Human Activity Recognition in Multiview Video

Sławomir Maćkowiak, Paweł Gardziński, Łukasz Kamiński and Krzysztof Kowalak
Poznań University of Technology, Poland
{smack, pgardzinski, lkaminski, kkowalak}@multimedia.edu.pl,
WWW home page: http://www.multimedia.edu.pl

Abstract

In this paper, a novel multiview video based human activity recognition system which automatic detects of such behavior as fainting, a fight or a call for help is presented. The approach proposed in this paper used a directed graphical model based on propagation nets, a subset of dynamic Bayesian networks approaches, to model the behaviors.

The performance of activity recognition is analyzed for three methods of characteristic points forming a behavior descriptor (four extreme points over contour,four extreme points over contour with different normalization process and n-evenly distributed points on the contour). The results prove high score of recognition of the system for "Calling for help", "Faint", "Fight", "Falling" and "Bend at the waist" behaviors.

1. Introduction

In the area of surveillance, automated systems to observe pedestrian traffic areas and detect dangerous action are becoming important. Many such areas currently have surveillance cameras in place. However, all of the image understanding and risk detection is left to human security personnel. This type of observation task is not well suited to humans, as it requires careful concentration over long periods of time. Therefore, there is clear motivation to develop automated, intelligent, vision-based monitoring systems that can aid a human user in the process of risk detection and analysis.

The scope of review is limited to some widely-used graphical models that have been applied on complex human activity modeling based in crowded public scenes. There is a great diversity of methods to activity modeling in single camera view including: probabilistic graphical models (*e.g. Bayesian networks* [10,11], *dynamic Bayesian networks* [12-14], *propagation nets* [15]), probabilistic topic models (e.g. *probabilistic latent semantic analysis model* [16], *latent Dirichlet allocation model* [17,18], *hierarchical Dirichlet processes model* [19, 20]), Petri nets [21], syntactic approaches [22, 23] or rule-based approaches [24]. The approach proposed in this paper belongs to a subset of Bayesian networks therefore these approaches are discussed in more detail. Detailed reviews on other approaches such as Petri nets, neural networks, synthetic approaches, and rule-based approaches can be found in a survey by Turaga et al. [25].

A Bayesian network or belief network is a directed acyclic graphical model with nodes representing variables of interest (e.g. consecutive states of an event) and the links encoding dependencies among the variables. The strength of a dependency is parameterized by conditional probabilities that are attached to each cluster of parent-child nodes in the network. Bayesian network has been a popular tool for activity modeling due to its powerful capabilities in representing and reasoning uncertain visual observations, as well as its computational feasibility.

A dynamic Bayesian network extends Bayesian network by incorporating temporal dependencies between random variables. Hidden Markov Model (HMM), the simplest dynamic Bayesian network with one hidden state variable and one observation variable at each time instance, has been extensively used for activity modeling and recognition. To model more complex activities, various topological extensions to the standard HMM have also been developed, which factorize the state space and/or observation space by introducing multiple hidden state variables and observation state variables [26].

Next approaches used in activity recognition are probabilistic topic models. In activity modeling, human events in video are often treated analogously as words in document analysis. Each video may be viewed as a mixture of dependencies that represent events. Typically, a video sequence is divided into a sequence of short clips, each of which is annotated by words that are constructed from extracted features accumulated over a temporal window. A trained probabilistic topic models can then be applied to evaluate normality of each event (i.e. word) whilst considering interactions (i.e. topic) between them. Probabilistic topic models have found wide applications in single view analysis especially in model human interactions within crowds for unusual event detection.

Apart from Bayesian network and probabilistic topic models, different graphical models have been proposed for activity modeling. For instance, propagation nets, a subset of dynamic Bayesian networks with the ability to explicitly model temporal interval durations, have been employed to capture the duration of temporal subintervals of multiple parallel streams of events.

The approach proposed in this paper is a directed graphical model with nodes corresponding to a consecutive states of an event in both space and time. The model exploited ability to explicitly model temporal interval durations between nodes. The nodes in a space represent consecutive parts of the event.

This paper is organized into 3 main sections. Section 2 presents the whole multiview video based human activity recognition system and explained required blocks of video processing. Section 3 presents the assumptions of the experiments and achieved results for several type of behaviors. Section 4 provides conclusions and suggests a number of areas to be pursued as further work.

