Towards High-Level Human Activity Recognition through Computer Vision and Temporal Logic

Joris Ijsselmuiden[1] and Rainer Stiefelhagen[1,2]

[1] Fraunhofer IOSB Karlsruhe
[2] Institute for Anthropomatics, Karlsruhe Institute of Technology
joris.ijsselmuiden@iosb.fraunhofer.de, rainer.stiefelhagen@kit.edu

Abstract. Most approaches to the visual perception of humans do not include high-level activity recognitition. This paper presents a system that fuses and interprets the outputs of several computer vision components as well as speech recognition to obtain a high-level understanding of the perceived scene. Our laboratory for investigating new ways of human-machine interaction and teamwork support, is equipped with an assemblage of cameras, some close-talking microphones, and a videowall as main interaction device. Here, we develop state of the art real-time computer vision systems to track and identify users, and estimate their visual focus of attention and gesture activity. We also monitor the users' speech activity in real time. This paper explains our approach to high-level activity recognition based on these perceptual components and a temporal logic engine.

Keywords: activity recognition, computer vision, temporal logic.

1 Introduction

There has been much progress lately in the visual perception of humans and their interaction with other people as well as with machines [1]. Most studies in this area however, focus on sub-problems like tracking or gesture recognition. What is often missing, is a second layer that spans a variety of perceptual components and fuses and interprets their outputs.

Our goal is to develop a framework for high-level human activity recognition. To achieve this, we fuse the available outputs from perception, cast them into a temporal framework, and use a logic engine containing context knowledge to deduce high-level facts. Regardless of the sensor setup or application domain, this framework should provide an abstract understanding of the given scene. For example, we can detect a group meeting or two people working together at a display. This can be used to adapt user interfaces accordingly, to automatically generate reports and visualizations, and to provide perceptual components with top-down knowledge. Application domains are: ambient intelligence and smart environments [2], but also robotics, surveillance and videosearch.

R. Dillmann et al. (Eds.): KI 2010, LNAI 6359, pp. 426–435, 2010.

This work is part of a larger project where computer vision and a wide array of other techniques are used to develop alternatives to the traditional mouse and keyboard controlled GUI. We are working towards real-world applications in control rooms for fire brigade, police, medical services, technical relief, military, and private security firms [3]. Besides for applications in human-machine interaction, we use computer vision, speech recognition, and high-level activity recognition, to automatically generate reports and visualizations.

The paper is organized as follows. In Section 2, we discuss some related work in activity recognition. Section 3 briefly explains the computer vision and speech recognition systems we use. Then, Section 4 describes our approach to multimodal fusion and activity recognition on top of these perceptual components. First experiments are presented in Section 5. And Section 6 provides the conclusion and some thoughts on future work.

2 Related Work

As the performance of computer vision systems increases, the interest in multimodal fusion and high-level activity recognition increases with it. A survey on existing approaches to activity recognition can be found in [4]. And [5] describes the state of the art in multimodal fusion for human-computer interaction.

Our methods for multimodal fusion and activity recognition were inspired by [6], where a temporal framework is used to define composite interactions between people in terms of atomic actions. Typical interactions they want to detect are: fighting, greeting, assault, and pursuit. Much like ourselves, they take advantage from the fact that perception is only concerned with learning and detecting the atomic actions. Complex compositions are provided by the layer on top. They also propose a probabilistic reasoning component that solves for missing and superfluous atomic actions.

In [7], instead of cameras and microphones, they use wearable motion and RFID sensors for recognizing activities of daily living. Nonetheless, we were able to draw inspiration from their work, since we strive for a general purpose framework, independent of sensor setup and application domain. They use an emerging patterns approach and sliding time windows to classify sequential, interleaved, and concurrent activities. A more classical approach is [8], where the perceptual outputs are fed to Hidden Markov Models after some preprocessing. Here, speech detection, ambient sound detection, tracking, and posture estimation are used to classify social activities in an ambient intelligence setting.

