Towards a Smart Control Room for Crisis Response Using Visual Perception of Users

Joris Ijsselmuiden
Fraunhofer IITB Karlsruhe
iss@iitb.fraunhofer.de

Rainer Stiefelhagen
Institute for Anthropomatics
University of Karlsruhe (TH)
Fraunhofer IITB Karlsruhe
stiefelhagen@ira.uka.de

Florian van de Camp
Fraunhofer IITB Karlsruhe
ca@iitb.fraunhofer.de

Alexander Schick
Fraunhofer IITB Karlsruhe
sci@iitb.fraunhofer.de

Michael Voit
Fraunhofer IITB Karlsruhe
vt@iitb.fraunhofer.de

ABSTRACT

Due to ever increasing challenges and complexity, there is a high demand for new human-machine interaction approaches in crisis response scenarios. In the framework of the five-year Fraunhofer internal project "Computer Vision for Human-Computer Interaction – Interaction in and with attentive rooms" we aim at building a smart crisis control room, in which vision-based perception of users will be used to facilitate innovative user interfaces and to support teamwork. Our smart control room is equipped with several cameras and has a videowall as the main output and interaction device. Using real-time computer vision, we can track and identify the users in the room and estimate their head orientations and pointing gestures. In order to build a useful smart control room for crisis response, we are currently focusing on situation modeling for such rooms, and we are investigating the target crisis response scenarios. This paper gives an overview of the project, presents our ongoing work and discusses future work.

Keywords

Computer vision, tracking, head pose estimation, gesture recognition, situation modeling, user modeling, human-machine interaction, intelligent interfaces, crisis response

INTRODUCTION

Human-centered interaction is becoming more and more important as computers are finding their way into every imaginable scenario. Consequently, the demand for nontraditional user interfaces is growing rapidly. In traditional human-computer interaction, users rely on keyboard and mouse to control any given application; they are bound to these devices. Other possibilities should be explored more extensively, because the traditional setup is suboptimal in cases where people work in teams or with large displays (Miller, Robinson, Wang, Chung and Quek 2006). The research group "Computer Vision for Human-Machine Interaction" at the Fraunhofer IITB Institute uses computer vision and a wide array of other techniques to develop alternatives to the traditional mouse/keyboard controlled GUI. We are working towards a real-world application: the *control room for crisis response* (i.e. fire brigade, ambulance, and police emergency command center). To clarify our goals, we start out with some concrete examples.

When somebody enters a control room, his or her identity can be obtained in a convenient, unobtrusive way through face recognition. A user model can then be used to generate a personal user interface, obeying the user's preferences, current tasks, and specialized knowledge. Using person tracking (Bernardin, Van de Camp and Stiefelhagen 2007), these interfaces can be displayed at a location close to the user. Equally convenient, the user manipulates objects on a videowall simply by pointing (Nickel and Stiefelhagen 2007) and directing his or her visual attention (Voit and Stiefelhagen 2008). When combined with a range of hand gestures, powerful, natural, and highly unobtrusive interaction is possible. Head pose and attention tracking can also be employed to analyze the interaction of the team, for example to determine who has been talking to whom.

SYSTEM OVERVIEW

Our experiments take place in a laboratory of about six by nine meters (See Figure 1). Four corner cameras and a fish-eye camera at the center of the ceiling observe the users in the room. A videowall serves as the prime output device. The software architecture, displayed in Figure 2, consists of three layers: *perceptual intelligence*, *situation modeling*, and *human-machine interaction*. The first layer, perceptual intelligence, contains three computer vision modules: *tracking and identification*, *visual focus of attention*, and *gestures and body pose*. These receive live video feeds from multiple cameras, and they provide the situation modeling layer with high-level information about the people in the control room: where are they located, who are they, what are they looking at, and what gestures are they making. Note that the fourth module in the perceptual intelligence layer, *speech recognition*, is a topic for future work (Section 6). It has not yet been added to the system.

Figure 1. Our laboratory for smart control room research (the circles highlight two of the corner cameras)

In the second layer, situation modeling, the high-level information generated by the perceptual intelligence layer is processed further into an elaborate model of the current situation in the control room. This model is used as input for the third layer: human-machine interaction. Its main task is to maintain intelligent user interfaces on the main output device: a videowall with a screensize of four meters in width and one meter fifty in height (4096 x 1536 pixels). It is worth noting that the perceptual intelligence layer and the human-machine interaction layer are also connected directly, without the situation modeling layer in between. This applies to straightforward interaction that does not require any high-level intelligence such as controlling a mouse cursor by pointing at the videowall.

