Stereo Tracking of Faces for Driver Observation

Markus Steffens[1,2], Stephan Kieneke[1,2], Dominik Aufderheide[1,2], Werner Krybus[1],
Christine Kohring[1], and Danny Morton[2]

[1] South Westphalia University of Applied Sciences, Luebecker Ring 2,
59494 Soest, Germany
{steffens,krybus,kohring}@fh-swf.de
[2] University of Bolton, Deane Road, Bolton BL3 5AB UK
d.morton@bolton.ac.uk

Abstract. This report contributes a coherent framework for the robust tracking of facial structures. The framework comprises aspects of structure and motion problems, as there are feature extraction, spatial and temporal matching, re-calibration, tracking, and reconstruction. The scene is acquired through a calibrated stereo sensor. A cue processor extracts invariant features in both views, which are spatially matched by geometric relations. The temporal matching takes place via prediction from the tracking module and a similarity transformation of the features' 2D locations between both views. The head is reconstructed and tracked in 3D. The re-projection of the predicted structure limits the search space of both the cue processor as well as the re-construction procedure. Due to the focused application, the instability of calibration of the stereo sensor is limited to the relative extrinsic parameters that are re-calibrated during the re-construction process. The framework is practically applied and proven. First experimental results will be discussed and further steps of development within the project are presented.

1 Introduction and Motivation

Advanced Driver Assistance Systems (ADAS) are investigated today. The European Commission states their capabilities to weakening and avoiding heavy accidents to approx. 70% [1]. According to an investigation of German insurance companies, a quarter of all deadly car accidents are caused by tiredness [2]. The aim of all systems is to deduce characteristic states like the spatial position and orientation of head or face and the eyeballs as well as the clamping times of the eyelids. The environmental conditions and the variability of person-specific appearances put high demands on the methods and systems. Past developments were unable to achieve the necessary robustness and usability needed to gain acceptance by the automotive industry and consumers. Current prognoses, as in [2] and [3], expect rudimental but reliable approaches after 2011. It is expected, that those products will be able to reliably detect certain lines of sight, e.g. into the mirrors or instrument panel. A broad analysis on this topic can be found in a former paper [4].

A.-B. Salberg, J.Y. Hardeberg, and R. Jenssen (Eds.): SCIA 2009, LNCS 5575, pp. 249–258, 2009.

In this report a new concept for spatio-temporal modeling and tracking of partially rigid objects (Figures 1) is presented as was generally proposed in [4]. It is based on methods for spatio-temporal scene acquisition, graph theory, adaptive information fusion and multi-hypotheses-tracking (section 3). In this paper parts of this concept will be designed into a complete system (section 4) and examined (section 5). Future work and further systems will be discussed (section 6).

2 Previous Work

Methodically, the presented contributions are originated in former works about structure and stereo motion like [11, 12, 13], about spatio-temporal tracking of faces such as [14, 15], evolution of cues [16], cue fusion and tracking like in [17, 18], and graph-based modeling of partly-rigid objects such as [19, 20, 21, 22]. The underlying scheme of all concepts is summarized in Figure 1.

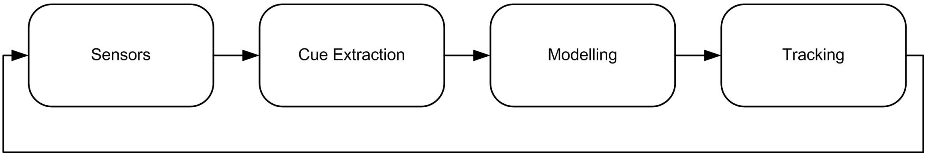

Fig. 1. General concept of spatio-temporal scene analysis for stereo tracking of faces

However, in all previously and further studied publications no coherent framework was developed like the one originally proposed here. The scheme was firstly discussed in [4]. This report originally contributes a more detailed and exact structure of the approach (section 3), a complete design of a real-world system (section 4), and first experimental results (section 5).