In [9], Fuzzy Metric-Temporal Horn Logic is used to generate natural language descriptions from vehicle trajectories. A similar approach, applied to human movement patterns, is presented in [10]. The work presented in [12] is concerned with storyline extraction from sports videos using and-or graph representations. This approach overcomes some of the limitations of Hidden Markov Models and Dynamic Bayesian Networks, because not only the model parameters are learned, but the model structures too. In [11], and-or graphs are used to generate text descriptions for a large dataset containing many different types of videos and

images. Note that these last four studies are not just concerned with classifying activities. They also generate corresponding reports in natural language.

The novelty and contribution of our work lies in the fact that we fuse and interpret a large variety of real-time perceptual components in a single model. Also, we use a high-level temporal logic based approach as opposed to low-level statistical methods. More novelty is provided by our application domain: human-machine interaction and teamwork support in crisis response control rooms.

3 Perception

Our experiments take place in a laboratory of six by nine meters, equiped with eleven cameras. Four are located in the room's upper corners and one fish-eye camera is mounted at the ceiling's center. Another four look down from the ceiling onto the area in front of our videowall. And two active pan-tilt-zoom cameras are mounted on the walls at head height. Each computer vision component uses a specific subset of these. For speech recognition, we use four close-talking microphones. And on the interaction side, a videowall serves as the prime user interface. All components described in this paper run on eight off-the-shelf PCs, connected through a dedicated LAN middleware. We currently use four perceptual components: tracks and identities, visual focus of attention, gestures, and speech (see Figure 1). They provide the following information: who are present in the room, where are they located, what are they looking at, what gestures are they performing, and what are they saying. To obtain a complete scene understanding, the perceptual information is supplemented by information that is not grounded in perception: what objects are in the room, where are they located, and what is currently being displayed on the room's user interfaces.

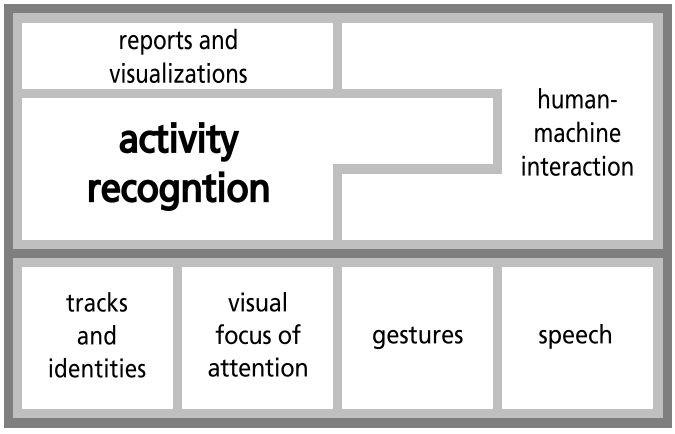

Fig. 1. Schematic representation of the system's components. The bottom layer provides the top layer with perceptual information about the people in the room.

Tracks and Identities. Users' locations and identities are arguably the most fundamental aspects of the observed scene. Tracking is performed using a multi-level particle filter approach on the four corner cameras and the central fish-eye camera [13]. For face identification, we use the two pan-tilt-zoom cameras on the walls and a DCT-based local appearance model [14].

Visual Focus of Attention. The person or object being looked at is another important factor in the perception of human interaction. Our laboratory's unconstrained camera placement yields low resolution images that put strong limitations on the perception of peoples' eyegaze, so we make approximations using head pose angles. The four corner cameras are used to generate head appearance hypotheses (i.e. head positions, sizes, and angles), and a combined particle filter framework rates the hypotheses with local shape descriptors and artificial neural networks. The successive deduction of a person's visual focus of attention (person or object) is then obtained by exploiting a linear relationship between measured head pose and actual gaze angles [15].

