11

Activity Classification

Kai Nickel[1], Montse Pardàs[2], Rainer Stiefelhagen[1], Cristian Canton[2], José Luis Landabaso[2], Josep R. Casas[2]

[1] Universität Karlsruhe (TH), Interactive Systems Labs, Fakultät für Informatik, Karlsruhe, Germany
[2] Universitat Politècnica de Catalunya, Barcelona, Spain

When a person enters a room, he or she immediately develops a mental concept about "what is going on" in the room; for example, people may be working in the room, people may be engaged in a conversation, or the room may be empty. The CHIL services depend on just the same kind of semantic description, which is termed *activity* in the following. The "Connector" or the "Memory Jog", for example, could provide support that is appropriate for the given context if it knew about the current activity at the user's place. This kind of higher-level understanding of human interaction processes could then be used, e.g., for rating the user's current availability in a certain situation.

The recognition of activities depends on many factors such as the location and number of people, speech activity, and the location and state of certain objects. The perceptual technologies like person tracking, identification, or acoustic event detection provide important information upon which the higher-level analysis of the activity can be based. Due to the complexity of the scene, there are, however, potentially relevant phenomena such as a door being half-opened, which – due to their high number and variability – cannot be addressed by manually designed detectors at large. Therefore, activity recognition may need to directly analyze the observation in order to find out what is relevant and what is not to detect a certain activity.

Activity recognition may be facilitated by the detection of *events*, which are semantic descriptions for actions that have a short duration and/or are limited to a small area of the room, such as "a person enters/leaves the room". In this case, activities are being recognized by their characteristic sequence of events.

This chapter describes three systems that have been implemented for (1) the recognition of events inside the CHIL room based on person tracking and a probabilistic syntactic approach, (2) person activity classification using body gestures, and (3) event recognition and room-level tracking in multiple office rooms based on low-level audiovisual features.

11.1 Visual Activities Recognition in a Smart-Room Environment Using a Probabilistic Syntactic Approach

In a smart-room environment, it is necessary to determine the type of interactions among the people in the room in order to perform context-dependent actions. Moreover, some specific situations need to be identified by the system. For this aim, a system based on [6], which analyzes activities using stochastic parsing, has been developed by adapting it to the specific requirements of smart-room environments. The fundamental idea of [6] is to divide the recognition problem into two levels. The lower-level detections are performed using standard independent probabilistic event detectors to propose candidate detections of low-level features. The outputs of these detectors provide the input stream for a stochastic, context-free, grammar-parsing mechanism. The system can be divided into three modules: the tracking system, the events generator, and the parser. The tracking system [7] takes as input the multicamera video sequence, reconstructs the 3D objects in the room using a foreground detection for each camera, and performs the tracking of the various detected objects. Thus, for each frame in the video sequence, we require N views from the calibrated cameras. Foreground regions are obtained for each camera using an algorithm based on Stauffer and Grimson's background learning and subtraction technique [12]. A Shape from Silhouette procedure is used next in order to generate a discrete occupancy representation of the 3D space (voxels) to decide whether a voxel is in the foreground or background by checking the spatial consistency of the N segmented silhouettes. Afterwards, a connectivity filter is introduced in order to remove isolated voxels and the remaining multiple RoIs are labeled in accordance with the results of a tracking procedure, as described in Chapter 3. The events generator and the parser are described in the following.

11.1.1 Events Generation

The main objective of the events generator is to provide the chain of events that the parser will take as input. The inputs to the events generator are the ones delivered by the tracking system: object identifier, number of frame, position (x, y, z), velocity, and volume of each object. Moreover, the output of the multicamera 3D person and object tracker is enriched by (1) an algorithm that is able to distinguish between an object and a person – assuming an average range of physical properties of adult humans – and (2) an algorithm that analyzes human body posture (standing, sitting, etc.) with a standard model of the human body that is aligned to the 3D regions of interest earlier classified as a person. This information is used together with some configuration information about the room (table, chair, and whiteboard position, dimensions of the room) to produce the events detection using simple grammars. The list of events detected by this module is the following: {"Person enters in the room", "Person exits from the room", "A person is lost in the room", "A person is found in the room", "A person sits down", "A person moves inside the room", "A person stops", "A person stands up", "A person is detected in the whiteboard area", "A new object is detected in the room", "An object disappears from the room", "The volume of

a person increases", "The volume of a person decreases", "A group of people is detected", "A person is divided in two"}.