3 Spatio-temporal Scene Analysis for Tracking

The overall framework (Figure 1) utilizes information from a stereo sensor. In both views cues are to be detected and extracted by a cue processor. All cues are modeled in a scene graph, where the spatial (e.g. position and distance) and temporal relations (e.g. appearance and spatial dynamics) are organized. All cues are tracked over time. Information from the graph, the cue processor, and the tracker are utilized to evolve a robust model of the scene in terms of features' positions, dynamics, and cliques of features which are rigidly connected. Since all these modules are generally independent of a concrete object, a semantic model links information from the above modules into a certain context such as the T-shape of the facial features from eyes and nose. The re-calibration or auto-calibration, being a rudimental part of all systems in this field, performs a calibration of the sensors, either partly or in complete. The underlying idea is that besides utilizing an object model, facial cues are observed without a-priori semantic relations.

4 System Design and Outline

4.1 Preliminaries

The system will incorporate a stereo head with verged cameras which are strongly calibrated as described in [23]. The imagers can be full-spectrum or infrared sensors. During operation, it is expected that only the relative camera motion becomes un-calibrated, that is, it is assumed that the sensors reside intrinsically calibrated.

The general framework as presented in Figure 1 will be implemented with one cue type, a simple graph covering the spatial positions and dynamics (i.e. velocities), tracking will be performed with a Kalman filter and a linear motion model, re-calibration is performed via an overall skew measure of the corresponding rays. The overall process chain is covered in Figure 2. Currently, the rigidity constraint is implicitly met by the feature detector and no partitioning of the scene graph takes place. Consequently, the applicability of the framework is demonstrated while the overall potentials are part of further publications.

4.2 Feature Detection and Extraction

Detecting cues of interest is one significant task in the framework. Of special interest in this context is the observation of human faces. Invariant characteristics of human

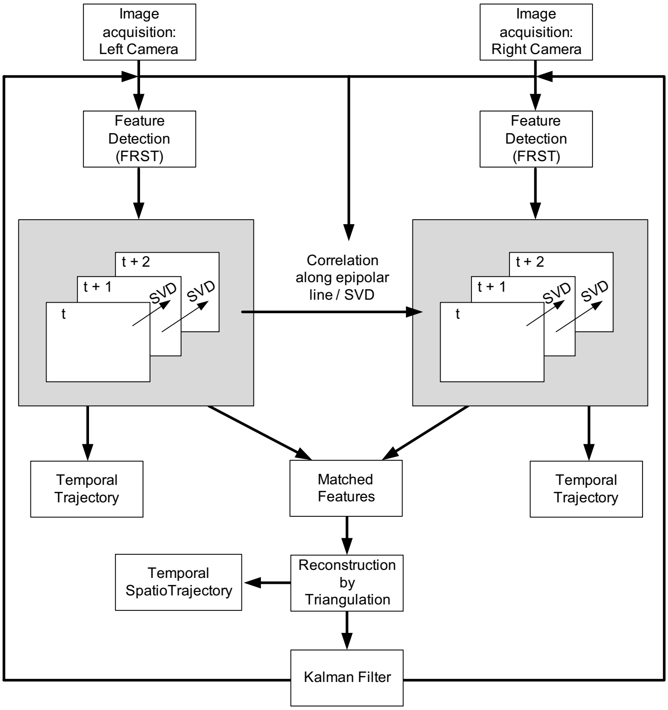
Image acquisition: Left Camera
Image acquisition: Right Camera
Feature Detection (FRST)
Feature Detection (FRST)
t + 2
t + 1
t
SVD
SVD
Correlation along epipolar line / SVD
t + 2
t + 1
t
SVD
SVD
Temporal Trajectory
Matched Features
Temporal Trajectory
Temporal SpatioTrajectory
Reconstruction by Triangulation
Kalman Filter