Gestures. Teamwork at large videowalls plays an important role in control room design [3]. Touch-sensitive surfaces seem to offer a natural way of interaction here. However, touch alone is not sufficient, because it forces users to walk along the videowall, and objects at the top can simply be out of reach. We overcome this limitation by adding the possibility of pointing at the videowall from a distance. The four cameras around the videowall are used for 3D body reconstruction through a visual hull approach implemented on the GPU [16]. The whole process is based on video cameras alone and it does not require a special surface.

Speech. Close-talking microphones and state of the art speech recognition software [17] provide us with an extra modality for activity recognition and human-machine interaction. We can currently use speech recognition combined with the pointing gestures described above to add tactical symbols to a digital map. And we can monitor who is speaking and who is silent for up to four users.

4 Multimodal Fusion and Activity Recognition

We use the perceptual information described above, knowledge from the domain of human interaction, and a temporal logic engine, to fuse the different sources and deduce high-level facts in real time. In other words, tracks, identities, visual focus of attention, gestures, and speech, but also information about furniture, user interfaces, and other objects, are fused and analysed to build an abstract model of the situation in the room. While keeping in mind our application domain [3], we strive for general purpose solutions to high-level activity recognition. Deduced high-level facts can be used to adapt user interfaces accordingly, to automatically generate reports and visualizations, and to provide the perceptual layer with top-down knowledge. Typical situations we want to detect are: two

people having a conversation, two people working together at the videowall, a group meeting, other subgroup-constellations, and users' roles in control room settings.

The perception layer also controls human-machine interaction directly, without a component for multimodal fusion and activity recognition in between. For example, we use tracks and identities to display user specific information close to the corresponding user. And a mixture of hand gestures and speech commands can manipulate objects on the videowall. In such cases, the human-machine interaction component performs its own limited fusion and interpretation. For handling complex situations however, a dedicated component for multimodal fusion and activity recognition is essential.

4.1 Model Taxonomy

The first step is to use the LAN middleware to subscribe to all the output streams generated by the perception layer. Each message received in this manner triggers the proper events as follows. During initialization, a scenario object is constructed containing a situation object for time $t = 0$, representing an empty room. Then, events are triggered that correspond to objects being put into the room, resulting in a situation object for $t = 1$. After initialization, a timestep and corresponding situation object is added to the scenario for each message batch received. Each situation is a copy of the last, augmented through events that are triggered by the incoming messages (see Figure 2).

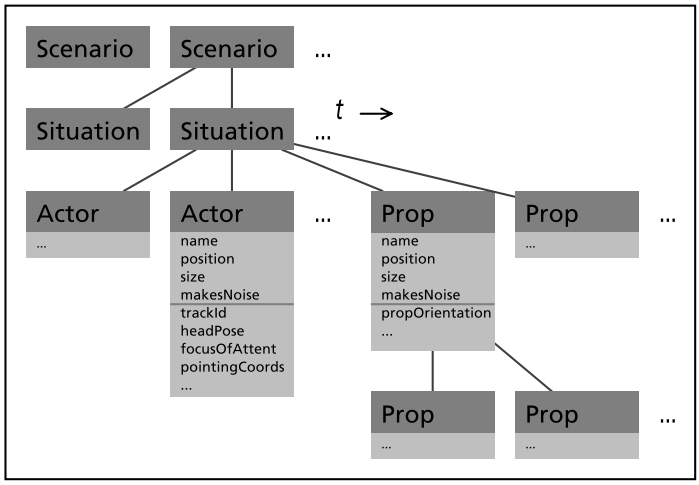

Fig. 2. Schematic representation of the model taxonomy. Actors represent the people in the room, and props can be user interfaces, furniture, or other objects. Situations are ordered in time and transitions between them are triggered by events in the perception layer. This provides the basis for temporal reasoning.

Each situation contains two types of entities: actors and props. The actor objects correspond to the people in the room. Currently, they contain identity, position, size, and knowledge about head pose, visual focus of attention, pointing gestures, and speech activity. All information needs to be assigned to the correct actor object, which requires context knowledge and reasoning. The props contained in each situation represent the objects in the room: furniture and their components, user interfaces and the elements displayed on them, and any other objects that might be there. Props' components are themselves props to allow for arbitrary complexity.