11.1.2 Video Activity Recognition

The video activity recognition is performed by the parser. Its function is as follows: Given a chain of events and a stochastic, context-free grammar, find the chain derivation with the maximum probability, if it exists. The parser we have used is based on the CYK algorithm [14]. It performs an ascendant analysis, considering subtrees from the leaves up to the root. The activities the system currently recognizes are the following:

- meeting,
- presentation,
- conversation between two people,
- leave an object ,
- take an object.

The video sequence is first analyzed with the tracking and event generators system. The generated chain of events is input to the parser, which analyzes the chain using the grammar corresponding to these classes. In order to generate more appropriate grammars, we have implemented a training system to generate the grammars for the different classes we want to learn. The Inside-Outside algorithm [8] is used to estimate the probability of the production rules of the stochastic, context-free grammars using the video sequences created for training. The production rules have been manually designed.

11.1.3 Experiments

To test the system, we have used 50 recordings where the five defined activities occur. The recordings have been done in a smart room with four fixed cameras in the corners plus a zenithal camera. An example of four frames corresponding to a "presentation" recording is shown in Fig. 11.1, together with a projection of the reconstructed blobs. The recognition results span from 60% recognition for the activities "conversation" and "take an object" to 87.5% for "presentation", with a mean correct recognition rate of 70%.

11.2 Person Activity Classification Using Gestures

Some activities of the persons in the room cannot be recognized using only the person tracking results and the 3D reconstructed objects. Human motion descriptors add the necessary information for classifying person activities that involve the motion of the body limbs. We have developed a view-independent approach to the recognition of human gestures of several people in low-resolution sequences from multiple

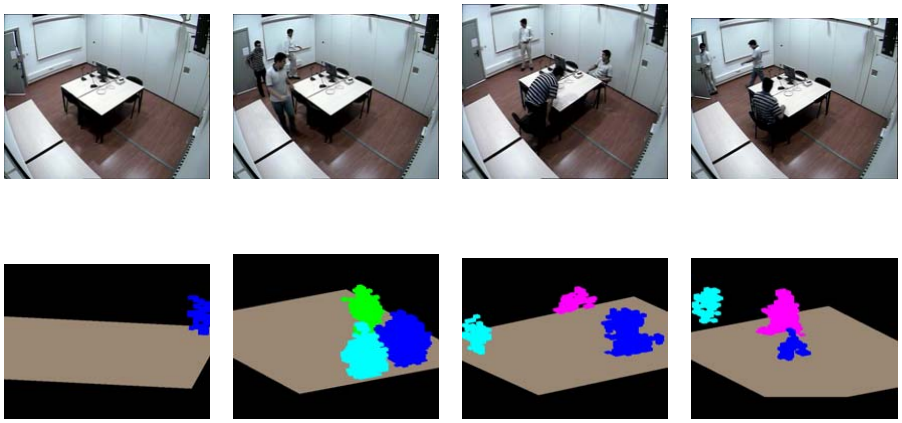


Fig. 11.1. Four frames from a presentation, showing different events that can be detected using the tracks. The first row shows the images from one camera and the second row a projection of the 3D detected blobs. The event detected in the images of the first and second columns is "person enters". The third column corresponds to "person at whiteboard" and "person sits", while in the image of the last column, a "person exits" is detected.

