Tracking People and Activities in Video Recordings of Classroom Presentations

João Pacheco[1]

Abstract This paper presents an intelligent system for video recording of classroom presentations. The system includes a tracking algorithm to track the speaker, recognizes six speaker's activities and records the presentation in video. The tracking algorithm considerably supports different indoor illumination conditions, tracks the speaker in both frontal and side views, and adapts to body scale. Speaker's face and hand regions are obtained by tracking skin regions and torso vertical boundaries are given by the median edge points. The activity classifiers achieved recall rates from 10 to 86.67%.

1 Introduction

Nowadays, many interactive presentations are given all over the world, whether in academic classrooms, business meetings or scientific conferences. For students, recorded classes would be useful so they could watch them, whether they were or not present. Furthermore, students attending distance courses would consider this essential to their learning, as well as those who wish to learn more about a particular topic. Thus, we can see that a recorded multi-view presentation, that includes the speaker and audience images/audio, and its slide show, is a fast and cheap way of sharing the knowledge.

Recognition of events (or activities) in video may be useful to to analyze the speaker's non-verbal communication; to automatically segment the presentation in several videos, according to specific events, and to automatically label the videos by their content.

In this paper, Section 2 reviews the previous work on systems for presentation rooms and similar, human tracking and human activity recognition. Section 3 describes the problems involved in this work and Section 4 describes the proposed system. In Section 5, the experimental results are presented.

2 Previous work

Many researches have focused on extracting information in a non invasive way from the video images of presentations, smart rooms and related environments. This section provides an overview on these works and also on the tracking and activity recognition approaches which enabled high level information extraction.

Instituto Superior Técnico, Universidade Técnica de Lisboa,
Av. Rovisco Pais,
1049-001 Lisboa, Portugal,
joao.d.pacheco@ist.utl.pt

2.1 Interactive Presentations and Meetings

Several systems have been developed for using in the environment of presentations and smart rooms. These are intended to extract information about what occurs inside the room. Often, they try to achieve some intelligent behavior, such as human tracking and/or face recognition [1] [2] [3] [4] [5] [6] and recognition of activities [7] [8] [9] [10] [11].

Bernardin et al. [2] presented a multiple human tracking system where the talking person is tracked based on the visual and audio information, and simultaneously, the system tries to identify that person. In the work of Wu and Nevatia [3], a multi-person tracking is achieved on a conference room. The authors detect each person from head and shoulder and their approach is insensitive to the camera motion, which is an important advantage. On the other hand, they use a single camera which gives a single point of view of the conference. In [12], Zhang et al. describe a single person tracker within a smart room. It performs a 3D tracking using four static cameras with overlapping fields of view.

Close to [2], but with static cameras, [4] develops a fusion between a face recognition system and a speaker identification system, based on video and speech. In [9] there is a system that uses three static cameras in an office environment to track humans and recognize some of their actions. Potamianos et al. [5] developed a system for a smart room where the talking person is tracked by PTZ cameras. Its main goal is tracking person's face and mouth from frontal and non-frontal views, as a visual way of knowing the person is talking. Most systems use static cameras, assume the existence of few or only one person in the scene, and that at least one of the cameras contains the person in its field of view.

2.2 Human Detection and Tracking

There are several techniques to track people, but some have been more frequently used for their effectiveness, ease of implementation or speed. Pfinder [13] tracks a person by tracking its blobs. Mean Shift [14] is a color based method for tracking and illumination sensitive, but there are methods which detect human body parts and combine them [15], reducing exposure to problems originated from illumination changes. Another approach is the creation of a human skeleton, by connecting some key points [16].

Skin detection for body part tracking is widely used. There are some skin modeling approaches, but the simplest and fastest is the heuristic approach which is based on a set of constraints [17]. It requires a large training set, it is sensitive to illumination changes and requires choosing a good color space. Face detection and tracking has been mainly improved by the work of Viola and Jones [18] due to its effectiveness, speed and motion independence.