These descriptions of actors and props are used as facts by a logic engine composed of domain knowledge rules. It deduces high-level facts, both within each situation and over sequences of situations, which amounts to temporal reasoning. The logic engine is implemented using the Castor Logic Library, which allows for seamless multiparadigm programming. The power and flexibility of C++ and its imperative paradigm can be mixed at will with the reasoning abilities and convenience of the Castor Logic Library and its declarative, PROLOG-like paradigm [18]. Castor is a small and elegant header-only template library that is open source and without dependancies. Truth conditions for atomic predicates can be of arbitrary complexity, using the full C++ language. And the predicates formed like this are immediately available for spatial, temporal, and logical composition.

4.2 Temporal Logic

We adopt the temporal interval relations before, meets, overlaps, starts, during, finishes, and equals, as defined in [19]. Temporal composition allows us to classify situations that involve changing behavior over time, and it filters out observations that are too short-lived to be of interest. The following example represents our current definition of coordinated interaction, where two people are interacting with a videowall and a third person is seated at a table, giving them instructions (Figure 3, bottom row, second column). For the experiments described in Section 5, we implemented five such rules.

$$\begin{aligned}
&\forall\, x, y, z \;\; (TalksTo(x, y, t, u) \;\vee\; TalksTo(x, z, t, u)) \;\wedge \\
&CloseTo(x, table, t', u') \;\wedge\; CloseTo(y, videowall, t', u') \;\wedge \\
&CloseTo(z, videowall, t', u') \;\;\wedge\;\; y \neq z \;\;\wedge\;\; t \;\text{ DURING }\; t' \\
&\rightarrow\; CoordinatedInteraction(x, y, z, t)
\end{aligned}$$

$$\begin{aligned}
CloseTo(x, y, d, t, u) &\iff CloseTo(x, y, d) \text{ in at least } u \text{ timesteps of interval } t \\
CloseTo(x, y, d) &\iff \text{Distance between } x \text{ and } y \text{ is smaller than } d \\
TalksTo(x, y, t, u) &\iff Talks(x, t, u),\ LooksAt(x, y, t', u'), \text{ and } t \text{ DURING } t' \\
Talks(x, t, u) &\iff Talks(x) \text{ in at least } u \text{ timesteps of interval } t \\
LooksAt(x, y, t, u) &\iff LooksAt(x, y) \text{ in at least } u \text{ timesteps of interval } t \\
t \;\text{ DURING }\; t' &\iff t_{begin} > t'_{begin} \;\text{ and }\; t_{end} < t'_{end}
\end{aligned}$$

5 First Experiments

We performed five group activities that also occur in crisis response control rooms: individual work, table meeting, presentation, coordinated interaction (see Section 4.2), and standing meeting. We recorded each activity seperately for three minutes, and we recorded two ten-minute sequences, containing all five activities (see Figure 3). There were three actors and four props involved: one director, two other staff members, a videowall, a table, and two posters representing individual workstations.

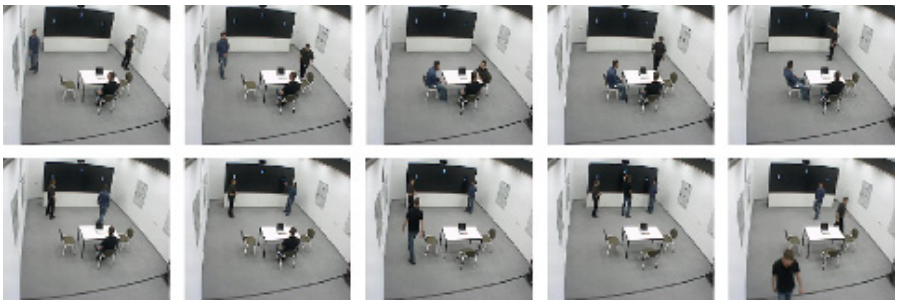

Fig. 3. Consecutively: individual work, transit 1, table meeting, transit 2, presentation, transit 3, coordinated interaction, transit 4, standing meeting, and transit 5

