, 2014.
8. J. Euzenat, O. Mbanefo, and A. Sharma. Sharing resources through ontology alignment in a semantic peer-to-peer system. *Cases on semantic interoperability for information systems integration: practice and applications*, pages 107–126, 2009.
9. G. Giray and M. O. Ünalir. A method for ontology-based semantic relatedness measurement. *Turkish Journal of Electrical Engineering & Computer Sciences*, 21:420–438, 2013.
10. T. Gruber. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(April):199–220, 1993.
11. A. Hawalah and M. Fasli. A graph-based approach to measuring semantic relatedness in ontologies. In *WIMS '11 Proc. of International Conference on Web Intelligence, Mining and Semantics*, page 29, 2011.
12. L. Hu and A. E. Randel. Knowledge Sharing in Teams: Social Capital, Extrinsic Incentives, and Team Innovation. *Group & Organization Management*, 39(2):213–243, Feb. 2014.
13. J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. on International Conference on Research in Computational Linguistics*, pages 19–33, Taiwan, 1997.
14. S. Kara, O. Alan, O. Sabuncu, S. Akpnar, N. K. Cicekli, and F. N. Alpaslan. An ontology-based retrieval system using semantic indexing. *Information Systems*, 37(4):294–305, June 2012.
15. Y. Li, Z. A. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15:871–882, 2003.
16. G. Lortal, M. Lewkowicz, and A. Todirascu-Courtier. Annotation: textual media for cooperation. *Proc. of International Workshop on Annotation for Collaboration (IWAC)*, pages 41–50, 2005.
17. G. Lortal, M. Lewkowicz, and A. Todirascu-Courtier. Enabling communication rationale via annotations: a document-based cooperation. *Proc. of COOP'06 (short paper)*, (March):75–82, 2006.
18. R. Maier and T. Hädrich. Knowledge Management Systems. In D. Schwartz, editor, *Encyclopedia of Knowledge Management*, pages 442–449. Idea Group Inc (IGI), 2006.
19. L. Mazuel and N. Sabouret. Semantic relatedness measure using object properties in an ontology. In A. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. Finin, and K. Thirunarayan, editors, *The Semantic Web - ISWC 2008*, volume 5318 of *Lecture Notes in Computer Science*, pages 681–694. Springer Berlin Heidelberg, 2008.

20. C. McInerney. Knowledge management and the dynamic nature of knowledge. *Journal of the American Society for Information Science and Technology*, 53(12):1009–1018, Oct. 2002.
21. M. Nakatsuji, Y. Miyoshi, and Y. Otsuka. Innovation Detection Based on User-Interest Ontology of Blog Community. *Proc. of International Semantic Web Conference (ISWC)*, pages 515–528, 2006.
22. G. Pirró and J. Euzenat. A feature and information theoretic framework for semantic similarity and relatedness. *The Semantic Web-ISWC 2010*, pages 615–630, 2010.
23. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. *Proc. of 14th Int. Joint Conf. on AI*, pages 445–453, 1995.
24. P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.
25. K. Rezgui, H. Mhiri, and K. Ghédira. Theoretical Formulas of Semantic Measure: A Survey. *Journal of Emerging Technologies in Web Intelligence*, 5(4):333–342, Nov. 2013.
26. L. A. Suchman. *Plans and situated actions: the problem of human-machine communication.* Cambridge university press, 1987.
27. G.-H. Wang, Y.-D. Wang, and M.-Z. Guo. An ontology-based method for similarity calculation of concepts in the semantic web. In *Machine Learning and Cybernetics, 2006 International Conference on*, pages 1538–1542, 2006.
28. S. Wang, R. a. Noe, and Z.-M. Wang. Motivating Knowledge Sharing in Knowledge Management Systems: A Quasi-Field Experiment. *Journal of Management*, 40(4):978–1009, July 2014.
29. G. Widen-Wulff. *The Challenges of Knowledge Sharing in Practice: A Social Approach.* Chandos Information Professional Series. Elsevier Science, 2014.
30. L. Zhao and R. Ichise. Ontology Integration for Linked Data. *Journal on Data Semantics*, May 2014.

supported by

Europäische Fonds EFRE, ESF und ELER
in Mecklenburg-Vorpommern

TREBING + HIMSTEDT

SKM Informatik

GI

MARINE SOFT

DVZ

BASIS GmbH

future TV

FORTecH Software

Fraunhofer
IGD

ISBN 978-3-8396-0960-6

9 783839 609606