2.3 Activity Recognition

In the literature on activity recognition there is a reasonable amount of single person activities that have already been covered, such as a person entering or exiting a room, a person at the computer, at the white board, sitting down, getting up, picking an object, walking, running, looking for an object, writing on the board or on a sheet of paper, swiveling left/right, doing sports or physical exercises, or doing specific hand gestures [9] [16] [19] [20] [21] [22] [23] [24] [25] [26]. Several algorithms and frameworks have been used for human activity recognition, such as HMM [24], Artificial Neural Networks (ANN) [9], SVM [22], Transferable Belief Model [23], optical flow [27], Fiedler Embedding [28], VSIP [29] and Pyramid Match Kernel [21]. HMM and ANN are the most used algorithms, although the others show satisfactory results (recognition rates above 60%)

3 Problem Description

This work focuses on three main problems. The first problem is the detection and tracking of speaker's body parts (face/head, torso and hands) within a classroom during a presentation. The speaker always moves in an indoor scene, but changes in lighting are expected, the audience moves and there are body occlusions. These factors difficult tracking the speaker.

The second problem is recognizing a set of activities that the speaker performs. Recognizing these activities heavily depends on the ability of the tracker to track speaker's body parts. The activities to recognize are: the speaker's face is visible (A1), the speaker is pointing to his/her right (A2), and the speaker has moved to his/her left (A3). Assuming the tracker is reliable, activity recognition requires the analysis of several frames, in order to understand speaker movements over the time.

The third problem is recording the presentation into two video files. One video contains the presentation global view, and the another contains a clipped view of the speaker, taken from the global view.

4 Solution

The system's architecture is divided into three main modules: human tracker, activity recognition and video recording. Starting from the capture image, the human body tracker estimates speaker's face, torso and hands. Second module uses tracker's estimation to perform activity recognition and it includes components for training and testing. Recording module records two videos to hard disk - the original video and a clipped video which contains speaker's body. The following sections describe these modules.

4.1 Tracking Algorithm

The human body segments considered in this paper is the face, torso and both hands. These are the chosen body parts to track, since they are the visible body parts for most of the presentation time, unlike legs and feet. Note that the arms are not distinguished from hands, so the hand regions may be associated with speaker's arms, if they are not covered by clothes.

The tracking algorithm is composed of four steps: background subtraction and tracking of face, torso and hands. Background subtraction provides the image region where the speaker is. Face is searched in that region through skin blobs. Torso vertical boundaries are obtained from the median edge points of the speaker's torso. Hand tracking consists in computing the existing skin blobs on the two sides of the face and below it.

In this paper any region $\mathscr{R} = (x, y, w, h)$ may be expressed by its center point in image coordinates (x, y), width w and height h. $\mathscr{R}_c$ denotes the region center and $\mathscr{R}_d$ denotes its w and h.

4.1.1 Background Subtraction

To distinguish the speaker's body from the background, we implemented Horprasert's algorithm [30], except the parallel processing. Some features have been improved or changed from the original algorithm, namely an adaptive updating of the background model, a technique to reduce the number of processed pixels and the pixel classification was simplified so that shadows and highlights are considered background. The background model $B(t)$ is updated for every time t and background pixel of coordinates (x,y) as follows:

$$B(x,y,t) = (1-\alpha_B)\, B(x,y,t-1) + \alpha_B\, I(x,y,t) \tag{1}$$

where I is the current image and $\alpha_B = 0.000199696$ is the update rate, so B is renewed every 10 minutes. In order to process only the region where the speaker moves, we define a region $R^{\mathscr{F}}$ centered in $c^{\mathscr{F}} = (c_x, c_y)$, where $\mathscr{F}$ is the foreground region and $c^{\mathscr{F}}$ is the centroid of $\mathscr{F}$. Width and height of $R^{\mathscr{F}}$ are denoted by $R_d^{\mathscr{F}} = (\upsilon_1 W, \upsilon_2 H)$. W and H denote the image's width and height, respectively, $\upsilon_1 \in [0.1, 0.5]$ and $\upsilon_2 \in [0.358, 1]$, given the expected average scale of the speaker in I.

4.1.2 Face detection and tracking

Face detection is accomplished through skin blobs. Skin blobs are rectangular regions computed from I characterized by their center (x,y) and by their width w and height h. In general, a skin blob is an image region where there is an 8-connected component of pixels classified as skin colors. Skin pixel classifier used is suitable to uniform daylight illumination [17]. Assuming that the speaker is standing and there is no other person in the image faced to the camera, the face is the highest skin blob of I (blob with the smallest y). A blob list Q is computed for region $R^{\mathscr{F}}$ and contains only blobs whose skin points number is between β_1 and β_2, where $\beta_1 = 9$ and $\beta_2 = 3\ \beta_1$. In order to avoid choosing as face a region which contradicts the human proportions, we only consider the blobs Q^i which satisfy