The actors were speaking and moving around freely and they pretended to interact with the videowall. Pointing activity however, was not used for the results presented in this paper. Also, we have only measured the head pose of the director for this experiment. Below are the five sentences generated upon detection, along with their truth conditions in simplified notation. Spatial, temporal, and logical composition for these activities is analogous to the example in Section 4.2. These rules would become more complex if we integrate gestures, head pose for all actors, and other perceptual components. Abbreviations are: dr = director, $s1$ = staff member 1, $s2$ = staff member 2, tb = table, vw = videowall, $p1$ = poster 1, and $p2$ = poster 2. Curly brackets signify disjunction.

- Director, S1, and S2 are doing individual work: $IndividualWork(dr, s1, s2)$
  if: $CloseTo(dr, tb), CloseTo(s1, p1), CloseTo(s2, p2)$
- Director, S1, and S2 are in a table meeting: $TableMeeting(dr, s1, s2)$
  if: $CloseTo(dr, tb), CloseTo(s1, tb), CloseTo(s2, tb), Talks(\{dr, s1, s2\})$
- S1 holds a presentation for Director and S2: $Presentation(s1, dr, s2)$
  if: $CloseTo(s1, vw), Cl.To(dr, tb), Cl.To(s2, tb), Talks(s1), LooksAt(dr, s1)$
- S1 and S2 are interacting under Director's supervision: $Coord.Int.(s1, s2, dr)$
  if: $CloseTo(s1, vw), CloseTo(s2, vw), CloseTo(dr, tb), TalksTo(dr, \{s1, s2\})$
- Director, S1, and S2 are in a standing meeting: $StandingMeeting(dr, s1, s2)$
  if: $CloseTo(dr, vw), CloseTo(s1, vw), CloseTo(s2, vw), Talks(\{dr, s1, s2\})$

Classifying these five activities is not a hard problem. In fact, one can make the correct classifications using only tracking information. But to show the potential of the presented system, the conditions for each activity were purposefully made harder to fulfill. For example, we set the constraint that three people sitting at a table do not form a meeting yet if nobody is speaking. Under the listed truth conditions, and with empirically chosen parameter values for d, t, and u, the system achieves reasonable classification results. A perfect classification of the five isolated three-minute recordings would be achieved if *only* the correct activity is detected throughout the entire corresponding recording. Over all five three-minute recordings, corresponding to the five activities listed above, we achieve an average precision score of 0.744, and an average recall score of 0.738. Performance on the two mixed recordings is similar.

False positives and false negatives can have several causes. First, the recordings were not annotated. We had to assume as ground truth that each activity was performed non-stop, throughout the corresponding recording. Second, the output from the perceptual layer is not always flawless. And third, the threshold d for $CloseTo(x, y, d)$, was set to $2.5m$ to make the tracker less powerful as a classifier. This had the effect that actors were often close to two props simultaneously, making two activities true at the same time, and thus decreasing the precision score without a notable increase in recall. Also note that t and u were given high values so that atomic predicates had to be true for a considerable amount of time before their temporal counterparts become true. This had the desirable effect of an increase in precision without a notable decrease in recall. However, if t and u were set too high, search spaces became too large for real-time operation.