Fig. 2. Applied concept for tracking of faces

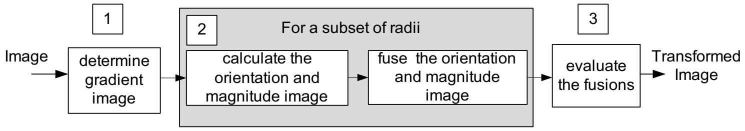

Fig. 3. Data flow of the Fast Radial Symmetry Transform (FRST)

faces are the pupils, eye corners, nostrils, top of the nose, or mouth corners. All offer an inherent characteristic, namely the presence of radial symmetric properties. For example a pupil has a shape as a circle and also nostrils have a circle-like shape. The Fast Radial Symmetry Transform (FRST) [5] is well suited for detecting such cues.

To reduce the search space in the images, an elliptic mask indicating the area of interest is evolved over the time [24]. Consequently, all subsequent steps are limited to this area and no further background model is needed.

The FRST further developed in [5] determines radial symmetric elements in an image. This algorithm is based on evaluating the gradient image to infer the contribution of each pixel to a certain centre of symmetry. The transform can be split into three parts (Figure 3). From a given image the gradient image is produced (1). Based on this gradient image, a magnitude and orientation image is built for a defined radii subset (2). Based on the resultant orientation and magnitude image, a resultant image is assembled, which encodes the radial symmetric components (3). The mathematical details would exceed the current scope; therefore have a look at [5]. The transform was extended by a normalization step such that the output is a signed intensity image according to the gradient's direction. To be able to compare consecutive frames, both half intervals of intensities are normalized independently yielding illumination invariant characteristics (Figure 6).

4.3 Temporal and Spatial Matching

Two cases of matches are to be established: the temporal (intra-view) and stereo matches. Applying FRST on two consecutive images in the left view, as well as in the right view, gives a bunch of features through all images. Further, the tracking module gives information of previous and new positions of known features. The first task is to find repetitive features in the left sequence. The same is true for the right stream. The second task is defined by establishing the correspondence between features from the left in the right view. Temporal matching is based on the Procrustes Analysis, which can be implemented via an adapted Singular Value Decomposition (SVD) of a proximity matrix $\mathbf{G}$ as shown in [7] and [6]. The basic idea is to find a rotational relation between two planar shapes in a least-squares sense. The pairing problem fulfills the classical principles of similarity, proximity, and exclusion. The similarity (proximity) $G_{i,j}$ between two features i and j is given by:

$$G_{i,j} = \left[e^{(-C_{i,j}-1)^2/2\gamma^2} \right] e^{-r_{i,j}^2/2\sigma^2} \quad (0 \le G_{i,j} \le 1) \tag{1}$$

where r is the distance between any two features in 2D and σ is a free parameter to be adapted. To account for the appearance, in [6] the normalized areal correlation index $C_{i,j}$ was introduced. The output of the algorithm is a feature pairing according to their locations in 2D between two consecutive frames in time from one view. The similarity factor indicates the quality of fit between two features.

Spatial matching takes place via a correlation method combined with epipolar properties to accelerate the entire search process by shrinking the search space to epipolar lines. Some authors like in [6] also apply SVD-based matching for the stereo correspondence, but this method only works well under strict setups, that are fronto-parallel retinas, so that both views show similar perspectives. Therefore, a rectification into the fronto-parallel setup is needed. But since no dense matching is needed [23], the correspondence search along epipolar lines is suitable. The process of finding a corresponding feature in the other view is carried out in three steps: First a window around the feature is extracted giving a template. Usually, the template shape is chosen as a square. Good results for matching are gained here for edge length between 8 and 11 pixel. Seconldy, the template is searched for along the corresponding epipolar line (Figure 5). According to the cost function (correlation score) the matched feature is found, otherwise none is found, e.g. due to occlusions. Taking only features from one view into account lead to less matches since each view may cover features which are not detected in the other view. Therefore, the previous process is also performed from the right to the left view.