$$\frac{Q_w^i}{Q_h^i} \leq \frac{Q_h^i}{Q_w^i} < r \tag{2}$$

where Q^i is the i-th blob of Q, Q_w^i is the blob width, Q_h^i is the blob height and $r = \frac{3}{2}$ is the maximum dimension ratio. These blobs are called square blobs. Blobs in the image's top or bottom are rejected because the face does not appear there. Finally, face region is $F(t) = Q^i$ where $i = \arg\min_{\hat{\mathbf{k}}} Q_y^{\hat{k}}$ and $Q_y^{\hat{k}}$ is the y coordinate of the blob.

Since it is assumed that the speaker does not move rapidly from $t-1$ to t, $F(t)$ should be in the neighborhood of $F(t-1)$. Thus, the approach for tracking the face is look for the skin blobs in the neighborhood of $F(t-1)$, and assign to $F(t)$ the blob whose similarity with $F(t-1)$ is the highest. In the tracking phase, β_1 is based on the area of $F(t-1)$ as a way of reducing the number of undesired blobs and it is computed as

$$\beta_1 = \lceil\ \beta_5\ F_w(t-1)\ F_h(t-1)\ \rceil \tag{3}$$

where $\lceil\ \rceil$ denotes the ceil function and $\beta_5 = 0.06$. When Q contains blobs, $F(t)$ is assigned to a blob $Q^i \in Q$ which presents the highest similarity with $F(t-1)$ where i is given by:

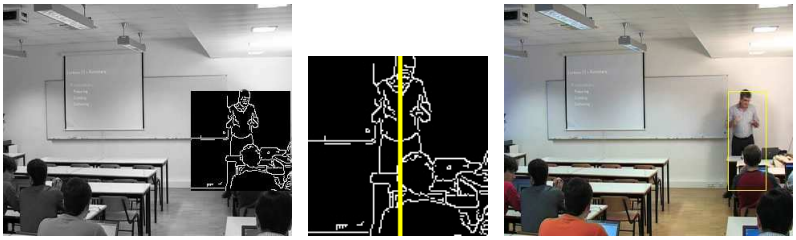
$$i = \arg\min_{\hat{\mathbf{k}}} \left(w_1\ |a(F(t-1)) - a(Q^{\hat{k}})| + w_2\ ||c(F(t-1)) - c(Q^{\hat{k}})||^2. \right) \tag{4}$$

where $w_1 = w_2 = 0.5$, operator a provides the pixels number of a blob and operator c gives the centroid of a region. If Q contains no blobs, the algorithm tries to detect a square blob again.

4.1.3 Torso tracking

Torso region $To(t)$ is detected in every image I and requires knowing $F(t-1)$. Its detection relies on torso's edge points and knowledge about the human proportions. Firstly, the detection method defines a region R^{To} where the torso is expected to be, depending on the face location and scale. The expected location of $To(t)$ is below the face, since it is assumed the speaker is standing and upright. Secondly, edge points within R^{To} are computed with Canny's edge detector from I as shown in Figure 1(a). Then, R^{To} is split into regions R_1^{To} and R_2^{To} (Figure 1(b)). Regions R_1^{To} and R_2^{To} are considered to contain right and left boundaries of the torso, respectively. Then, the approximate right and left boundaries are the median of x coordinate in the corresponding regions. Once obtained right and left

boundaries, the remaining top and bottom are computed from the knowledge of human proportions. Figure 1(c) shows the computed torso region. In Figure 1(c), the computed left boundary is a little away from the speaker's torso left boundary because the algorithm considered a large amount of objects' edges. Still, the algorithm's results and simplicity compensate the errors from other objects' edges, given the goals of recognizing activities and achieving a real time algorithm.



(a) Image after detecting the edge points of the speaker.

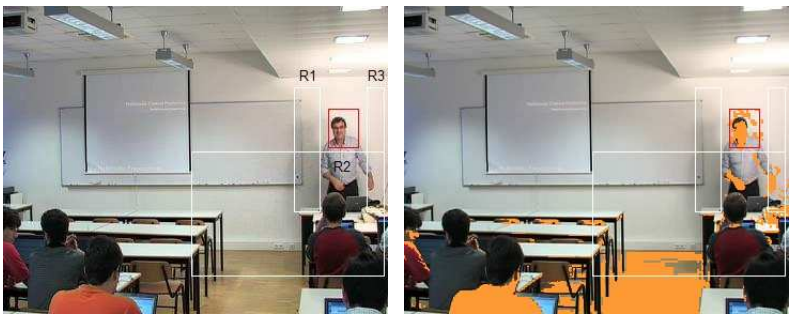
(b) Division of region R^T into two equal regions.