2. System description

System presented in this paper uses N stationary cameras. First, the system must be calibrated to properly detect and track objects in 3D space. Next, to extract front side of human silhouette a face detection algorithm is used. Voxel model is reconstructed based on the light fields from multi camera system. Next, the reconstructed model is rotated by an angle which is determined from face recognition step. Finally, according to the type of the behaviors one of the three methods to describe object poses is used to recognize the activity of humans. Proposed system is shown in Fig. 1.

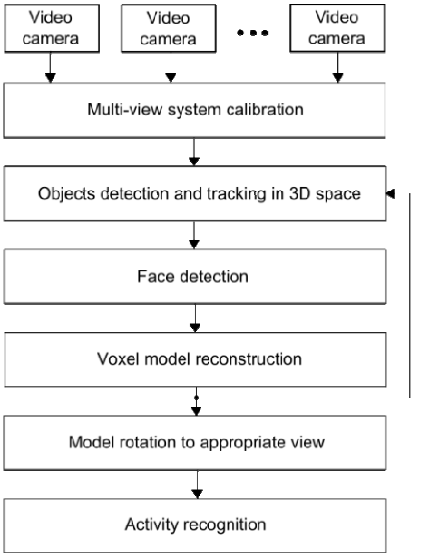
Video camera
Video camera
Video camera
Multi-view system calibration
Objects detection and tracking in 3D space
Face detection
Voxel model reconstruction
Model rotation to appropriate view
Activity recognition

Figure 1: Human activity recognition system.

2.1. Multiview system calibration

Cameras which are used in described system should observe as much combined area as possible. System calibration is performed by using a planar marker. Marker position is unambiguously interpeted in each view to calculate the position and rotation of each camera in space around the scene. Camera calibration allows to clearly identify objects present in each view.

2.2. Object detection and tracking in 3D space

For each camera in the system a background subtraction algorithm is applied. Background subtraction is used to extract foreground image which contains active regions in the images. These regions are mostly moving objects.

Object tracking is carried out in a virtual voxel space. It uses a feedback information from the voxel model reconstruction step to distinguish between the objects. It also uses Kalman filter to smooth trajectories of moving objects. Correct object tracking is needed to properly recognize activity of humans.

2.3. Human recognition and face detection

Each moving object is classified by a SVM classifier. It uses HOG features descriptors. If object is classified as human, the face recognition step is applied. Face detection is performed by cascade classifier with Haar-like features descriptors.

2.4. Voxel model reconstruction

Camera positioning is known, therefore it is possible to conduct the voxel model reconstruction to represent observed scene as a 3D model (Figs. 2-4). The process is well known and recreates a simplified light trajectory using pinhole camera model according to:

$$r = \frac{Size_{voxel}}{\rho_{step}}, \tag{1}$$

$$\alpha = \frac{\alpha_{Cam} * \pi}{180} - \tan^{-1}\frac{y}{f_{Cam}}, \tag{2}$$

$$\beta = \frac{\beta_{Cam} * \pi}{180} - \tan^{-1}\frac{x}{f_{Cam}}, \tag{3}$$

$$\Delta\colon (dx, dy, dz) = (r * \sin\beta\,, -r * \cos\alpha\,, \\ r * \cos\beta) \tag{4}$$

Where:

- r is an elementary projection step;
- x is a horizontal coordinate of currently computed pixel;
- y is a vertical coordinate of currently computed pixel;

- Δ is vector designating direction and length of the projection step in coordinate system defined in earlier chapters.

In that way, the visible voxels are determined by marking them from all views.

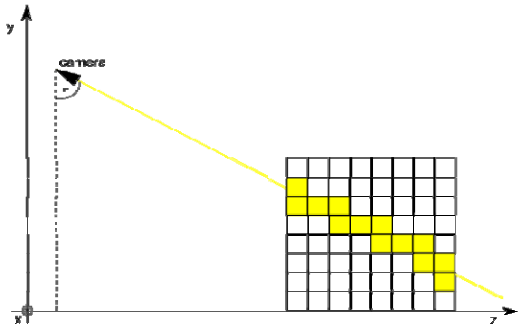

Figure 2: Light projection and voxel marking.