4.4 Reconstruction

The spatial reconstruction takes place via triangulation with the found consistent correspondences in both views. In a fully calibrated system, the solution of finding the world coordinates of a point can be formulated as a least-square problem which can be solved via singular value decomposition (SVD). In Figure 9, the graph of a reconstructed pair of views is shown.

4.5 Tracking

This approach is characterized by feature position estimation in 3D, which is carried out by a Kalman filter currently [8] as shown in Figure 4. A window around the estimated feature, back-projected into 2D, reduces the search space for the temporal as well as the spatial search in the successive images (Figure 5). Consequently, computational costs for detecting the corresponding features are limited. Furthermore, features which are temporarily occluded can be tracked over time in case they can be classified as belonging to a group of rigidly connected features. The graph and the cue processor estimate their states from the state of the clique to which the occluded feature belongs.

The linear Kalman filter comprises a simple process model. The features move in 3D, so the state vector contains the current X-, Y- and Z-position as well as the feature's velocity. Thus, the state is the 6-vector $\mathbf{x} = [X, Y, Z, V_X, V_Y, V_Z]$. The process matrix $\mathbf{A}$ maps the previous position with the velocity multiplied by the time step to the new position $\mathbf{P}_{t+1} = \mathbf{P}_t + \mathbf{V}_t \Delta t$. The velocities are mapped identically. The measurement matrix $\mathbf{H}$ maps the positions from $\mathbf{x}$ identically to the world coordinates in $\mathbf{z}$.

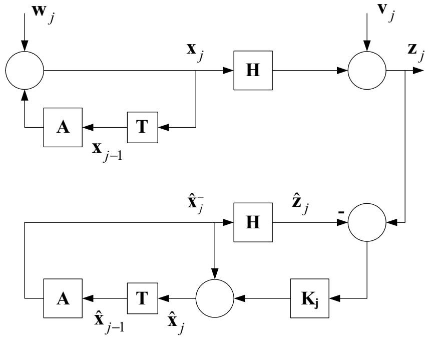

Fig. 4. Kalman Filter as block diagram [10]

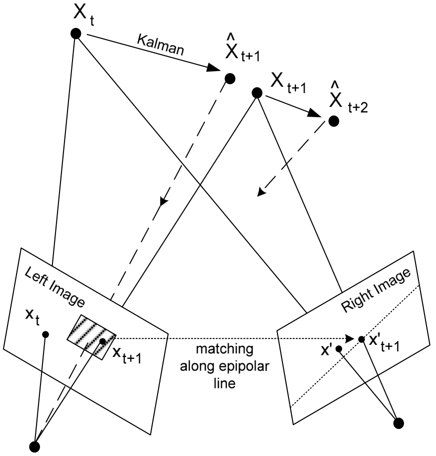

Fig. 5. Spatio-Temporal Tracking using Kalman-Filter

5 Experimental Results

An image sequence of 40 frames is taken exemplarily here. The face moves from the left to the right and back. The eyes are directed into the cameras, while in some frames the gaze is shifting away.

5.1 Feature Detection

The first part of the evaluation proves the announced property and verifies the robust ability of locating radial symmetric elements. The radius is varied by a fixed radial strictness parameter α . The algorithm yields the transformed images in Figure 6. The parameter for the FRST is a radii subset of one up to 15 pixels. The radial strictness parameter is 2.4. With exceeding a radius of 15 pixels, the positions of the pupils are highlighted uniquely. The same is true for the nostrils. By exceeding the radius of 6, the nostrils are extracted accurately. The influence of the strictness parameter α yields comparably significant results. The higher the strictness parameter, the more contour fading can be noticed. The transform was further examined under varying illumination and line-of-sights. The internal parameters were optimized accordingly with different sets of face images. The results obtained are conforming to those in [5].