calibrated cameras [3]. In contrast to other multi-ocular gesture recognition systems based on generating a classification on a fusion of features coming from different views, our system performs a data fusion (3D representation of the scene) and then a feature extraction and classification. Motion descriptors introduced by Bobick and Davis. [1] for 2D data are extended to 3D and a set of features based on 3D invariant statistical moments is computed. A simple ellipsoid body model is fit to incoming 3D data to capture in which body part the gesture occurs, thus increasing the recognition ratio of the overall system and generating a more informative classification output. Classification is thus performed by jointly analyzing the motion features and the body position data obtained by fitting the ellipsoid body model. Finally, a Bayesian classifier is employed to perform recognition over a small set of actions. The actions that are more relevant to the smart-room scenario are raising hand, sitting down, and standing up. However, we have tested the system including other actions such as waving hands, crouching down, punching, kicking, and jumping. The approach taken relies on the 3D reconstruction of the detected persons. Thus, the system uses as input the same data described in Section 11.1, that is, the multiple RoIs labeled coherently along time, corresponding to the persons in the room. In the following, we describe the approach taken to analyze the person's activity using these input data.

11.2.1 Motion and Body Analysis

In order to achieve a simple and efficient low-level, view-dependent motion representation, [1] introduced the concept of motion history image (MHI) and motion energy image (MEI). We extended this formulation to represent view-independent 3D

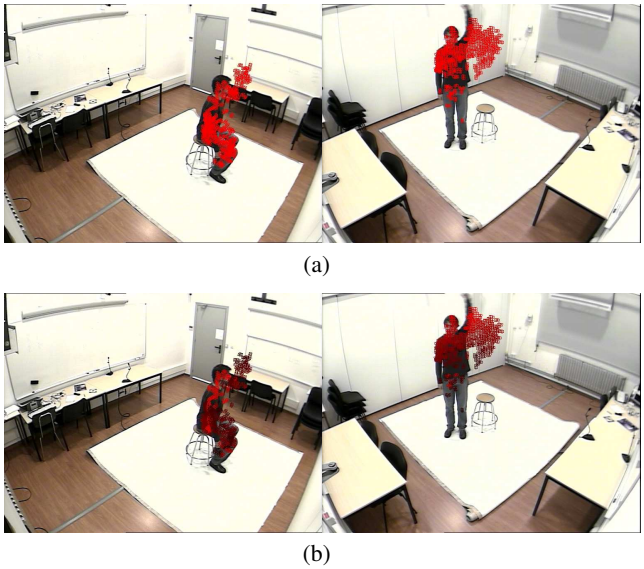


Fig. 11.2. Example of motion descriptors. In (a) and (b) are depicted the 2D projections of MEV and MHV, respectively, for *sitting down* and *raising hand*.

motion. In this way, ambiguities generated by occlusions are overcome. Analogously to [1, 2], the binary motion energy volume (MEV) $E_\tau(\mathbf{x},t)$ captures the 3D locations, where there is motion in the last τ frames. Motion detection can be coarsely estimated by a simple forward differentiation among voxel frames, still leading to satisfactory results while preserving a reduced computational complexity. Figure 11.2(a) depicts an example of MEV.

To represent the temporal evolution of the motion, we define the motion history volume (MHV) $H_\tau(\mathbf{x},t)$, where each voxel intensity is a function of the temporal history of the motion at that 3D location. An example of MHV is shown in Fig. 11.2(b).

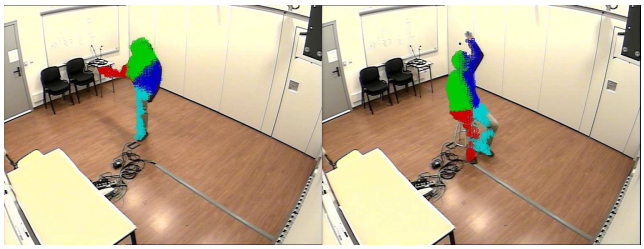
In order to extract a set of features describing the body of a person performing an action, a geometrical configuration of the human body must be considered. An ellipsoid model of the human body has been adopted and, in spite of this fairly simple approximation compared with more complex human body models, classification results proved the validity of our assumption, as shown in Section 11.2.3.