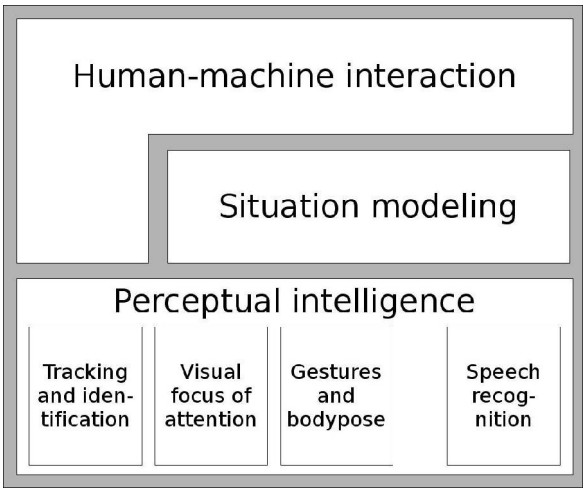

Figure 2. Software architecture, consisting of three layers

PERCEPTUAL INTELLIGENCE

To obtain multiple camera coverage of the entire room, every upper corner is fitted with a fixed, calibrated camera and the center of the ceiling with a fisheye camera. Using such a camera setup, the location, identity,

body pose and actions of all users can be perceived using state of the art computer vision techniques. In the following sections we describe these perceptual components. See (Stiefelhagen, Bernardin, Ekenel and Voit 2008) for a more detailed description of such components.

Tracking and Identification

User's locations and identities are arguably the most fundamental aspects of the situation in a smart crisis response control room. They are vital to the other perceptual intelligence modules as well as the situation modeling and human-machine interaction layers on top of them. To determine somebody's head pose or hand gestures, the system needs to know where he or she is located. Similarly, to say anything about users' roles, capabilities, and preferences, the system needs to know their identities. Figure 3 provides an image from a ceiling-mounted fisheye camera with estimations of people's locations. Our group developed multiple components for tracking people and their identities.

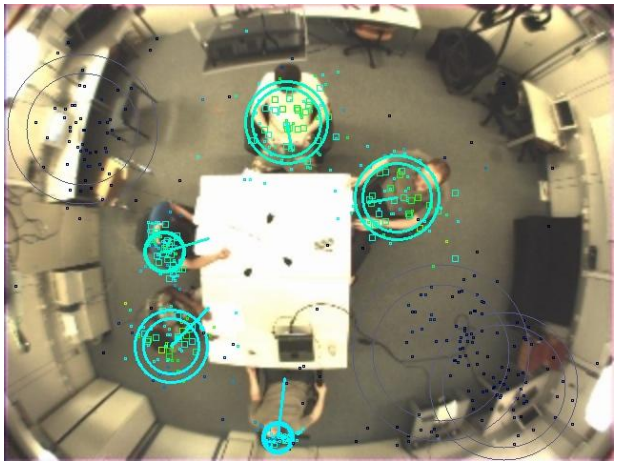

Figure 3. Person tracking as seen from a fish-eye camera

Visual tracking is done in 3D with a multi-level particle filter approach which can be combined with additional modalities like audio (Bernardin, Gehrig and Stiefelhagen 2007). We use a novel local appearance-based face identification approach (Ekenel and Stiefelhagen 2005). This approach has been proven to work successfully in smart environments and has also achieved state-of-the-art results on the commonly used face recognition benchmarks (Ekenel, Jin, Fischer and Stiefelhagen 2007). We are currently working on integrating all available information more tightly. In a joint framework for person tracking and identification, we have shown that the probabilistic integration of multiple modalities significantly improves performance (Bernardin, Stiefelhagen and Waibel 2008).

Other ongoing work includes the incorporation of prior knowledge about the room and the objects in it. For example, early experiments have shown significant improvement in the accuracy of track generation and track disappearance, when explicitly modeling entry and exit zones derived from door locations. Our current focus lies on integrating such additional information in a way that the system can profit from all of it while not depending on any of it.

