in: Proc. 55th Int. Scientific Colloquium, Ilmenau, Germany, 2010, pp. 612-617, Verlag ISLE 2010

REAL-TIME ACTIVITY RECOGNITION ON A MOBILE COMPANION ROBOT

Michael Volkhardt, Steffen Müller, Christof Schröter, Horst-Michael Gross

Neuroinformatics and Cognitive Robotics Lab,
Ilmenau University of Technology, Germany

ABSTRACT

Recently, there has been an increasing research effort in supporting people by mobile robots in home environments. In this scope, activity recognition can tremendously enhance the social interaction skills of a robot by taking into account the user's state. Additionally, the system can adapt to the user's preferences and habits or detect deviations from daily routines. This paper presents a novel real-time activity recognition system on a mobile robot. The system continuously tracks the pose and motion of the user and combines them with structural knowledge like the current room or objects in proximity. All extracted features are modeled as probability distributions and processed by Bayesian Networks to reason about different activities. First experimental results on real data show the usefulness of our approach.

Index Terms— Activity Recognition, Companion Robot, Home Environment, Bayesian Networks

1. INTRODUCTION

For a long time, there has been a great interest in developing robots that support people in their daily routine and increase their quality of life. In contrast to static solutions like smart homes, mobile robots can increase usability by offering service where it is needed. Example features of an intelligent system include daytime management, video calls, monitoring, and natural human-machine interaction. In this scope, activity recognition can enhance the social behavior skills of a robot by taking in to account the individual preferences of the user. For instance, a person should not be disturbed by noncritical tasks, when s/he is resting or occupied. Furthermore, the system must recognize situations, that call for proactive reaction like coming home or leaving the home, incoming phone calls or critical situations. To enable the robot to reason about the user's current activity, it must extract features in their current context: a fallen person on the ground is much more alarming than a resting person on the couch – although in both cases the person is in a lying pose. To achieve this distinction, we track the person's position, pose, and motion by using range and image data and enrich these features with structural knowledge of the home environment (Fig. 1). Although this knowledge could be given by additional external information cues like infrared presence sensors or wall-mounted cameras, we seek for a solution that enables the robot to function autonomously in any home environment. Our system models all properties of the user and the environment as probability distributions to account for noisy or missing data.

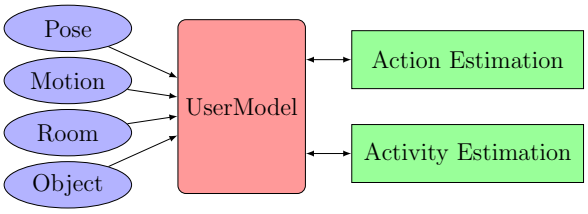

Fig. 1. *Overview of the proposed system. Key idea is the combination of visual features of the user and knowledge of the environment for action estimation. Activity recognition is based on the estimated actions.*

Multiple characteristic properties are combined to an action of short temporal duration. For instance, an action could be defined as a motion in the upper body while situated in the kitchen near the oven. An activity is then defined as a temporal sequence of multiple actions. The actions of the activity 'cooking' could vary in the upper body motion and the objects in proximity (oven, sink, table). By applying Bayesian Networks to labeled training data, the system learns the dependencies between different features, actions and activities. Once trained the Bayesian Models can be used to reason about the activity of the user in future situations.

Note that we are not aiming for long-term behavior tracking to detect deviations from daily routines. As a consequence, we can simplify the recognition by not detecting the starting point, duration or history of activities. Therefore, we focus on the estimation of those activities that reflect the current situation to improve human machine interaction. Challenges to this aim are introduced by the limited observability of the user, which

This work has received funding from the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 216487 (CompanionAble Project).
michael.volkhardt@tu-ilmenau.de

can easily leave the room or get occluded by furniture. It is also worth noting that the employed method must operate in real-time: Since the robot needs to interact and, therefore, react on a person's behavior, a retroactive analysis is not sufficient. Hence, the contribution of this paper is the development of a non-invasive real-time capable short-term activity recognition system on a mobile robot. The remainder of this paper is organized as follows. The next section presents methods closely related to our work. Sec. 3 describes our activity recognition system in more detail. In Sec. 4 we present first experimental results and Sec. 5 summarizes our conclusions.