After obtaining the set of voxels describing a given person, we fit an ellipsoid shell to model it. This information is then fed to a body-tracking module that refines this estimation by taking into account body anthropometric restrictions, imposing some motion and size constraints compatible with human bodies [4]. For example, the height of a person restricts the possible locations of arms and legs according to the average lengths of body parts. Finally, time consistency of the ellipsoid parameters is achieved by a Kalman filter.

Once the parameters of the ellipsoid representing the human body are computed, a simple body part classification can be derived. Voxels can be labeled as belonging

to four categories: left/right arm/leg (see Fig. 11.3). These data will be used while classifying an action jointly with motion information.

(a)



(b)

Fig. 11.3. Body analysis module output. In (a), original images for *kick* and *arise hand*. In (b), voxels belonging to the body of the person are labeled as belonging to right/left arm/leg categories.

11.2.2 Feature Extraction and Gesture Classification

Data produced by the motion and body analysis modules are processed to extract a vector of features for classification.

Informative features derived from the analyzed data (MHV and MEV in our case) are required to represent motion in a low-dimensional space. Statistical moments invariant to scaling, translation, rotation, and affine mappings were introduced by [5]. Three-dimensional invariant statistical moments [9] were used in our case. For each data set $E_\tau(\mathbf{x},t)$ and $H_\tau(\mathbf{x},t)$, two invariant moment-based feature vectors were computed, ψ_{MEV} and ψ_{MHV}, each comprising five components.

Information from body parts provided by the body analysis module can be used to generate additional features. Let ψ_{BODY} denote the four features describing the relative amount of motion voxels located in each body part. Given the computed moment-based motion features and the body features obtained for each of the actions to classify ω_j, $0 \leq j < K$, we define a full 14-dimensional feature vector $\Gamma = [\psi_{\mathrm{MEV}}\, \psi_{\mathrm{MHV}}\, \psi_{\mathrm{BODY}}]$. The dimensionality of Γ can be further reduced through principal components analysis (PCA). By analyzing the training data, we noticed

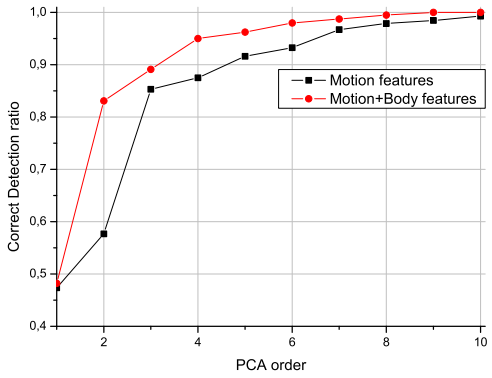


Fig. 11.4. Classifier performance evaluated with motion and body features depending on the order of the PCA analysis.

that 90% of the variance of the data was achieved by doing a dimension reduction to $d = 7$. Let us refer to the data set obtained after PCA analysis as $\hat{\Gamma}$.

The classification method is based on a Bayesian classification criterion assuming that $p(\hat{\Gamma}|\omega_j)$ is normally distributed and estimating the mean and covariance matrix of each class with the training data.

11.2.3 Experiments

In order to evaluate the performance of the proposed algorithm, we collected a set of 70 training and 30 testing multiview sequences of each action to be recognized. The gesture category set was formed by eight common actions of interest in the field of human-computer interfaces such as raising hand, sitting down, waving hands, crouching down, standing up, punching, kicking, or jumping. Moreover, to show the effectiveness of our method and its robustness against rotations, occlusions, and position, actions were recorded in different positions inside the room and facing various orientations.

In average, we got a $p(\text{error}) = 0.0154$. Experiments have been carried out with and without these features to show the influence of body part features on the overall performance. Figure 11.4 depicts the behavior of the classifier for diverse orders of the PCA analysis showing that body features increase the performance of the system. The experimental results prove the efficiency of our method, proposing an alternative to the classical methodology to multi-ocular and mono-ocular motion-based gesture analysis [1, 11, 2].