Visual Focus of Attention

In the interaction between humans, the visual focus of attention is an important clue. It is therefore a natural choice to also integrate this mode of interaction into a smart control room. What people are looking at gives us an indication of their intentions and current awareness. Additional user-specific information can be displayed on the videowall, at the user's current visual focus of attention. Also, we can make people aware of what they haven't seen yet, if we follow their gaze's trajectory. Another advantage is that one can analyse the interaction that takes place between users, by monitoring their visual focus of attention. Finally, in combination with further modalities, such as pointing gesture recognition, the inclusion of the visual focus increases expressive power, robustness (coping with ambiguities) and ease of use (Kortum 2008).

The unobtrusiveness of the room's sensor setup prevents the classification of visual foci from eye gaze patterns directly. The room is too large to obtain reliable close-ups of all users' pupils without falling back to head-mounted or otherwise close by positioned cameras. Our approach therefore utilises an approximation of eye gaze by estimating head orientation (Voit, Nickel and Stiefelhagen 2007) and an individual mapping

from the recognised viewing frustums to most likely focus targets (Voit and Stiefelhagen 2008) (Figure 4). Typical focus targets are other people, specific areas on the videowall, and other objects in the room.

Figure 4. Estimating head poses of people in a meeting and deriving the most likely focus targets

Our system takes advantage of all available cameras, by applying a neural network-based estimator on each view to gather individual likelihood distributions from all camera angles concerning the person's head rotation. The single hypotheses are then joined via a Bayesian filter into a final estimate. The successive derivation of the most likely focus target, given this estimated viewing frustum's direction, then builds on observations of the physiological relationship between eye gaze and head movements during fixation on visual targets as presented in (Freedman and Sparks 1997). A parallel and online adaptation of each individual's mapping parameters further allows us to cope with dynamic and unforeseeable interaction patterns, such as a sudden change of potential focus targets. This can occur for example when other people enter or leave the room or relatively static objects are moved (Voit et al. 2007; Voit and Stiefelhagen 2008).

Gestures and Bodypose

In a smart control room, users must be able to interact with objects on the videowall to solve the crisis situation. We are developing a gesture interface that supports multiple users and multiple simultaneous inputs per user without using a pointing-device like a mouse. With existing touchscreens supporting multiple inputs, users are bound to the screen and the system does not know which input belongs to which user. Our solution allows (identified) users to move freely through the room and still interact with the videowall.

We use a voxel-based approach to approximate the visual hulls of the people in the control room (Figure 5). To achieve real-time operation, we compute the voxels belonging to the visual hulls on a graphic card's GPU using NVidia's CUDA framework. Tracking the whole body instead of only the hands increases the robustness of our system by integrating more information about the users.

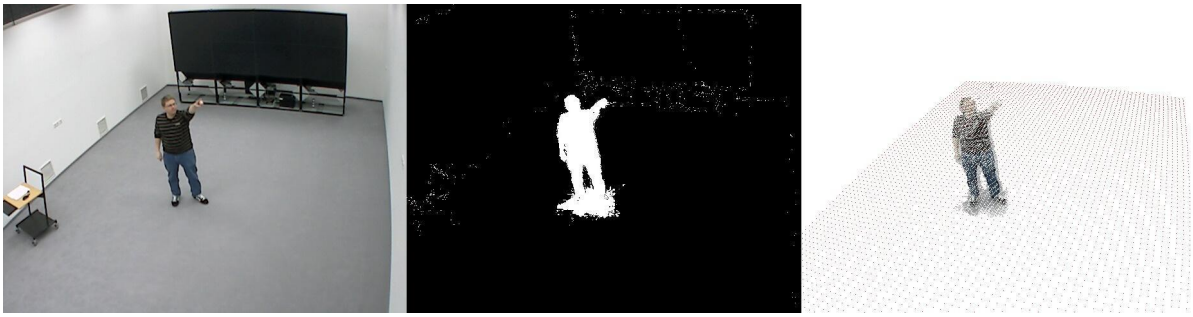

Figure 5. One of the camera views with the corresponding segmentation and voxel back-projection

When using very small voxels, the visual hulls are similar to point clouds. These 3D point clouds can be clustered using Euclidean distance and neighbouring relations to extract separate people in the scene. In the next processing step, the articulated structure of each person is computed using a variation of the iterative

closest point registration algorithm (Ziegler, Nickel and Stiefelhagen 2006). To add robustness to the system, we will integrate colour information for every visible voxel. This allows us to identify body parts like arms even if they are close to the body (Caillette and Howard 2004).