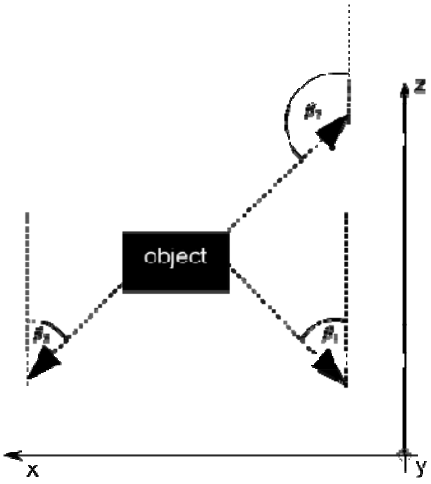

Figure 3: Vizualisation of $\boldsymbol{\beta}$ in reconstructed scene.

2.5. Model rotation to appropriate view

Reconstructed objects are placed in 3D scene. For each activity reconstructed model is rotated by a different angle. The central point of rotation is set individually for each object in scene. This step allows to reduce number of descriptors used in *Activity recognition* step.

2.6. Activity recognition

The behavior of a person can be described by a set of trajectories of characteristic points of the person, as shown on Fig.5.A set of characteristic points at a given time defines a pose. A set of poses defined for consecutive time points or a set of time vectors for individual points forms a descriptor.

The set of points to define a pose may have a different configuration for different types of behavior to be detected. In other words, for at least one type of behavior, a set of points is generated having a configuration different than a set of points for another type of behavior. For consecutive frames, the positions of points belonging to the set are traced and form trajectories of points.

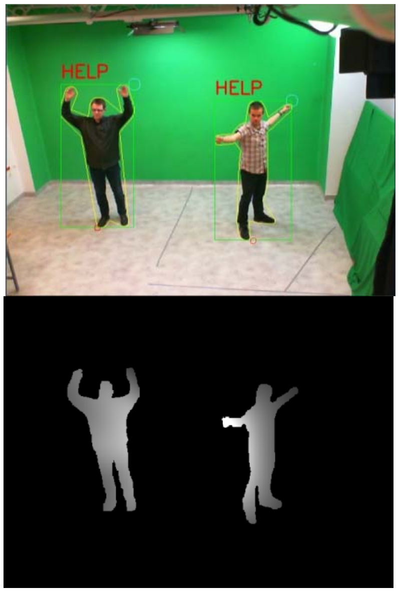

Figure 4: Scene with two objects and reconstructed voxel model.

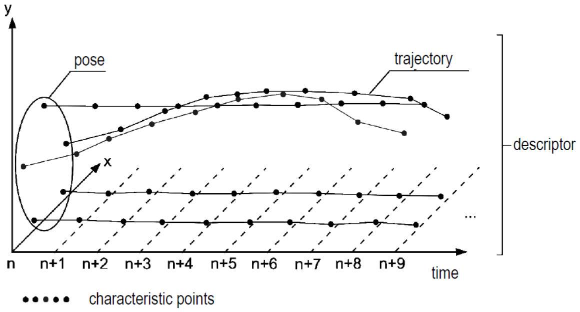

Figure 5:Behavior as a set of poses, characteristic points form trajectories in time.

The comparison is performed by calculating the Euclidean distance for pairs of corresponding points. Each trajectory shall fit within a predetermined range. For example, assuming that 4 points of a person are traced (e.g. two palms and two feets), the characteristic point designated as the right palm must, for each frame, be located in a distance D not larger than δ from the reference “right hand” for each behavior, namely:

$$D = \sqrt{(x_{obs} - x_{ref})^2 + (y_{obs} - y_{ref})^2} \qquad (5)$$

$$D \leq \delta \qquad (6)$$

wherein x_{obs} , y_{obs} designate the position of characteristic points (note: these are not spatial coordinates) and x_{ref},y_{ref} designate corresponding reference values.

There are three methods created to describe object poses:

- four extreme points over contour,
- four extreme points over contour (different normalization process),
- n-evenly distributed points.

2.6.1 *Midpoint and four extreme points over contour*

The first method, according to one example embodiment of the system described in this paper, is called “Midpoint and four extreme points”. This is well suited for detecting behavior, in which human limbs are widely positioned, for example, while waving person’s arms or crying for help.

The method sets four points {A, B, C, D}, wherein the Euclidean distance from the geometric center of the contour P is the greatest. These points are computed, one in each quadrant of the coordinate system having a center located at point P, by the formula:

$$A_x = \frac{p_x^i - P_x}{w} \qquad (7)$$

$$A_y = \frac{p_y^i - P_y}{h} \qquad (8)$$

Wherein the following references refer to:

A_x, A_y– coordinates x and y of the calculated point,
w – hull's width on 0X axis,
h – hull's height on 0Y axis,
P_x, P_y– coordinates *x* and *y* of the hull's central point,
p_x^i, p_y^i– coordinates *x* and *y* of the *i*-th hull's point.