2. RELATED WORK

Plenty of research has been done in the field of activity recognition in various applications (see [1] for a survey). Yet, very few approaches consider activity recognition on a mobile robot platform. Most solutions that reason about activities in home environments use multiple, heterogeneous sensors like light-, infrared- or pressure sensors, and static cameras to monitor the user – these installations are commonly known as smart homes. [2] fuse multiple sensor readings of a smart home to offer different services to the user – suitable to the current estimated activity. The system is updated on-line with very little labeling effort to account for changes in the environment or preferences of the user. [3] reasons about long-term daily activities like sleeping, eating, dressing up and detects abnormal behavior by incorporating the information of different infrared sensors. The recorded data is processed off-line while the user is sleeping. [4] includes the sensor information of a smart home while putting strong emphasis on interaction with objects in the apartment. For that purpose even objects like the toaster or knifes are augmented with sensors. In contrast, we are not able to recognize these fine-granular interactions with objects and cannot rely on sensory input from external sources. Other approaches like [5], [6] estimate activities through inertial sensors worn by the user. The recorded data from accelerometers attached to the hip, thigh, wrist and ankle allows to detect different activities like walking, running, lying-down or sitting. [7] use colored gloves and skin color tracking to determine the position of the head and hands and to recognize activities related to a dictionary related semantic.

Most of these aforementioned sensors are not available on our mobile platform. Generally, we seek for a non-invasive solution – not bothering the user to wear any devices – that requires no changes in the home environment. Although, our robot perceives range data, sounds, camera images, and context information of the environment, in this work we focus on the latter two input cues. Most approaches that use camera images rely on silhouettes of the user and recognize activities by extracting features from single instances or sequences [8], [9]. The extraction step usually depends on static cameras to apply background segmentation or image difference. [10] uses adaptive background subtraction to extract motion features and directionally based feature vectors from silhouettes. [11] analyzes temporal and spatial variations of activities by applying time warping transformations on silhouettes. Unfortunately, the moving robot does not allow to apply background subtraction or to extract the human contours.

Besides the comparison of applied sensors and extracted features, the methods can be divided by the used classification method. Different classifiers like MLPs [12], SVMs, or Bayesian Networks are evaluated in [13], which also inspects the optimal number of different features like pose, limbs and interactions with objects. Other approaches use classifiers which incorporate time, like HMMs or Hidden Conditional Random Fields [8], [9]. Although Bayesian Networks reach lower classification results than MLPs or SVMs in some scenarios [13], they are proven to be a powerful and easily expandable tool for activity recognition [4], [7], [11]. The activity recognition system used in this work is based on Bayesian Networks as well and described next.

3. ACTIVITY RECOGNITION SYSTEM

As stated before, we compose daily activities as a sequence of actions, which themselves are conditioned for different observable features. These features incorporate multiple characteristic properties of the user and the environment and describe the current situation. Because the measurements of the features are usually noisy, each observation is coded as a discrete probability distribution. Missing observations are accounted for by uniforming distributions over all possible realizations. We assume that all observations are conditioned by different actions of short temporal duration. To estimate the unobserved actions, we use Bayesian Networks that integrate the evidence given by the observations. After that, a temporal sequence of the estimated actions is built up. Finally, for each activity we evaluate the occurrences of certain representative actions. For example, the actions of the activity 'cooking' could vary in the upper body motion of the person and the objects in proximity (oven, sink, table), while the pose and room usually remain static.

An activity becomes active once their specific actions have a high probability in a characteristic time span assigned to the activity. In contrast to classifiers that explicitly model time like HMMs, this has the advantage that the order of actions is not relevant and duration of actions and activities may vary. Therefore, the order of the actions 'motion near cupboard', 'motion near the oven', 'upper body motion next to the table' may be switched for activity 'cooking'. Furthermore, the actions could appear multiple times. That means,

(a) room map (b) object map

Fig. 2. *(a) Top down view of 2D room labels overlaid with a 2D occupancy map. Red – robot, green – person hypothesis, yellow – living room, blue – kitchen, orange – bedroom, etc. (b) Overlaid object labels. Blue – sofa, orange – table, yellow – dishwasher, pink – cupboard.*

if an actions lasts longer, it might be detected in multiple time steps. Last but not least, the occurrences of actions could vary in a short or longer period of time as long as they are still in the time slot assigned to the activity. The following sections describe the components of the system in more detail.