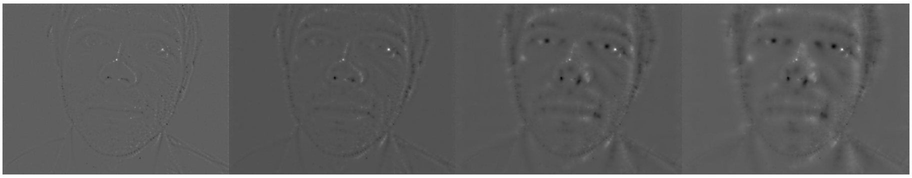

Fig. 6. Performing FRST by varying the subset of radii and fixed strictness parameter (radius increases). Dark and bright pixels are features with a high radial symmetric property.

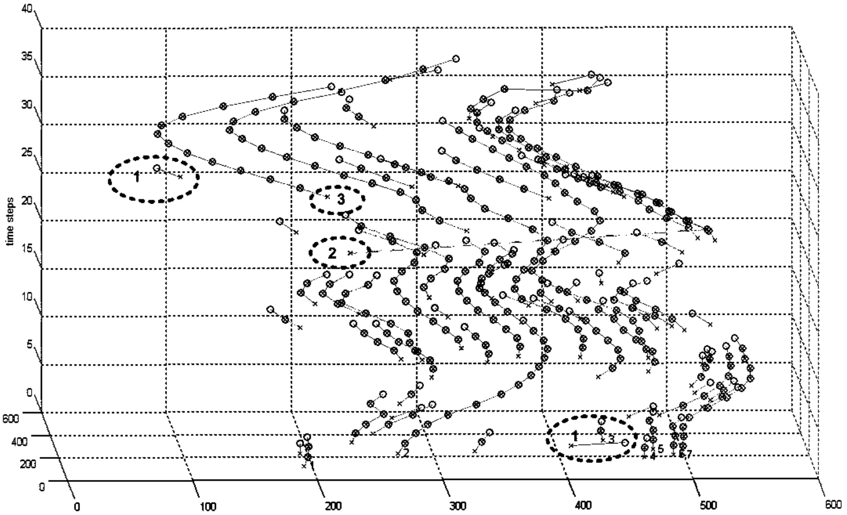

Fig. 7. Trajectory of the temporal tracking of the 40-frame sequence in one view. A single cross indicates the first occurrence of a feature, while a single circle indicates the last occurrence.

5.2 Matching

The temporal matching is performed as described. Figure 7 presents the trajectory of the sequence with the mentioned FRST parameters. A trajectory is represented by a line. Time is passing along the third axis from the bottom up. A cross without a circle indicates a feature appearing the first time in this view. A circle without cross encodes the last frame in which a certain feature appeared. A cross combined with a circle declares a successful matching of a feature in the current frame with the previous and following frame. Temporarily matched correspondences are connected by a line.

At first one is able to recognize an upstanding similar movement of most of the features. This movement has a shape similar to a wave. This correlates exactly to the real movement of the face in the observed image sequence. In Figure 10, there are four positions marked, which highlight some characteristics of the temporal matching. The first mark is a feature which was not traceable for more than one frame. The third mark is the starting point of a feature which is track-able for a longer time. In particular, this feature was observed in 14 frames. Noteworthy is the fact, that in this sequence no feature is tracked over the full sequence. It is not unusual due to the matter of the radial symmetric feature characteristic in faces. For example a recorded eye blink leads to a feature loss. Also, due to head rotations, certain features are rotated out of the image plane. The second mark shows a bad matching. Due to the rigid object and coherent movement, such a feature displacement is not realistic. The correlation threshold was chosen relatively low to 0.6, while it is working fine for this image sequence. For demonstrating the spatial matching, 21 characteristic features are selected. Figure 8 represents the results for an exemplary image pair.