(c) Computed torso region.

Fig. 1 Example of the torso detection.

4.1.4 Hand tracking

Hand tracking is performed through skin blobs, after knowing $F(t)$. Since hands cannot be too far from the face, F_c and F_d are used to define a search region for the hands. Hands may be over the torso plane, over the legs plane, and at each side of the head and torso. As a result, three searching regions (R_1^{Ha}, R_2^{Ha} and R_3^{Ha}) are defined as shown in Figure 2(a) as R1, R2 and R3 respectively. From Figure 2(b), one can observe the existence of many skin blobs within the three regions. The



(a) Image regions where the hands are searched (R1, R2 and R3).

(b) Image with hand search regions and skin pixels (orange pixels).

Fig. 2 Regions around the speaker where hand blobs are searched.

algorithm chooses at most two skin blobs and rejects the others based on the following rules:

1. Blob's area must be greater than 1 and less than the double of face's area.
2. Distance between F_c and the blob farther boundary in x coordinate should be less than 4.25 ω.
3. Distance between F_c and the blob farther boundary in y coordinate should be less than 4.25 ω.
4. Blob cannot intersect $F_c(t)$.
5. Blob width must be less than 2 ω.

$\omega = \frac{4}{5}\ F_h^{\tilde{d}}$ is the deduced face height from the default face $F^{\tilde{d}}$ where $F_h^{\tilde{d}} = \lceil \gamma_1\ F_h(t) \rceil$ and $\gamma_1 = 1.1$. Once the hands are detected, tracking is very similar to face tracking. Therefore, each hand is searched in the neighborhood of the previous hand region. To compute the most similar skin blob, we defined a default hand $Ha^d(t)$ which provides a hand model based on both hands sizes. Default hand width is given by:

$$Ha_w^d(t+1) = \theta\ Ha_w^d(t) + (1-\theta)\ \frac{LH_w(t) + RH_w(t)}{2} \tag{5}$$

where $\theta = 0.5$. Its height $Ha_h^d(t+1)$ is analogous to the component h.

4.2 Activity Recognition

In a presentation, the speaker can simultaneously perform several activities. As a consequence, it is not possible to recognize all the activities with a single classifier, leading to create a separate binary classifier for each activity. Each activity classifier was trained with positive and negative examples of its activity. The classifier's parameters are a classification algorithm, a feature set, a sliding window size and a method for building the feature vector.

Classification algorithms used were SVM and Normal Bayes. Features used include mostly body points locations, distance between points and body point displacements over the time. The sliding window size K is the number of frames used to collect features to insert into the feature vector $\mathscr{V}_\alpha$. Two methods for building the feature vector were developed. Let $M = j - i + 1$ be the frames number of a training example, where i and j are the numbers of the first and last frames of the example, respectively. Let also f_l be the feature vector of frame l and K is the sliding window size. Each feature vector $\mathscr{V}_\alpha$ is given by:

$$\mathscr{V}_\alpha = f_u \cup f_{u+1} \cup \cdots \cup f_{u+K-2} \cup f_{u+K-1}. \tag{6}$$

In Method 1, $u = [i + \alpha\ K,\ i + M - K]$ and $\alpha = [0, \frac{M}{K} - 1]$ ($\alpha \in \mathbb{N}$), while in Method 2, $u = [i, i + M - K]$ and $\alpha = u$. In both methods, $\mathscr{V}_\alpha$ is only built if $M \geq K$ because it is required to collect features from at least K frames. The test vector $\dot{\mathscr{V}}$ is defined as

$$\dot{\mathscr{V}} = f_{\dot{u}-K+1} \cup f_{\dot{u}-K+2} \cup \cdots \cup f_{\dot{u}-1} \cup f_{\dot{u}} \tag{7}$$

where $\dot{u}$ is the current frame. An activity may be detected in Δ consecutive feature vectors, hence only one activity is considered. Accordingly, an activity a is considered to have started at frame $\dot{u} - K + 1$ and ended at frame $\dot{v} = \dot{u} + \Delta - 1$.