3.1. Features

First, we track the position of the user by a multi-cue tracker based on the Kalman Filter [14]. The tracker applies a leg detector, a motion detection module, and the well-known AdaBoost face-detector [15]. Second, the height of the user's head relative to the floor is calculated by applying an upper body HOG detector in the bounding box of the user given by the tracker [16]. The pose of the user is then classified into 'standing' or 'sitting' by a simple height threshold. We are currently working on a more sophisticated HOG pose detector that is able to classify arbitrary user poses by applying a cascade of linear SVMs [17]. Furthermore, a module is in development to detect a fallen user on the ground via an intelligent floor segmentation. Third, we classify the motion of the user into different classes like 'upper body motion', 'full body motion', and 'no motion'. For that purpose, a difference image is calculated between successive sub-sampled gradient images in the bounding boxes of the user. The classification into the motion classes is done by comparing the integral and the statistical moments of the activation in the motion histogram.

Finally, structural knowledge is extracted by localizing the user with respect to predefined room and object maps of the environment. This step is easy, because we already rely on an occupancy map of the environment for localization of the robot and tracking the user in world coordinates. The current room plays an important role in the possible activities of the user, like 'cooking' in the kitchen, 'sleeping' in the bedroom, and 'watching TV' in the living room (Fig. 2(a)). Objects in proximity like a sofa, a table, the dish washer, the kitchen sink, or a bed define the current activity more precisely. The current room and objects are evaluated by calculating the integral of the Gaussian position estimation of the user in each region of the room- and object map. Fig. 2(b) shows an example where the robot and the user are located in the kitchen. Because of the uncertainty in the position estimation of the user, the states 'dishwasher' and 'cupboard' are both active in the respective probability distribution of the objects. The final observations are written to the user model representing the current system state (Fig. 1).

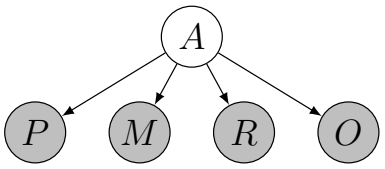

Fig. 3. *Bayesian Network for action inference. The observable features (gray) are depended on an unobserved action (white). A – action, P – pose, M – motion, R – room, O – objects.*

3.2. Actions

Since actions of a person are unobserved variables, we can only reason about them via the observed features. We model a Bayesian Network for each action we want to infer. The state for action A is then $p(A, P, M, R, O)$, where P, M, R, O are the observed features of the user's pose, the user's motion, the current room, and objects, respectively. Assuming that the features are independent given A the state space factorizes to:

$$p(A, P, M, R, O) = p(P|A)p(M|A)p(R|A) \quad (1)$$
$$p(O|A)p(A) \, .$$

The Bayesian Network for this factorization is shown in Fig. 3, where the variables of the system are displayed as *variable nodes* and the dependencies are visualized as arrows. The concept of factor graphs explicitly models the factors of (1) in *factor nodes* [18]. The dependencies of the features on an action are then coded in the factor potentials. These probability distributions can either be learned by using labeled training data and applying maximum a posteriori estimation or by applying hand-made rules from expert knowledge.

Once trained the Bayesian Networks are used to decide if a certain action is present or not. In our case, seven different Bayesian Networks are applied to estimate seven actions. Therefore, all observations from the user model are integrated into the Bayesian Networks and the sum-product algorithm is applied [18]. This infers the current unnormalized action likelihood $\mathcal{L}(\mathcal{A})$ in the current state of the system:

$$\mathcal{L}(A) = \sum_P p(P|A)p(P) \sum_M p(M|A)p(M) \quad (2)$$
$$\sum_R p(R|A)p(R) \sum_O p(O|A)p(O) \, .$$

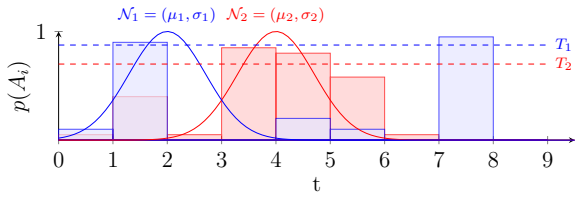

Fig. 4. *Action sequence for an activity that depends on two actions A_1 (blue) and A_2 (red). The activity becomes active once the actions are above thresholds T_i. Optionally each $p(A_i)$ could be weighted by a Gaussian prior to account for the beginning of the action and ordering.*

By normalizing $\mathcal{L}(\mathcal{A})$ one gets the probability of the action $p(A)$. The resulting probability distributions of the actions are written back to a sequence vector in the user model. This sequence of actions like 'motion in the upper body near the sink in the kitchen', 'walking from A to B', and 'sitting on the couch' is the basis for activity recognition.