In (Nickel and Stiefelhagen 2007), it is shown that the three phases of pointing gestures; begin, hold, and end, can be detected using one single HMM for each phase. We will apply this technique to our extracted hand and arm trajectories to recognize gestures and, in combination with the head orientation, determine the location pointed at. By integrating these techniques, the users will be able to manipulate objects on the videowall using intuitive gestures from any point in the smart control room. To obtain richer and more robust interaction, we will supplement the pointing interface with other types of hand gestures such as grabbing and circling motions.

SITUATION MODELING

The second layer, situation modeling, maintains an elaborate, real-time model of everything that takes place in the room in order to appropriately react and adapt to the users. The perceptual intelligence layer's output consists mainly of user descriptions containing identity, location, head orientation, and hand gestures. The situation modeling layer uses these descriptions as input and is not concerned with the underlying mechanics. As already mentioned, the user descriptions are also forwarded directly to the human-machine interaction layer to allow for straightforward interaction. Pointing gestures for example, can be directly used to control a mouse pointer. Similarly, someone's location can be used to determine the best location for a personal interface without using situation modeling.

But we want to have a bit more than straightforward interaction. We would like the room to really be aware of the situation. To improve awareness we could for example derive which people are working together (subgroup constellations). Also, we want the room to possess user, situation and context dependence. This requires reasoning and user modeling to process the incoming facts into higher level facts. We also require the situation modeling layer to learn from previously seen situations.

Many things need to be considered here. Uncertainty and fuzziness as well as people's mental states have to be modeled. Like the perceptual intelligence layer, the situation modeling should largely run in real time. To shrink search-spaces and minimize complexity issues, it is key to properly define prior knowledge of how humans and objects typically move and behave. Further prior knowledge can be formalized from the task domain. An appealing approach is to use qualitative (logic-based) methods (Doyle and Thomason 1999) rather than classical, quantitative reasoning. See (Brdiczka, Crowley, Curín and Kleindienst 2009; Crowley 2006) for modeling approaches similar to our plans. The current design leads in the direction of a multimodal predicate logic incorporating time and user's mental states.

HUMAN-MACHINE INTERACTION

In our experiments, people will control the computer using their locations, identities, visual focus of attention, and hand gestures. A videowall will be the prime output device (Figures 1 and 6). We are analysing the applicability of these new forms of interaction to *computer supported cooperative work* and we will improve current solutions for it. By improving the intelligence and context dependance of the user interface through situation modeling, we can obtain a better understanding of what is going on in the room.

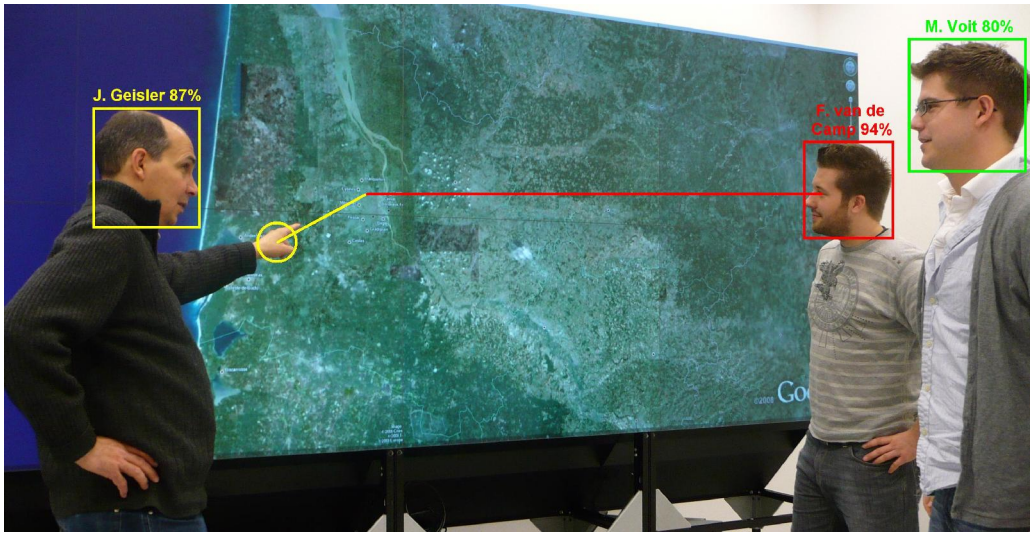

Figure 6. Example: interaction with our videowall through hand gestures and head pose