Activity A1 ("the speaker's face is visible") occurs when the speaker's face is visible in the image . Generally, A1 takes from several minutes to the whole sequence and the features used are the face location in x and/or y. Activity A2 ("the speaker is pointing to his/her right") occurs whenever the speaker points to his/her left, but it is only considered as pointing if the arm is stretched (or almost) to that side or the hand is pointing between left and up. Main features of A2 are the distances between the left hand and $F(t)$ and $To(t)$, and the left hand relative coordinates to $F(t)$. A2 takes 0.48-2 seconds. Activity A3 ("the speaker has moved to his/her left") occurs when the speaker moves to the left side, ie, in the positive direction of the X axis of the coordinate system. However, activity A3 restricts this motion to the movements that overcome one face width. Therefore, slight movements are not considered. A3 takes at least 0.44 seconds.

4.3 Video Recording

Video recording module includes two video recording operations. This module continuously records the full captured image (original resolution). Then, another video sequence showing the image region which contains speaker's body is recorded. This particular region is given by merging face, hands and torso regions into a single region that contains these four regions.

5 Experimental Results

In this section, we present the experimental procedure to collect the results and then we present the experimental results for the tracking algorithm, activity classifiers and for the whole system speed.

5.1 Experimental Procedure

The proposed system was evaluated on a database of 73 video sequences with a total duration of 2 hours, 20 minutes and 16 seconds. All video sequences have a 360x288 resolution at 25 fps and they were captured with a camera Canon XL2. 5 video sequences were used to measure the tracker's performance with different presentations and speakers (Group 1) and other 5 sequences were collected to measure the tracker's performance under different illumination conditions (Group 2). The remaining 63 sequences (Group 3) are the training and test set for activity recognition. Training and test sets of each activity were defined by 5-fold cross validation where each fold contained 2-4 sequences. Some sequences from Group 3 were used to train/test more than one activity classifier.

Sequences of Groups 1 and 2 were randomly selected from presentations containing different speakers and illumination conditions, respectively. Sequences of Group 3 were chosen after verifying that the tracker could track fairly well during the activities occurrence, providing good features for the training and testing. These sequences were also chosen because they contained more often each activity than the other presentation subsequences. Sequences from Group 3 were taken from presentations given in the same room and by the same speaker.

5.1.1 Tracker's Ground Truth

Tracker's Ground Truth (GT) was produced by two human experts who labeled 50 randomly chosen image samples in every sequence of Groups 1 and 2. The experts performed the labellings through an application developed by one of the authors, where each expert had to click over the samples to indicate the body part locations (face center, torso boundaries and hands centers). The non-author expert was told to choose the face and hands locations as their approximate center, and the torso limits as the boundaries between the speaker and the room background, according to his opinion. The experts could also label a body part as non-visible, so that it could be matched with a non detection from the tracking algorithm. GT for each frame was given by averaging the coordinates of each body part provided by the experts. Each expert took between 2 and 5 hours to label the complete data set.

5.1.2 Activity's Ground Truth

Labeling of activity examples for testing was performed by three human experts (including the author). The complete data sets of activities A1, A2 and A3 were labeled twice by two experts. Data set

of A2 was also labeled twice, but one of the labellings was shared by two experts. The time required to label the complete data set of each activity mainly depended on the sequence's length, the number of activity occurrences and on the expert's speed. To label the data set of activities A1-A3, each expert took 5-10 minutes (A1), 20-30 minutes (A2) and 13-25 hours (A3). The experts were taught to recognize each activity from image examples and from counterexamples given by us. Each expert has carefully observed the selected video sequences in VirtualDub and very often the experts had to forward and rewind the video frames in order to check the "correct" start and end frames of the activities. To label an activity occurrence the expert stored the sequence name, the occurrence label (positive or negative), the first frame where the activity is observed and the last frame of the activity. For training, we have selected a subset of the labeled examples where the tracker is able to track the speaker. This avoided providing features which did not correspond to an activity occurrence or at least reduced the number of wrong examples. Since the experts have only labeled positive examples, we have labeled some negative examples from Group 3. Activity A1 was trained with 10 positive (+) and 4 negative (-) examples. Activities A2 and A3 were trained with (12+, 12-) and (142+, 74-), respectively. The number of activities in the GT for A1-A3 is 12, 14 and 450, respectively.