3.3. Activities

Human activities usually have a complex structure. Example activities we want to detect include coming home, leaving, reading newspaper, resting, cooking, and going to bed. Because it is very hard to detect the starting point, temporal process and duration of activities, we assign a characteristic time span $\mathbf{S}$ and specific actions $A_i \in \mathcal{A}$ to each activity we want to estimate. An activity $p(Act)$ then becomes active, if the probabilities $p(A_i)$ of the actions assigned to the activity are above an experimentally defined threshold T_i in time span $\mathbf{S}$:

$$p(Act) = \begin{cases} 1 & , \max\limits_{\mathbf{S}} p(A_i) > T_i \, , \forall A_i \in \mathcal{A} \\ 0 & , \text{else} \, . \end{cases} \quad (3)$$

Thus, the order of actions is irrelevant and the duration of the activity could vary to a certain amount (Fig. 4). On the other hand, for some activities it is more applicable to restrict the actions to a certain starting point and apply a weak ordering constraint. This is done by weighting the occurrences of the actions by Gaussian prior distributions (Fig. 4). Therefore, each $p(A_i) \in \mathcal{A}$ in time span $\mathbf{S}$ is weighted by a Gaussian:

$$p(A_i)\mathcal{N}_i, \quad \text{with } \mathcal{N}_i = (\mu_i, \sigma_i) \, . \quad (4)$$

Note that we do not use probabilistic models to estimate activities but rely on this simple heuristic, because the high variance in the assignment of actions to activity would require a huge amount of training data. To improve recognition performance, a prior on the daytime could be added to distinguish activities that share similar actions. Activities like 'cooking', 'watching TV' or 'going to bed' roughly occur at the same point in time each day. Another improvement is the causal dependency to other activities. After the activity 'cooking' usually follow 'eating' and 'cleaning the dishes'. These causal priors are only applicable if one is not explicitely interested to detected deviations from daily routines and pre-recorded profiles. These two steps are not yet included in the current system, but we hope that they will significantly improve recognition performance in the future.

4. EXPERIMENTAL RESULTS

The complete system for activity recognition is not finished, yet. In this work, we evaluate a part of the system that includes the proposed features (Sec. 3.1) and the inference of different actions (Sec. 3.2). The evaluation of activity recognition (Sec. 3.3) based on a temporal sequence of actions is subject of future work. We are using the Bayesian Networks shown in Fig. 3 to estimate the actions of a person. Recall that these actions do not represent the final activities of a person. Yet, for descriptiveness, we assign the label of the current activity to the actions we are estimating. Hence, we create a Bayesian Network for each labeled activity the training and test set. In other words, each activity is then defined by only one distinctive action $p(A_i)$ and the time span $\mathbf{S}$ is one.

For training, we use sequences with single persons performing different activities in an apartment. We labeled each frame with the robot's position, the bounding box of the user and its current activity (Fig. 5(a)). Based on that knowledge, pose estimation, motion classification, room and object extraction is done. All these features are used to train the factor potentials of the Bayesian Networks using maximum a posteriori estimation. The training set contains 3,116 frames with seven different activities. After training, we apply the Bayesian Networks to a test sequence showing another person performing similar tasks in the same apartment. In this case, only the user's bounding box and the robot's position are labeled. Pose, Motion, the current room, objects are estimated (Sec. 3.1). These observation are processed by the Bayesian Networks to reason about the current activity (Sec. 3.2). The test set contains 2,234 frames with six different activities.

In the following, we focus on four exemplary activities shown in Table 1. The table also summarizes the occurrences of the activities in both data sets. On the one hand, 'reading', 'drinking' and 'watching TV' are relative short-term and static activities, where the user is often sitting with weak motion. On the other hand, 'cleaning the dishes' is a dynamic long-term activity including different positions and actions. For evaluation we calculated the ROC curve on the test sequence (Fig. 6). This curve plots the true positives rate vs. the false positives rate as the discrimination threshold of the system is varied. We vary the threshold paramater

(a) Training set (b) Test set

Fig. 5. *(a) Training set with labeled bounding boxes and activity of a person. All other features shown in the bounding box are estimated by the system. (b) Test set of a different person in the same apartment with labeled bounding boxes only.*

activity	training set	test set
drinking	372	95
watching TV	93	35
cleaning the dishes	305	86
reading a book	101	148

Table 1. *Number of frames for different exemplary activities in training and test sequence.*

T in (3) from zero to one for the ROC curve. This parameter defines which probability of action $p(A_i)$ must be reached to activate the coresponding activity. We count true positives, if an active activity matches the labeled activity. False positives are counted if an activity is active, but another activity label is given by the ground truth. Fig. 6 displays the overall performance of the system including all labeled activities. The performance of the system is rather weak reaching a true positive rate of 60% with 10% false postives. This is due to the fact that the inference of short temporal actions based on features is not enough to capture the variance of complex, long-term activities. This becomes apparent in Fig. 7, that displays the ROC curves for exemplary activities. In this case, simple short-term activities like 'watching TV' are better classified than complex, long-term activities like 'cleaning the dishes'.