11.3 Activity Recognition and Room-Level Tracking in an Office Environment

The previously described approaches use the output from a person tracker to infer people's activities. In this work [13], we bypass the tracker and try to infer human activity directly from the camera image. This is motivated by the fact that person tracking is computationally expensive, requires a rich sensor setup, and is still not 100% reliable. Furthermore, the recognition of human activity may not depend only on the location and pose of the human body, but also on the state of objects like doors or chairs. Therefore, we are following an appearance-based approach and develop an activity recognition system that operates directly on the data from a single fixed camera and a single microphone per room.

We decompose activities in two classes, namely events and situations, both carrying a semantic meaning. In our case, events are defined to be visible or audible short-term phenomena that are spatially limited to a small area. In the presented application, we detect events like PERSON SITTING AT A DESK or PERSON ENTERING/LEAVING AN OFFICE; i.e., we focus on events that are triggered by humans. In contrast to events, we define situations to range over a longer period of time and space. Situations that are to be distinguished by our system span the entire room: MEETINGS, DISCUSSIONS, PAPERWORK, PHONE CALLS, or NOBODY PRESENT. They were chosen manually by observing the recorded data. The objective was to cover a maximum share of daily office activities in a real-world setting.

The experimental activity recognition system spans four office rooms, each occupied by one or two members of the lab, as well as the local lab room (see Fig. 11.5). Each room is equipped with a sparse sensor setup consisting of a single camera and one omnidirectional microphone.

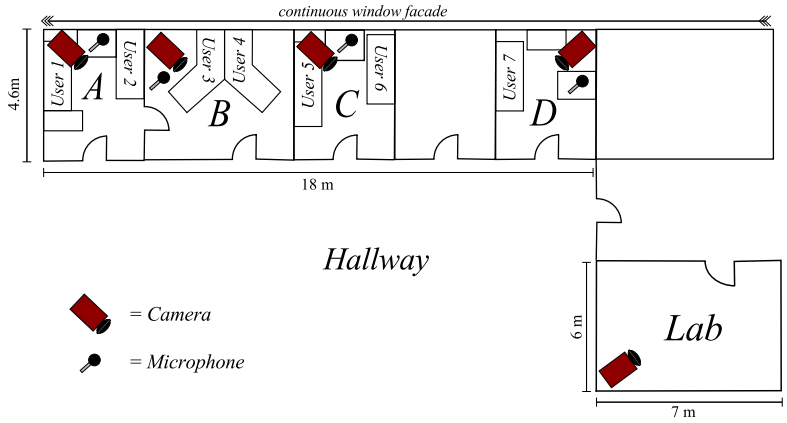


Fig. 11.5. Plan view of the rooms that were monitored for activity recognition.

Using only simple low-level features such as motion and optical flow for the video modality and signal energy, zero crossing rate, and pitch for audio, we employ a multilevel HMM activity recognition framework (see Fig. 11.6). Decomposing the parameter space into several layers reduces the amount of training data required and gives a better intuition on the learning process [10, 15]. The lower level detects events and passes them on to the situation layer. The situation layer infers room situations based on the sequence of detected events. Each layer can be trained on labeled data on its own by employing the well-known Baum-Welch parameter estimation algorithm.

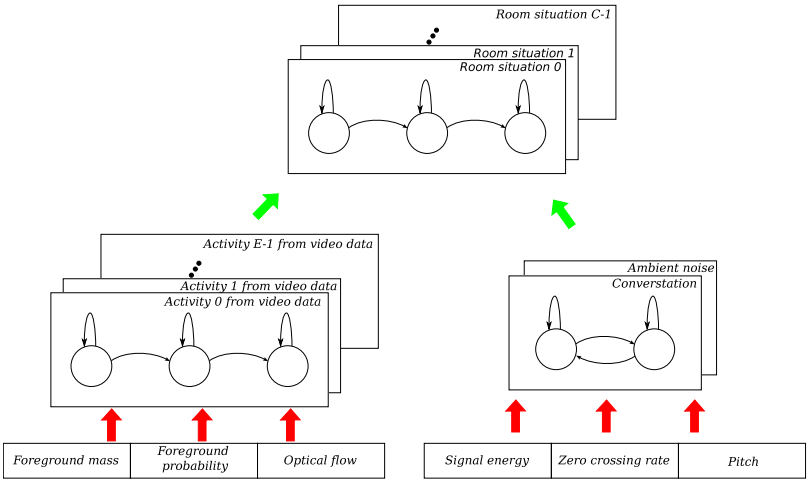


Fig. 11.6. Structure of the multilayer HMM for a single office. The lower level recognizes events, whereas the higher level represents *room situations*.