GT was obtained by a selective merging of the experts' labellings (classifications). Firstly, the classifications where the experts' opinions completely differed were manually analyzed. Differences in classifications may be caused by different criteria or the expert may have unintentionally missed an activity occurrence. Those activity occurrences only labeled by an expert and which do not correspond to the activity characteristics were not considered. Secondly, the remaining classified occurrences of each video sequence and activity were merged and converted into discrete intervals which depict the starting and ending frames of an activity occurrence. Each interval is denoted by a_{GT}.

5.2 Tracking Performance

The tracking algorithm performance was measured through the average errors $e = (e_x, e_y)$, standard deviation of errors (σ_x, σ_y), covariance matrix and confusion matrix. The errors (in pixels) were measured in the 360x288 resolution and against the GT.

The error values for face tracking and system's default image resolution (90x72) are low ($e_x = 0.84 \pm 4.41$ and $e_y = 4.15 \pm 5.15$). For face and torso there are no false detections and their detection probability (DP) is between 76% and 86% for the tested image resolutions (360x288, 180x144, 90x72). DP of hands is between 52% and 83%, but it decreases to the range [50%, 64%] when the skin classifier is provided with dark or bright skin pixels. Average error values for torso's right and left boundaries are considerably low ($e_x = 0.23 \pm 6.07$ and $e_y = -5.84 \pm 5.7$ in 90x72), given that there are not two unique x coordinates for torso boundaries. Moreover, one of the test sequences contains many skin-like regions which difficult tracking and greatly increase the errors.

5.3 Activity Recognition Evaluation

Activity classifiers were evaluated with a confusion matrix and precision and recall rates. A GT activity occurrence $a_{GT} = [a_1, b_1]$ is correctly recognized if there is an automatic detection $a = [a_2, b_2]$ that satisfies

$$\frac{\#(a \cap a_{GT})}{\#(a \cup a_{GT})} \geq \mathscr{O}\%. \tag{8}$$

where $[a_1, b_1]$ and $[a_2, b_2]$ are the frame intervals where a_{GT} and a occurred, respectively. $\mathscr{O} = 50$ and $\#(\tilde{a})$ gives the cardinality of $\tilde{a}$. Best results of activity classifiers are shown in Table 1. Using Normal Bayes has shown many false positives (FP), while SVM with RBF kernel has shown to be more precise. The classifiers have correctly detected a high number of activity occurrences, but most of the occurrences were detected in several small frame intervals. Therefore, a smaller amount of a

satisfied (8) which greatly increased the FP and false negatives probabilities. By setting $\mathcal{O} = 1$, the classifiers' recall rates increase to 93.33 (activity A1), 90.00 (A2) and 86.23 (A3).

Table 1 Best results for each activity classifier. FP Prob. is the false positive probability and FN Prob. is the false negative probability. A1 to A3 identify the activities.

Activity / Rates (%)	FP Prob.	FN Prob.	Precision	Recall
A1	10.00	13.33	86.67	86.67
A2	43.61	35.00	32.50	65.00
A3	81.15	70.98	6.41	29.02

5.4 Speed Evaluation

The goal of tracking in real time was clearly achieved with a tracking algorithm that runs at 63.02 fps, 50.64 fps and 22.23 fps for 90x72, 180x144 and 360x288 image resolutions. The complete system runs at 43.37, 33.54 and 19.57 fps for the same resolutions.

6 Conclusions

This paper presents an intelligent system for video recording of classroom presentations. The detection probability of the speaker is above 76% and hand detection is above 52%, which is an important achievement, given the amount of skin-like regions around the speaker. We observed that the classifiers, in fact, recognize a high number of activity occurrences, but most of the activities are split into several frame intervals. Still, the proposed system presents a solid basis for further development. Future work should focus on enhancing face and hand detection robustness and on experimenting new classifiers' settings.