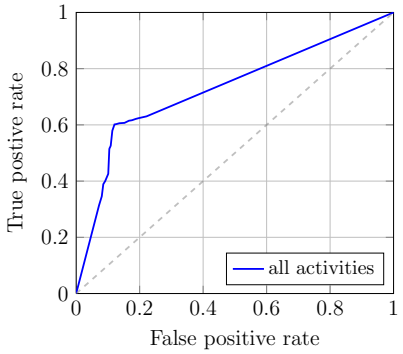

Fig. 6. *ROC curve for all labeled activities with varying threshold T.*

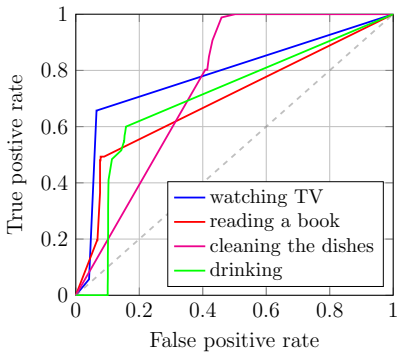

Fig. 7. *ROC curve for exemplary activities with varying threshold T.*

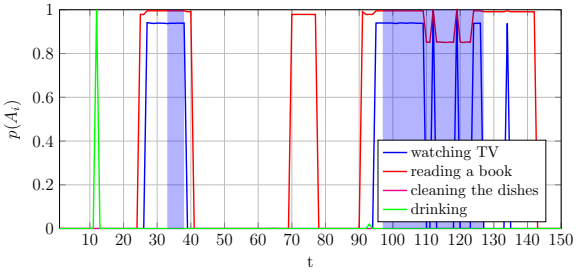

Fig. 8. *Probabilities of different activities. Ground truth of watching TV is overlaid by blue rectangles. Similar activities like reading a book and watching TV cannot be differentiated, yet.*

Therefore, we hope that the integration of time spans and the assignment of multiple actions to one activity will improve the system performace. Another aspect supporting this reasoning is the fact, that the actions of similar activities cannot be differentiated by the system so far. Due to the current implementation, that only assigns one action to each activity, the activity 'watching TV' cannot be distinguished from a similar activity like 'reading a book' (Fig. 8). In both cases, the person is mostly sitting on the sofa in the living room and has weak upper body motion. The figure illustrates the chronological sequence of the probability of the activity 'watching TV' with overlaid ground truth labels. For comparison also the estimated probabilities of the activities 'reading a book', 'drinking' and 'cleaning the dishes' are visualized. As can be seen in the figure 'watching TV' and 'reading book' are usually both active at the same time. This is a limitation of the current system as we are not able to distinguish between similar activities. On the other hand, the activity 'drinking' can be distinguished because the person has stronger upper body motion and the pose of the person must not necessarily be sitting. The probability of activity 'cleaning the dishes' is zero for the whole sequence, because the person was not situated in the kitchen.

The current system runs in real-time on a standard 2.4 GHz CPU with 20Hz allowing enough processing

power for other tasks required by the mobile system. Most of the processing time is consumed by the HOG detector to estimate the user's head position in the bounding box of the user, while the estimation of the actions is rather cheap. Note that, the running time for the final system will be higher because the action sequence evaluation is not included, yet. Furthermore, we worked on labeled bounding boxes. In the final application, robot specific modules like the person tracker, localization and navigation run in parallel. However, because of the real-time capability of these system components and the little processing time of the preliminary system, we are still confident that the final system will run in real-time.

5. CONCLUSION AND DISCUSSION

This paper presented a real-time method to reason about user's activities on a mobile robot in the scenario of a home environment. The system extracts different features like the user's pose and motion from camera images and combines them with expert knowledge of the environment. Thereby, the current room and objects in proximity of the user can be estimated. All features are processed by Bayesian Networks to reason about different user actions. A sequence of these actions defines the activity of a user. In first experiments, we showed that the proposed features are suitable to infer different actions.