2.6.2 *Midpoint and four extreme points over contour(different normalization process)*

Another example method, also utilizing the concept of the “Midpoint and four extreme points”, applies a different normalization process. The method differs from the previous one in that it applies normalization of *x* coordinates, calculated according to the formula:

$$A_x = \frac{p_x^i - P_x}{h} \qquad (9)$$

2.6.3 *n-Evenly distributed points (nEDP)*

Anauthors'proposed method is n-Evenly Distributed Points (nEDP). It means that points are evenly spread on the hull. The method of points selection is based on selection of evenly spread, arbitrary number of points from the hull (typically of a silhouette). Such approach allows for gradual selection of the level of hull mapping. The method has been depicted in figure 6, wherein the following references refer to:

- *Step* – step of selection of consecutive points;
- *Buff* – a buffer storing reminder of a division in order to minimize an error caused by calculations on integer numbers. It is a case when the Step is not an integer;
- *rest(lPktKont/lPkt)* – a function calculating a reminder of a division.

The selected hull points define a descriptor of the hull (typically of a silhouette of a person) in a given video frame and are buffered in an output vector as shown in figure 6.

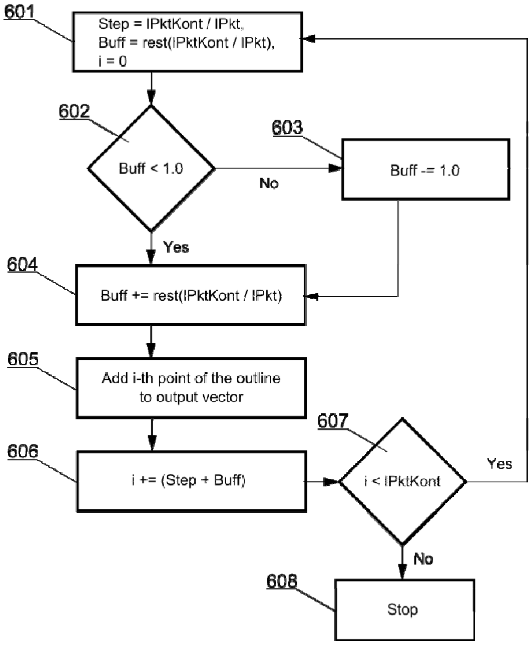

Figure 6:Evenly Distributed Points - algorithm block diagram.

More specifically, the procedure in figure 6 starts from step 601, where the Step variable is set a value of *lPktKont/lPkt*, wherein *lPkt* denotes chosen number of equally distributed points on object contour and lPktKont denotes a total number of points in object contour, the *Buff* variable is set to *rest(lPktKont / lPkt)* and the *i* variable is set to 0. Next, at step 602, it is verified whether the value of the *Buff* variable is less than 1. In case it is not, at step 603 the value of *Buff* variable is decreased by 1.0 before moving to step 604. Otherwise, the step 603 is skipped and step 604 is executed where the value of the *Buff* variable is increased by the rest of the quotient *(lPktKont/lPkt).*Subsequently, at step 605, the *i*-th point of the hull is added to an output vector as a selected point. Next, at step 606, the variable *i* is set to a value of *Step* + *Buff.* Next, it is verified 607 whether *i* is lower that *lPktKont* and in case it is the process returns to step 601 in order to process the next point or otherwise the procedure ends at step 608.

All three methods were tested and used with behaviors that suited most and gave best results, which are as follows:

- *Calling for help* with method 2.6.1
- *Faint* with method 2.6.1
- *Fight* with method 2.6.2
- *Falling* with method 2.6.1
- *Bend at the waist* with method 2.6.3

Those experimental results are described in section 3.

3. Experimentalresults

The experiments were performed for many different behaviors, such as Calling for help, Faint, Fall, Fight and *Bend at the waist*. In our experiments all sequences in resolution 640x480 were divided with respect to the subjects into a training set (9 persons), a validation set (8 persons) and a test set (9 persons)[1]. The classifiers were trained on the training set while the validation set was used to optimize the parameters of each method (parameter mentioned in Section 3). The recognition results were obtained on the test set. Authors were looking for well-known databases needed to perform experiments but commonly used test sequences doesn't meet required conditions. Objects in each frame of video sequences were analyzed independently of a result obtained in previous frame. All experiments were conducted for active regions. Efficiency of classification was analyzed and for evaluating. For evaluating classification algorithms, it was used the *precision* and *recall* performance metrics, defined as follows:

$$precision = \frac{TP}{TP + FP}, \tag{10}$$

$$recall = \frac{TP}{TP + FN}, \tag{11}$$

where *TP* is the set of true positives, *FP* is the set of false positives and *FN* is the set of false negatives. According to achieved results (97.89% precision ratio and 74.27% recall ratio) the SVM with HOG have been chosen as diagnostic features for further experiments.