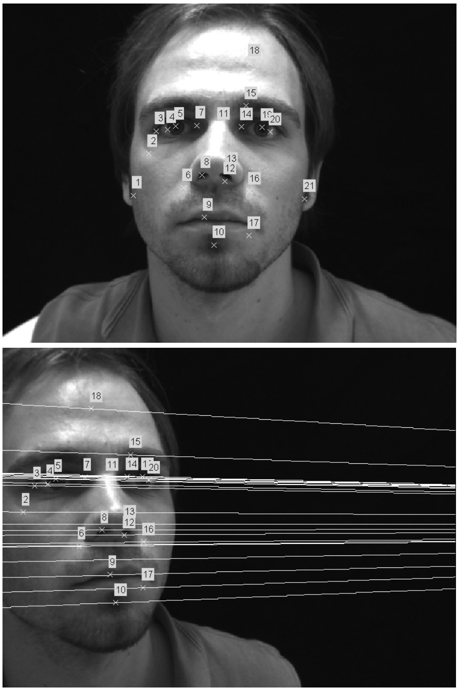

Fig. 8. Left Image with applied FRST, serves as basis for reconstruction (top); the corresponding right image (bottom)

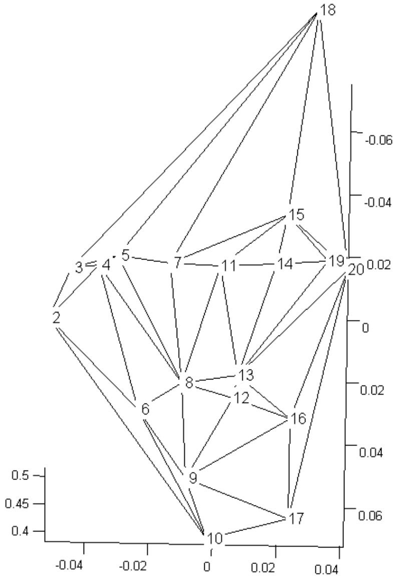

Fig. 9. Reconstructed scene graph of world points from a pair of views selected for reconstruction (scene dynamics excluded for brevity). Best viewed in color.

5.3 Reconstruction

The matching process on the corresponding right image is performed by applying areal correlation along epipolar lines [9]. The reconstruction is based on least-squares triangulation, instead of taking the mean of the closest distance between two skew rays.

Figure 8 shows the left and right view, which is the basis for reconstruction. Applying the FRST algorithm, 21 features are detected in the left view. The reconstruction based on the corresponding right view is shown in Figure 9. As one can see, almost the entire bunch of features from the left view (Figure 8, top) is detected in the right view. Due to the different camera positions, features 1 and 21 are not covered in the right image and consequently not matched. Although the correlation assignment criteria is quite simple, namely the maximum correlation along an epipolar line, this method yields a robust matching as shown in Figures 8 and 9. All features, except feature 18, are assigned correctly. Due to the wrong correspondence, a wrong triangulation and consequently a wrong reconstruction of feature 18 is the outcome as can be inspected in Figure 9.

5.4 Tracking

In this subsection the tracking approach will be evaluated. The previous sequence of 40 frames was used for tracking. The covariance matrices are currently deduced experimentally. This way the filter works stable over all frames. The predictions by the filter and the measurements lie on common trajectories. However, the chosen motion model is only suitable for relatively smooth motions. The estimates of the filter were further used during fitting of the facial regions in the images. The centroid of all features in 2D was used as an estimate of the center of the ellipse.

6 Future Work

At the moment there are different areas under research. Here, only some important should be named: robust dense stereo matching, cue processor incorporating fusion, graphical models, model fusion of semantic and structure models, auto- and re-calibration, and particle filters in Bayesian networks.

7 Summary and Discussion

This report introduces current issues on driver assistance systems and presents a novel framework designed for this kind of application. Different aspects of a system for spatio-temporal tracking of faces are demonstrated. Methods for feature detection, for tracking in the 3D world, and reconstruction utilizing a structure graph were presented. While all methods are at a simple level, the overall potentials of the approach could be demonstrated. All modules are incorporated into a working system and future work is indicated.