Knowing who is where on a room-level scale is the natural complement to situation recognition within the individual rooms. We track the users by inferring their locations from the sequence of events recognized by the first activity layer. We are thus exploiting common constraints like the fact that office workers tend to have a dedicated workplace that they use most of the time exclusively.

The method does not rely on conventional person identification or tracking techniques, which often pose restrictions for practical deployment due to their high sensory and computational requirements. Our approach exploits a Bayesian filter framework in a discrete state space, where the state vector contains the belief that a certain person is in a certain room or out of sight. The key problem of this multiperson tracking task is to assign observed events to the correct track, which is known as *data association* problem. A separate tracker is run for each person and data association is performed in two stages: A nearest-neighbor filter is applied to consider only the observations for the belief update that are occurring close to the highest belief

state. Moreover, the observation model is designed in a way that persons can only be observed at certain places depending on the room.

It is again important to note that this tracking approach does not use appearance models to identify the individual users directly, but instead aims to infer the user's locations only from the sequence of events that are recognized by the first layer of the activity recognition model. The output of the system is the room in which the person is currently located – and not its precise location within that room.

11.3.1 Features

Due to the changing light conditions, any illumination-dependent cue such as histogram backprojection to identify skin color would be error-prone. Moreover, persons are perceived from just one camera view. Depending on their orientation, we get either frontal, side, or back views of their head, so that face detectors can hardly be employed to determine the number of persons in the room. Therefore, we are concentrating on simple but fast video features that are robust against varying lighting conditions: adaptive background subtraction and optical flow.

On the audio side, we use speech activity detection, which is an important cue to determine people's current occupation. In our office scenario, it helps, for example, to separate visually similar classes like PAPERWORK and DISCUSSIONS. In order to detect speech activity, we calculate signal power, zero crossing rate, and pitch and process them with an audio classification HMM.

The key problem is to decide which regions of the input image are relevant for certain events. In our approach, we consider the relevant foreground regions of a certain activity to be the components of a Gaussian mixture. This allows a data-driven learning approach with the well-known EM algorithm. Features are then extracted from the enclosed areas of each mixture component within three standard deviations. Together these features are capable of describing a scene by the amount of motion with the dominant direction, while preserving rough location information.

Fig. 11.7. Three Gaussian mixture components obtained from data-driven clustering. They represent areas where users often sit.

11.3.2 Experiments

For the activity recognition part, we collected data from six work days with a total length of about 34.8 hours, of which we used four days for training and two days for evaluation. To obtain ground-truth labels, the data were annotated manually. On this data set, the recognition rate of the events ranged between 63% for "visitor behind user's desk" and 100% for "somebody enters". The situations were recognized with a recognition rate of $70-96\%$ depending on the type of situation; for details, see [13].

As this set of data contained only a few events of people changing offices, we recorded a second set with a scripted sequence of 44 events with a length of about one hour and we used it to evaluate room-level tracking. As the hallway was not monitored due to privacy reasons, blind gaps occurred between the cameras. On average, we could track the location of all seven people in 91.5% of the frames, and 36 of 44 transitions were correctly recognized. Table 11.1 shows ground-truth and tracking results for one of the tracked persons.

Ground Truth			**Tracking Results**		
Begin	*End*	*Place*	*Place*	*Begin*	*End*
0	32	Office B	Office B	0	41
46	68	Lab	Lab	49	75
77	800	Office B	Office B	81	809
809	1110	Office D	Office D	810	1102
1117	2194	Office B	Office B	1103	2197
2194	2484	Office A	Office A	2198	2496
2496	2660	Office D	Office D	2497	2671
2668	3248	Office B	Office B	2672	3263
3248	3389	Out of view	Out of view	3264	3382
3389	3685	Office B	Office B	3383	3687
3685	3699	Lab	Lab	3688	3705
3709	3719	Office A	Office D	3709	3925
3719	3926	Office B			

Table 11.1. Example trajectory for user #4 (times are given in seconds); for the sake of readability, OUT OF SIGHT is not listed for state durations of less than 30 seconds.