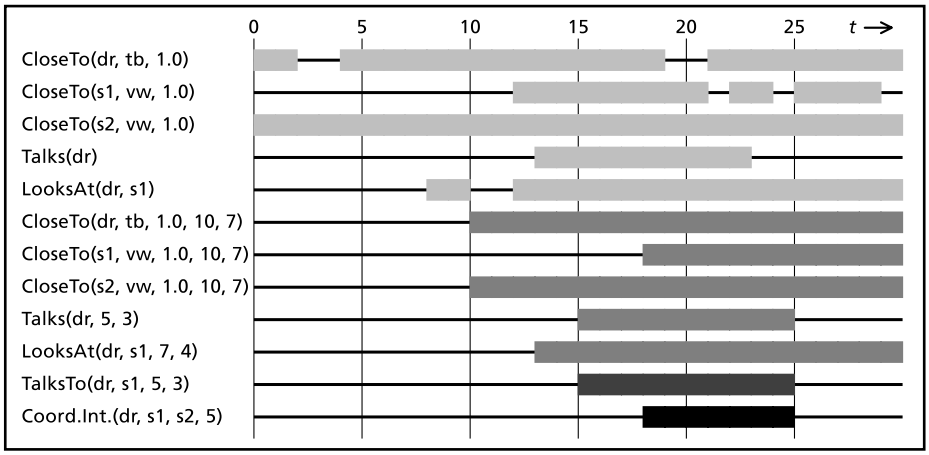

Fig. 4. A detailed view of the system's output. The 30 timesteps on the time axis correspond to three seconds of recorded perceptual information (see Figure 3 and Section 4.2). Abbreviations are: dr = director, $s1$ = staff member 1, $s2$ = staff member 2, tb = table, vw = videowall, $p1$ = poster 1, and $p2$ = poster 2

Figure 4 provides a detailed view of the system's output. The 30 timesteps that are displayed here, correspond to a three second interval from the coordinated interaction recording. The graph illustrates how the spatial, temporal, and logical composition from Section 4.2 reacts to a sequence of real perceptual information. Although we used an interval from the middle part of the recording, the system was started at $t = 0$, having no knowledge about what was perceived before. This is why none of the temporal predicates are true during the first ten timesteps in Figure 4. For the sake of clarity, we chose lower values for d, t, and u during the generation of this graph: $d = 1.0m$, $t = 10$ timesteps, and $u = 7$ timesteps for $CloseTo(x, y, d, t, u)$ and even smaller values for t and u in other predicates. For practical applications, behaviors that spread over longer time intervals are of course more interesting. Also note that $t = 10$ is a short way of saying that t is an interval containing ten timesteps.

6 Conclusion and Future Work

In this paper, we presented our progress towards a framework for high-level human activity recognition. Regardless of the sensor setup or application domain, the presented system fuses the outputs of several perceptual components, casts them into a temporal framework, and uses a logic engine containing context knowledge to deduce high-level facts, thus providing an abstract understanding of the given scene. Application domains include ambient intelligence, smart environments, robotics, surveillance, and videosearch. The novelty and contribution of our work lies in the fusion and interpretation of a large variety of real-time perceptual components into a single model. And in the use of a high-level temporal logic based approach as opposed to low-level statistical methods. More novelty comes from our current application domain: new tools for human-machine interaction and teamwork support in crisis response control rooms.

Future versions of the presented system will deal with more complex and detailed scenarios. Our focus in the future will lie on thorough empirical evaluations using videos with ground truth annotations and systematic evaluation metrics. We will benefit from the fact that the underlying perceptual components are constantly being improved and extended. For example, articulated body models and detailed knowledge of display activity will be available in the near future. We are also looking into the integration of more large screens, traditional workstations, table-displays, tablets, handhelds, and speakers. In parallel to these developments, the amount of possible facts and rules will increase. Also, we are envestigating the integration of alternative methods to obtain more subtle classifications, such as n-valued or fuzzy logic, possible worlds models, default logic, parameter evolution, and statistical learning. Finally, we will pursue the generation of natural language reports and 3D scene visualizations, as well as alternative application domains.

Acknowledgments. This work is supported by the FhG Internal Programs under Grant No. 692 026.