A typical crisis response control room for fire brigade, ambulance, and police is operated by a team of five or more staff members. The key is to find system setups that significantly improve the human-machine interaction of a group of users (i.e. fire brigade, ambulance, and police commanders) operating a crisis response control room. The current situation in such rooms is being analysed (Ivergard and Hunt 2008) and from there, stepwise, modular improvements are made at the three layers of our architecture (See Figure 2). The comparison of different input/output-modality combinations will be evaluated in *user studies*. Similarly, different perception and modeling approaches are interchanged and tested for performance.

The crisis response control room is a very challenging application due to team based operation, a high cost of failure, time pressure, dense and complex information, the user acceptance problem, and the demand for unobtrusiveness (Ivergard and Hunt 2008). The system setup we aim for improves expressive power, ease of use, intuitiveness, speed, reliability, adaptability, and cooperation while reducing physical and cognitive workload.

CONCLUSIONS AND FUTURE WORK

In this paper, we have presented research-in-progress towards building a smart crisis control room. Our work combines computer vision, situation modeling, and human-machine interaction to improve current control room solutions. A number of vision-based real-time perceptual components to perceive the locations, identities, head orientations and body poses of the users are already in place, but need to be improved further. These components can already be used for simple location and gesture based interaction with the videowall in the room.

A large part of the project must still be classified as future work. First of all, there is the speech recognition module. It forms an essential part of the targeted multimodal interaction setup and it will provide the situation modeling layer with information about who is saying what (Figure 2). As we focus on computer vision research, we plan to use third-party speech recognition software instead of developing our own. Clearly, speech recognition can increase the user's expressive power and ease of use in robust multimodal settings with the three visual modules (Kortum 2008). For instance, a constrained set of speech commands can provide the categorical information (what should happen), while pointing gestures take care of the deictic aspect (where should it happen) (Kortum 2008; Miller et al. 2006). Likewise, gesture recognition combined with visual focus of attention results in more reliable measurements because of their redundancy (Chai, Hong and Zhou 2004; Kortum 2008, Miller et al. 2006).

Second, we will integrate active pan-tilt-zoom cameras in the near future. They will be able to perform better on detailed analysis such as face recognition than the five fixed, calibrated cameras already installed. Active pan-tilt-zoom cameras can be controlled by the system itself and thus provide the control room with more detailed views and active vision (Bernardin et al. 2007).

Third, we need to design and implement the right situation modeling layer for the job. Artificial intelligence has a long history full of potentially useful methods. Choosing the right ones and using and adapting these for this entirely new application forms an important part of our future work.

Fourth, to obtain a full-fledged smart control room, additional output devices will be integrated to support the videowall: A *digital situation table* (Figure 7), tablet PCs, and a speaker system (for speech synthesis for example). During the past years, Fraunhofer IITB has been developing a digital situation table with multitouch capability and tablet PCs for crisis response (Bader, Meissner and Tschnerney 2008). Our task for the future is to integrate this in a new setting. Another addition could be cameras that look over the videowall surface to also provide that screen with multitouch functionality.

Figure 7. Digital situation table and tablet PCs in a map application

Finally, a lot of work remains to be done to apply everything to a real-world crisis response control room. Section 5 introduced some design issues that need to be considered here. Furthermore, we are thoroughly studying current control rooms to optimally make use of existing knowledge and experience (Ivergard and Hunt 2008). The end-user should be able to interact with the control room in much the same way as humans interact with each other; by using speech, hand gestures, and directing one's gaze. Therefore, the human body itself is the most important "input device". Iterative design through user studies with expert participants from the field of crisis response as well as novice users is the way to go forward here.

ACKNOWLEDGMENTS

This work is supported by the FhG Internal Programs under Grant No. 692 026. More information can be found through our webpage (Stiefelhagen and Ijsselmuiden 2009).