The effectiveness of the proposed approach was validated using descriptors which were created for each behavior independently. Each of those descriptors were tested in parallel configuration to distinguish between behaviors.

Table 1. Human activity experimental results.

Human activity	Precision	Recall
Calling for help	96.55%	93.33%
Fall	96.67%	96.67%
Faint	100.00%	93.33%
Bend at the waist	100.00%	93.33%
Fight	100.00%	80.00%

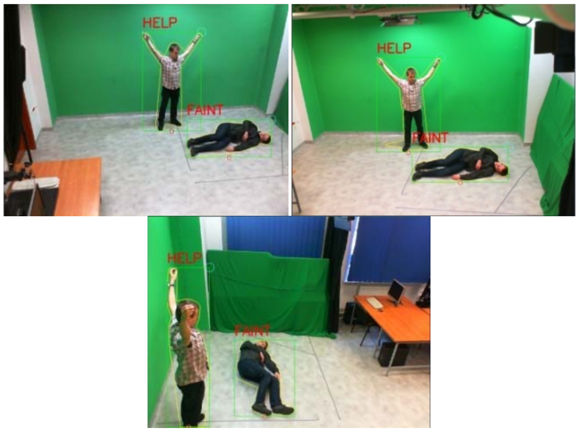

Figure 7: Results of detection two human activity ("Calling for Help" and "Faint") in one scene for 3 cameras.

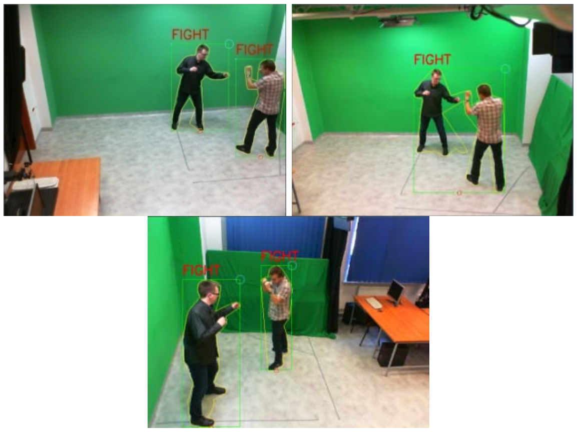

Figure 8: Results of "Fight" activity detection for 3 cameras.

During the tests, five activities were analyzed: calling for help, fall, faint, stomach ache and fight. Each of activities was detected independently of the other. Activity "Fight" is the only one which may be detected only when 2 or more humans are participating in it. The rest of the behaviors are detected independently for each objects. Results are shown in Table 1.

The results acquired during experiments demonstrate a high effectiveness level of activity recognition system. For each activities in which single object are participating, recall ratio exceeds 90%. In case of "Fight" activity recall ratio reaches 80%. It is caused by interpretation of interaction between an objects in scene. Methods for activity recognition described in this paper can't correctly describe these interactions.

[1] The test sequences used in the experiments are available at http://www.multimedia.edu.pl/human-behavior

4. Conclusions

In this paper, an innovative multiview activity recognition system is proposed. The proposed system is a complex solution which incorporates many techniques of object detection and moving objects tracking, as well as 3D scene reconstruction and human activity recognition. System uses voxel reconstruction method to reconstruct a scene in 3D space. System allows to free manipulate with scene to rotate a view to suitable fit and detect an activity of objects using one of three contour based algorithms. Results prove that proposed approach allows to achieve high human activity detection level. Further works focuses on extend an activity recognition algorithms by interactions between objects.

Acknowledgment

Research project was supported by the National Science Centre, Poland according to the Grant no. 4775/B/T02/2011/40.

References

[1] M. Wai Lee; R. Nevatia; „Body Part Detection for Human Pose Estimation and Tracking”, IEEE Workshop on Motion and Video Computing (WMVC’07), 2007.