References

1. C. Busso, S. Hernanz, C. wei Chu, S. il Kwon, S. Lee, P. G. Georgiou, I. Cohen, and S. Narayanan, "Smart room: Participant and speaker localization and identification," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.
2. K. Bernardin, H. K. Ekenel, and R. Stiefelhagen, "Multimodal identity tracking in a smartroom," in *AIAI*, pp. 324–336, 2006.
3. B. Wu and R. Nevatia, "Tracking of multiple humans in meetings," in *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, (Washington, DC, USA), p. 143, IEEE Computer Society, 2006.
4. H. Ekenel, M. Fischer, Q. Jin, and R. Stiefelhagen, "Multi-modal person identification in a smart environment," in *Biometrics07*, pp. 1–8, 2007.
5. G. Potamianos and P. Lucey, "Audio-visual asr from multiple views inside smart rooms," in *Multisensor Fusion and Integration for Intelligent Systems, 2006 IEEE International Conference on*, pp. 35–40, September 2006.
6. V. Vilaplana, C. Martinez, J. Cruz, and F. Marques, "Face recognition using groups of images in smart room scenarios," in *ICIP06*, pp. 2069–2072, 2006.
7. I. McCowan, D. Gatica-Perez, S. Bengio, and G. Lathoud, "Automatic analysis of multimodal group actions in meetings," 2003.
8. R. Stiefelhagen, "Tracking focus of attention in meetings," in *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*, pp. 273–280, 2002.

9. T. C. C. Henry, E. G. R. Janapriya, and L. C. de Silva, "An automatic system for multiple human tracking and actions recognition in office environment," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '03)*, vol. 3, pp. 45–48, April 2003.
10. B. Ozer and W. Wolf, "A smart camera for real-time human activity recognition," in *Signal Processing Systems, 2001 IEEE Workshop on*, pp. 217–224, 2001.
11. A. Waibel, T. Schultz, M. Bett, M. Denecke, R. Malkin, I. Rogina, R. Stiefelhagen, and J. Yang, "Smart: the smart meeting room task at isl," in *Acoustics, Speech, and Signal Processing (ICASSP '03). 2003: IEEE*, pp. 752–755, 2003.
12. Z. Zhang, G. Potamianos, S. M. Chu, J. Tu, and T. S. Huang, "Person tracking in smart rooms using dynamic programming and adaptive subspace learning," in *ICME*, pp. 2061–2064, 2006.
13. C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *T-PAMI*, vol. 19, pp. 780–785, 1997.
14. K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *Information Theory, IEEE Transactions on*, vol. 21, no. 1, pp. 32–40, 1975.
15. D. Ramanan, D. A. Forsyth, and A. Zisserman, "Tracking people by learning their appearance," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 29, pp. 65–81, 2007.
16. H. Fujiyoshi and A. J. Lipton, "Real-time human motion analysis by image skeletonization," in *Proceedings of IEEE WACV98*, pp. 15–21, 1998.
17. J. Kovac, P. Peer, and F. Solina, "Human skin color clustering for face detection," in *Proc. The IEEE Region 8 EUROCON Int'l Conference*, vol. 2, pp. 144–148, 2003.
18. P. Viola and M. Jones, "Robust real-time object detection," in *International Journal of Computer Vision*, 2001.
19. M. Brand and V. Kettnaker, "Discovery and segmentation of activities in video," *PAMI*, vol. 22, pp. 844–851, August 2000.
20. H. Nait-charif and S. J. Mckenna, "Activity summarisation and fall detection in a supportive home environment," in *International Conference on Pattern Recognition*, pp. 323–326, 2004.
21. F. Lv and R. Nevatia, "Single view human action recognition using key pose matching and viterbi path searching," in *CVPR07*, pp. 1–8, 2007.
22. C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *Proc. ICPR*, pp. 32–36, 2004.
23. E. Ramasso, C. Panagiotakis, D. Pellerin, and M. Rombaut, "Human action recognition in videos based on the transferable belief model," *Pattern Anal. Appl.*, vol. 11, no. 1, pp. 1–19, 2008.
24. M. Al-Hames, C. Lenz, S. Reiter, J. Schenk, F. Wallhoff, and G. Rigoll, "Robust multi-modal group action recognition in meetings from disturbed videos with the asynchronous hidden markov model.," in *ICIP (2)*, pp. 213–216, IEEE, 2007.
25. A. F. Bobick, J. W. Davis, I. C. Society, and I. C. Society, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 257–267, 2001.
26. A. D. Wilson and A. F. Bobick, "Parametric hidden markov models for gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 884–900, 1999.
27. A. A. Efros, E. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *ICCV*, pp. 726–733, 2003.
28. J. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features," in *CVPR08*, pp. 1–8, 2008.
29. F. Bremond, M. Thonnat, and M. Zuniga, "Video understanding framework for automatic behavior recognition," *Behavior Research Methods*, vol. 3, no. 38, pp. 416–426, 2006.
30. T. Horprasert, D. Harwood, and L. S. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," in *ICCV Frame-Rate WS*, 1999.