In future work, we want to apply more sophisticated pose recognition to augment the user's pose state with 'lying' and 'fallen on the ground' to detect activities like 'sleeping' or 'critical situation'. The complete activity recognition system based on sequences of actions is then evaluated on challenging real datasets of home scenarios recorded by a mobile robot. Limitations of the current system include the dependency on training data or expert knowledge from hand-made rules. To capture the high inter- and intraclass variation of actions and activities, one requires either huge mass of training data or many hand-made rules. Getting one of these is very hard and often expensive. Therefore, we seek for a dialog-driven system that learns the addressability and attention of the user on-line by incorporating feedback from the user. Hence, the meaning of different activities is unimportant to the system, but the dependency of activities to the user's addressability is learned. By combining the activity estimation proposed in this work with an adaptive system that reasons about the user's will to interact, we hope to develop a system that can sustainably improve human machine interaction.

6. REFERENCES

[1] M. Ahad, J.K. Tan, H.S. Kim, and S. Ishikawa, "Human activity recognition: Various paradigms," in *International Conference on Control, Automation and Systems*, 2008, pp. 1896–1901.

[2] Yu-chen ho, Ching-hu lu, I-han chen, Shih-shinh huang, Ching-yao wang, and Li-chen fu, "Active-learning assisted self-reconfigurable activity recognition in a dynamic environment," in *IEEE ICRA*, 2009, pp. 813–818.

[3] G. Virone and A. Sixsmith, "Monitoring activity patterns and trends of older adults," in *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2008, pp. 2071–2074.

[4] V. Osmani, S. Balasubramaniam, and D. Botvich, "A bayesian network and rule-base approach towards activity inference," in *IEEE 66th Vehicular Technology Conference*, 2007, pp. 254 – 258.

[5] Tâm Huynh and Bernt Schiele, "Unsupervised discovery of structure in activity data using multiple eigenspaces," in *2nd International Workshop on Location- and Context- Awareness*, Dublin, Ireland, 2006.

[6] Maja Stikic and Kristof Van Laerhoven, "Recording housekeeping activities with situated tags and wrist-worn sensors: Experiment setup and issues encountered," in *Proceedings of the 1st International Workshop on Wireless Sensor Networks for Health Care*, 2007.

[7] J. Lokman, J.-i. Imai, and M. Kaneko, "Understanding human action in daily life scene based on action decomposition using dictionary terms and bayesian network," in *2nd International Symposium on Universal Communication*, 2008, pp. 67 –74.

[8] F. Niu and M. Abdel-Mottaleb, "Hmm-based segmentation and recognition of human activities from video sequences," in *IEEE International Conference on Multimedia and Expo*, 2005, pp. 804 –807.

[9] Fawang Liu and Yunde Jia, "Human action recognition using manifold learning and hidden conditional random fields," in *The 9th International Conference for Young Computer Scientists*, 2008, pp. 693 –698.

[10] M. Singh, A. Basu, and M.K. Mandal, "Human activity recognition based on silhouette directionality," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 9, pp. 1280 –1292, 2008.

[11] A. Veeraraghavan, A. Srivastava, A.K. Roy-Chowdhury, and R. Chellappa, "Rate-invariant recognition of humans and their activities," *IEEE Transactions on Image Processing*, vol. 18, no. 6, pp. 1326–1339, 2009.

[12] Hui Li, Qingfan Zhang, and Peiyong Duan, "A novel one-pass neural network approach for activities recognition in intelligent environments," in *7th World Congress on Intelligent Control and Automation*, 2008, pp. 50 –54.

[13] M. Losch, S. Schmidt-Rohr, S. Knoop, S. Vacek, and R. Dillmann, "Feature set selection and optimal classifier for human activity recognition," in *IEEE International Symposium on Robot and Human interactive Communication*, 2007, pp. 1022–1027.

[14] St. Müller, E. Schaffernicht, A. Scheidig, H.-J. Böhme, and H.-M. Gross, "Are you still following me?," in *European Conference on Mobile Robots*, 2007, pp. 211–216.

[15] Paul Viola and Michael Jones, "Robust real-time object detection," in *International Journal of Computer Vision*, 2001.

[16] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[17] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[18] F.R. Kschischang, B.J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *Information Theory, IEEE Transactions on*, vol. 47, no. 2, pp. 498 –519, feb 2001.