References

[1] European Commission, Directorate General Information Society and Media: Use of Intelligent Systems in Vehicles. Special Eurobarometer 267 / Wave 65.4. 2006

[2] Büker, U.: Innere Sicherheit in allen Fahrsituationen. Hella KGaA Hueck & Co., Lippstadt (2007)

[3] Mak, K.: Analyzes Advanced Driver Assistance Systems (ADAS) and Forecasts 63M Systems For 2013, UK (2007)

[4] Steffens, M., Krybus, W., Kohring, C.: Ein Ansatz zur visuellen Fahrerbeobachtung, Sensorik und Algorithmik zur Beobachtung von Autofahrern unter realen Bedingungen. In: VDI-Konferenz BV 2007, Regensburg, Deutschland (2007)

[5] Lay, G., Zelinsky, A.: A fast radial symmetry transform for detecting points of interest. Technical report, Australien National University, Canberra (2003)

[6] Pilu, M.: Uncalibrated stereo correspondence by singular valued decomposition. Technical report, HP Laboratories Bristol (1997)

[7] Scott, G., Longuet-Higgins, H.: An algorithm for associating the features of two patterns. In: Proceedings of the Royal Statistical Society of London, vol. B244, pp. 21–26 (1991)

[8] Welch, G., Bishop, G.: An introduction to the kalman filter (July 2006)

[9] Steffens, M.: Polar Rectification and Correspondence Analysis. Technical Report Laboratory for Image Processing Soest, South Westphalia University of Applied Sciences, Germany (2008)

[10] Cheever, E.: Kalman filter (2008)

[11] Torr, P.H.S.: A structure and motion toolkit in matlab. Technical report, Microsoft Research (2002)

[12] Oberle, W.F.: Stereo camera re-calibration and the impact of pixel location uncertainty. Technical Report ARL-TR-2979, U.S. Army Research Laboratory (2003)

[13] Pollefeys, M.: Visual 3Dmodeling from images. Technical report, University of North Carolina - Chapel Hill, USA (2002)

[14] Newman, R., Matsumoto, Y., Rougeaux, S., Zelinsky, A.: Real-Time Stereo Tracking for Head Pose and Gaze Estimation. In: FG 2000, pp. 122–128 (2000)

[15] Heinzmann, J., Zelinsky, A.: 3-D Facial Pose and Gaze Point Estimation using a Robust Real-Time Tracking Paradigm, Canberra, Australia (1997)

[16] Seeing Machines: WIPO Patent WO/2004/003849

[17] Loy, G., Fletcher, L., Apostoloff, N., Zelinsky, A.: An Adaptive Fusion Architecture for Target Tracking, Canberra, Australia (2002)

[18] Kähler, O., Denzler, J., Triesch, J.: Hierarchical Sensor Data Fusion by Probabilistic Cue Integration for Robust 3-D Object Tracking, Passau, Deutschland (2004)

[19] Mills, S., Novins, K.: Motion Segmentation in Long Image Sequences, Dunedin, New Zealand (2000)

[20] Mills, S., Novins, K.: Graph-Based Object Hypothesis. Dunedin, New Zealand (1998)

[21] Mills, S.: Stereo-Motion Analysis of Image Sequences. Dunedin, New Zealand (1997)

[22] Kropatsch, W.: Tracking with Structure in Computer Vision TWIST-CV. Project Proposal, Pattern Recognition and Image Processing Group, TU Vienna (2005)

[23] Steffens, M.: Close-Range Photogrammetry. Technical Report Laboratory for Image Processing Soest, South Westphalia University of Applied Sciences, Germany (2008)

[24] Steffens, M., Krybus, W.: Analysis and Implementation of Methods for Face Tracking. Technical Report Laboratory for Image Processing Soest, South Westphalia University of Applied Sciences, Germany (2007)