REFERENCES

1. Bader, T., Meissner, A., Tschnerney, R. (2008) Digital Map Table with Fovea-Tablett®: Smart Furniture for Emergency Operation Centers *Proceedings of the 5th International Conference on Information Systems for Crisis Response and Management*, Washington, DC.
2. Bernardin, K., Gehrig, T., Stiefelhagen, R. (2007) Multi-Level Particle Filter Fusion of Features and Cues for Audio-Visual Person Tracking *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*, 70-81, Springer LNCS 4625.
3. Bernardin, K., Van de Camp, F., Stiefelhagen, R. (2007) Automatic Person Detection and Tracking using Fuzzy Controlled Active Cameras *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN.
4. Bernardin, K., Stiefelhagen, R., Waibel, A. (2008) Probabilistic Integration of Sparse Audio-Visual Cues for Identity Tracking *Proceeding of the ACM International Conference on Multimedia*, New York, NY.
5. Brdiczka, O., Crowley, J. L., Curín J., Kleindienst, J. (2009, to appear) Chapter: Situation Modeling, in Waibel, A., Stiefelhagen, R. (Eds.) *Computers in the Human Interaction Loop*, 121-132, Springer, Human-Computer Interaction series.
6. Caillette, F., Howard, T. (2004) Real-Time Markerless Human Body Tracking with Multi-View 3-D Voxel Reconstruction *Proceedings of the 3rd IEEE and ACM International Symposium on Mixed and Augmented Reality*, Arlington, VA.

7. Chai, J. Y., Hong, P., Zhou, M. X. (2004) A Probabilistic Approach to Reference Resolution in Multimodal User Interfaces *Proceedings of the International Conference on Intelligent User Interfaces*, New York, NY.
8. Crowley, J. L. (2006) Social Perception *ACM Queue*, 4, 6, 34-43.
9. Doyle, J., Thomason, R. (1999) Background to Qualitative Decision Theory *AI Magazine*, 20, 2, 55-68.
10. Ekenel, H. K., Stiefelhagen, R. (2005) A Generic Face Representation Approach for Local Appearance Based Face Verification *Face Recognition Grand Challenge Experiments Workshop, at the IEEE Conference on Computer Vision and Pattern Recognition,* San Diego, CA.
11. Ekenel, H. K., Jin, Q., Fischer, M., Stiefelhagen, R. (2007) ISL Person Identification Systems in the CLEAR 2007 Evaluations *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*, 256-265, Springer LNCS 4625.
12. Freedman, E., Sparks, D. (1997). Eye-Head Coordination during Head-Unrestrained Gaze Shifts in Rhesus Monkeys *Journal of Neurophysiology*, 77, 2328-2348.
13. Ivergard, T., Hunt, B. (2008) Handbook of Control Room Design and Ergonomics: A Perspective for the Future, Second Edition, CRC.
14. Kortum, P. (Ed.) (2008) HCI Beyond the GUI: Design for Haptic, Speech, Olfactory and Other Nontraditional Interfaces, Morgan Kaufmann.
15. Miller, C., Robinson, A., Wang, R., Chung, P., Quek, F. (2006) Interaction Techniques for the Analysis of Complex Data on High-Resolution Displays *Proceedings of the 8th International Conference on Multimodal Interfaces*, New York, NY.
16. Nickel, K., Stiefelhagen, R. (2007) Visual Recognition of Pointing Gestures for Human-Robot Interaction *Image and Vision Computing*, 25, 12, 1875-1884.
17. Stiefelhagen, R., Bernardin, K., Ekenel, H. K., Voit, M. (2008) Tracking Identities and Attention in Smart Environments - Contributions and Progress in the CHIL Project *Proceedings of the IEEE International Conference on Face and Gesture Recognition*, Amsterdam, Netherlands.
18. Stiefelhagen, R., Ijsselmuiden, J. (2009) Website: Computer Vision for Human-Machine Interaction at Fraunhofer IITB *http://www.iitb.fraunhofer.de/servlet/is/20718*.
19. Voit, M., Nickel, K., Stiefelhagen, R. (2007) Head Pose Estimation in Single- and Multi-View Environments - Results on the CLEAR'07 Benchmarks *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*, 307-316, Springer LNCS 4625.
20. Voit, M., Stiefelhagen, R. (2008) Deducing the Visual Focus of Attention from Head Pose Estimation in Dynamic Multi-view Meeting Scenarios *Proceedings of the 10th International Conference on Multimodal Interfaces*, Chania, Greece.
21. Ziegler, J., Nickel, K., Stiefelhagen, R. (2006) Tracking of the Articulated Upper Body on Multi-View Stereo Image Sequences *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, New York, NY.