11.4 Conclusion

In the course of the CHIL project, three different approaches for automatic activity recognition have been implemented and evaluated. They are different in terms of the set of activities they classify, the features they use, and the actual classification method.

The first system is oriented to the recognition of room-level activities and thus uses only the detected volumes and the room configuration information as input.

It can detect interactions between people or between people and objects as well as classify the kind of activity that takes place in the smart room according to the predefined classes (meeting, conversation, etc).

The second system is oriented to the recognition of activities on a person level. For these kinds of activities, motion descriptors as well as a simple human body model are used, together with the foreground volumes. With the experiments carried out, we have concluded that a set of human body activities can be efficiently distinguished without requiring a complex human body model analysis that implies a high computational cost.

The third system works with a sparse sensor setup of one camera and one microphone per room. Events are detected based on a data-driven analysis of the sensor data. Based on the sequence of events, both office activities as well as the location of people on a room-level scale could then be inferred.

It is obviously hard, if not impossible, to find a common definition and methodology for activity recognition that fits all application domains. The current systems are dedicated to certain domains like office situations and meetings. They define a small set of activities that are specific and relevant within their domain, and they proved to be able to recognize the activities on in-domain test data. Future work, on the one hand, could try to extend the application domains, while, on the other hand, it could aim for a more detailed analysis of the activities within one domain.

One lesson learned in the CHIL project was that the development of an activity recognition system needs to be application-driven: The consumer of the information – for example, the Connector or the Memory Jog service – defines the domain and the set of meaningful activities due to its specific need. Only with that knowledge can an appropriate activity recognizer be designed.

References

1. A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.
2. G. Bradski and J. Davis. Motion segmentation and pose recognition with motion history gradients. *Machine Vision and Applications*, 13(3):174–184, 2002.
3. C. Canton-Ferrer, J. R. Casas, and M. Pardàs. Human model and motion based 3D action recognition in multiple view scenarios (invited paper). In *14th European Signal Processing Conference, EUSIPCO*, University of Pisa, Florence, Italy, 4–9 Sept. 2006.
4. S. Dockstader, M. Berg, and A. Tekalp. Stochastic kinematic modeling and feature extraction for gait analysis. *IEEE Transactions on Image Processing*, 12(8):962–976, 2003.
5. M. Hu. Visual pattern recognition by moment invariants. *IEEE Transactions on Information Theory*, 8(2):179–187, 1962.
6. Y. A. Ivanov and A. F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:852–872, 2000.
7. J. L. Landabaso and M. Pardas. Foreground regions extraction and characterization towards real-time object tracking. In *Machine Learning for Multimodal Interaction (MLMI)*, LNCS 3869, pages 241–249. Springer, 2006.

8. K. Lari and S. Young. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer, Speech and Language*, 4:35–56, 1990.
9. C. Lo and H. Don. 3-D oment forms: Their construction and application to object identification and positioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(10):1053–1064, 1989.
10. N. M. Oliver, B. Rosario, and A. Pentland. A Bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):831–843, 2000.
11. R. Rosales. Recognition of human action using moment-based features. *Boston University Computer Science Technical Report, BU*, pages 98–120, 1998.
12. C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.
13. C. Wojek, K. Nickel, and R. Stiefelhagen. Activity recognition and room level tracking in an office environment. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, Heidelberg, Germany, Sept. 2006.
14. D. H. Younger. Recognition and parsing of context-free languages in time n3. *Information and Control*, 10:189–208, 1967.
15. D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan. Semi-supervised adapted HMMs for unusual event detection. In *Computer Vision and Pattern Recognition*, pages 611–618, 2005.