References

1. Waibel, A., Stiefelhagen, R. (eds.): Computers in the Human Interaction Loop. Springer, London (2010)
2. Nakashima, H., Aghajan, H., Augusto, J.C. (eds.): Handbook of Ambient Intelligence and Smart Environments. Springer, New York (2010)
3. Ivergard, T., Hunt, B.: Handbook of Control Room Design and Ergonomics: A Perspective for the Future, 2nd edn. CRC Press, London (2008)
4. Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine Recognition of Human Activities: A Survey. Circ. Syst. Vid. Techn. 18(11), 1473–1488 (2008)
5. Thiran, J.-P., Marques, F., Bourlard, H. (eds.): Multimodal Signal Processing, Theory and Applications for Human-Computer Interaction. Academic P., Oxford (2010)
6. Ryoo, M.S., Aggarwal, J.K.: Semantic Representation and Recognition of Continued and Recursive Human Activities. Int. Jour. of Computer Vision 82, 1–24 (2009)
7. Gu, T., Wu, Z., Tao, X., Pung, H.K., Lu, J.: epSICAR: An Emerging Patterns based Approach to Sequential, Interleaved and Concurrent Activity Recognition. In: 7th Conf. on Pervasive Computing and Communications. IEEE P., New York (2009)
8. Brdiczka, O., Langet, M., Maisonnasse, J., Crowley, J.L.: Detecting Human Behavior Models from Multimodal Observation in a Smart Home. IEEE T. Automation Science and Engineering 6(4), 588–597 (2009)
9. Gerber, R., Nagel, H.-H.: Representation of Occurrences for Road Vehicle Traffic. Artificial Intelligence 172, 351–391 (2008)
10. Gonzalez, J., Rowe, D., Varona, J., Xavier Roca, F.: Understanding dynamic scenes based on human sequence evaluation. Im. Vis. Comput. 27(10), 1433–1444 (2009)
11. Yao, B.Z., Yang, X., Lin, L., Lee, M.W., Zhu, S.-C.: I2T: Image Parsing to Text Description. Proceedings of the IEEE 99, 1–24 (2010)
12. Gupta, A., Srinivasan, P., Shi, J., Davis, L.S.: Understanding Videos, Constructing Plots; Learning a Visually Grounded Storyline Model from Annotated Videos. In: Conf. on Computer Vision and Pattern Recog., pp. 2004–2011. IEEE P., New York (2009)
13. Bernardin, K., Gehrig, T., Stiefelhagen, R.: Multi-Level Particle Filter Fusion of Features and Cues for Audio-Visual Person Tracking. In: Stiefelhagen, R., Bowers, R., Fiscus, J.G. (eds.) RT 2007 and CLEAR 2007. LNCS, vol. 4625, pp. 70–81. Springer, Heidelberg (2008)
14. Ekenel, H.K., Jin, Q., Fischer, M., Stiefelhagen, R.: ISL Person Identification Systems in the CLEAR 2007 Evaluations. In: Stiefelhagen, R., Bowers, R., Fiscus, J.G. (eds.) RT 2007 and CLEAR 2007. LNCS, vol. 4625, pp. 256–265. Springer, Heidelberg (2008)
15. Voit, M., Stiefelhagen, R.: Deducing the Visual Focus of Attention from Head Pose Estimation in Dynamic Multi-view Meeting Scenarios. In: 10th International Conference on Multimodal Interfaces, pp. 173–180. ACM Press, New York (2008)
16. Schick, A., van de Camp, F., Ijsselmuiden, J., Stiefelhagen, R.: Extending Touch: Towards Interaction with Large-Scale Surfaces. In: Interactive Tabletops and Surfaces 2009, pp. 127–134. ACM Press, New York (2009)
17. Soltau, H., Metze, F., Fugen, C., Waibel, A.: A One-pass Decoder Based on Polymorphic Linguistic Context Assignment. In: 2001 Automatic Speech Recognition and Understanding Workshop, pp. 214–217. IEEE Press, New York (2001)
18. Naik, R.: Blending the Logic Paradigm into C++ (2008), `http://mpprogramming.com`
19. Allen, J.F., Ferguson, G.: Actions and Events in Interval Temporal Logic. Journal of Logic and Computation 4(5), 531–579 (1994)