[2] L. Zhao; „Dressed Human Modeling, Detection, and Parts Localization”, The Robotics Institute, Carnegie Mellon University, Pittsburgh, 2001.

[3] E. Corvee; F. Bremond; „Body Parts Detection for People Tracking Using Trees of Histogram of Oriented Gradient Descriptors”, AVSS – 7th IEEE International Conference on Audio Video and Signal Based Surveillance, 2010.

[4] K. Mikolajczyk; C. Schmid; A. Zisserman; , „Human Deteciton Based on a Probabilistic Assembly of Robust Part Detectors”, T. Pardla and J. Matas (Eds.): ECCV 2004, LNCS 3021, pp. 69-82, 2004.

[5] R. Cucchiara, C. Grana, G. Neri, M. Piccardi, and A. Prati, “The Sakbot System for Moving Object Detection and Tracking,” Video-Based Surveillance Systems—Computer Vision and Distributed Processing, pp. 145-157, 2001.

[6] W. Lao; J. Han; P. H.N. de With.; , „Fast Detection and Modeling of Human-Body Parts from Monocular Video”, F.J. Perales and R.B. Fisher (Eds.): AMDO 2008, LNCS 5098, pp. 380 389, 2008.

[7] Zivkovic Z., Improved Adaptive Gaussian Mixture Model for Background Subtraction, Proceedings of ICPR, 2004.

[8] N. Dalal; B. Triggs; , “Histograms of Oriented Gradients for Human Detection”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Montbonnot, France, 2005.

[9] G. Welch, G. Bishop,: "An Introduction to the Kalman Filter", TR 95-041 Department of Computer Science University of North Carolina at Chapel Hill, 2006.

[10] H. Buxton and S. Gong, "Visual surveillance in a dynamic and uncertain world", Artificial Intelligence, 78(1-2):431–459, 1995.

[11] S. S. Intille and A. F. Bobick. "A framework for recognizing multi-agent action from visual evidence", AAAI Conference on Artificial intelligence, pages 518–525, 1999.

[12] Y. Du, F. Chen, W. Xu, and Y. Li., "Recognizing interaction activities using dynamic Bayesian network", International Conference on Pattern Recogntion, pages 618–621, 2006.

[13] T. Duong, H. Bui, D. Phung, and S. Venkatesh. "Activity recognition and abnormality detection with the switching hidden semi-Markov model", IEEE Conference on Computer Vision and Pattern Recognition, pages 838–845, 2005.

[14] S. Gong and T. Xiang, "Recognition of group activities using dynamic probabilistic networks", IEEE International Conference on Computer Vision, pages 742–749, 2003.

[15] Y. Shi, A. Bobick, and I. Essa, "Learning temporal sequence model from partially labeled data", IEEE Conference on Computer Vision and Pattern Recognition, pages 1631–1638, 2006.

[16] J. Li, S. Gong, and T. Xiang, "Global behaviour inference using probabilistic latent semantic analysis", British Machine Vision Conference, pages 193–202, 2008.

[17] T. Hospedales, S. Gong, and T. Xiang, "A Markov clustering topic model for mining behaviour in video", IEEE International Conference on Computer Vision, pages 1165–1172, 2009.

[18] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behaviour detection using social force model", IEEE Conference on Computer Vision and Pattern Recognition, pages 935–942, 2009.

[19] D. Kuettel, M. D. Breitenstein, L. V. Gool, and V. Ferrari, "What’s going on? Discovering spatio-temporal dependencies in dynamic scenes", IEEE Conference on Computer Vision and Pattern Recognition, pages 1951–1958, 2010.

[20] X.Wang, X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models", IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(3):539–555, 2009.

[21] M. Albanese, R. Chellappa, V. Moscato, A. Picariello, V. S. Subrahmanian, P. Turaga, and O. Udrea, "A constrained probabilistic Petri net framework for human activity detection in video", IEEE Transactions on Multimedia, 10(6):982–996, 2008.

[22] M. Brand, "Understanding manipulation in video", International Conference on Automatic Face and Gesture Recognition, pages 94–99, 1996.

[23] 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(8):852–872, 2000.

[24] H. Dee and D. Hogg, "Detecting inexplicable behaviour", British Machine Vision Conference, pages 477–486, 2004.

[25] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities - a survey", IEEE Transactions on Circuits and Systems for Video Technology, 18(11):1473–1488, 2008.

[26] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition", IEEE Conference on Computer Vision and Pattern Recognition, pages 994–999, 1997.