Flexible human behavior analysis framework for video surveillance applications

Hindawi Publishing Corporation
International Journal of Digital Multimedia Broadcasting
Volume 2010, Article ID 920121, 9 pages
doi:10.1155/2010/920121

Research Article

Flexible Human Behavior Analysis Framework for Video Surveillance Applications

Weilun Lao,[1,2] Jungong Han,[1] and Peter H. N. de With[1,3]

[1] *Eindhoven University of Technology, Den Dolech 2, 5600MB Eindhoven, The Netherlands*
[2] *Guangdong Power Grid Company, 510620 Guangzhou, China*
[3] *Cyclomedia, 4180BB Waardenburg, The Netherlands*

Correspondence should be addressed to Weilun Lao, w.lao@tue.nl

Received 5 October 2009; Accepted 9 January 2010

Academic Editor: Ling Shao

We study a flexible framework for semantic analysis of human motion from surveillance video. Successful trajectory estimation and human-body modeling facilitate the semantic analysis of human activities in video sequences. Although human motion is widely investigated, we have extended such research in three aspects. By adding a second camera, not only more reliable behavior analysis is possible, but it also enables to map the ongoing scene events onto a 3D setting to facilitate further semantic analysis. The second contribution is the introduction of a 3D reconstruction scheme for scene understanding. Thirdly, we perform a fast scheme to detect different body parts and generate a fitting skeleton model, without using the explicit assumption of upright body posture. The extension of multiple-view fusion improves the event-based semantic analysis by 15%–30%. Our proposed framework proves its effectiveness as it achieves a near real-time performance (13–15 frames/second and 6–8 frames/second) for monocular and two-view video sequences.

1. Introduction

Visual surveillance for human-behavior analysis has been investigated worldwide as an active research topic [1]. In order to have automatic surveillance accepted by a large community, it requires a sufficiently high accuracy and the computation complexity should enable a real-time performance. In the video-based surveillance application, even if the motion of persons is known, this is not sufficient to describe the posture of the person. The postures of the persons can provide important clues for understanding their activities. Therefore, accurate detection and recognition of various human postures both contribute to the scene understanding. The accuracy of the system is hampered by the use of a single camera, in case of complex situations and several people undertaking actions in the same scene. Often, the posture of people is occluded, so that the behavior cannot be realized in high accuracy. In this paper, we contribute to improve the analysis accuracy by exploiting the use of second camera and mapping the event into a 3D scene model, that enables analysis of the behavior in the 3D domain. Let us now discuss related work from the literature.

1.1. Related Work. Most surveillance systems have focused on understanding the events through the study of trajectories and positions of persons using *a priori* knowledge about the scene. The Pfinder [2] system was developed to describe a moving person in an indoor environment. It tracks a single nonoccluded person in complex scenes. The VSAM [3] system can monitor activities over various scenarios, using multiple cameras that are connected as a network. It can detect and track multiple persons and vehicles within cluttered scenes and manage their activities over a long period of time. The real-time visual surveillance system W4 [4] employs the combined techniques of shape analysis and body tracking, and models different appearances of a person. This single-camera system detects and tracks groups of people and monitors their behaviors, even in the presence of partial occlusion and in outdoor environments. However,

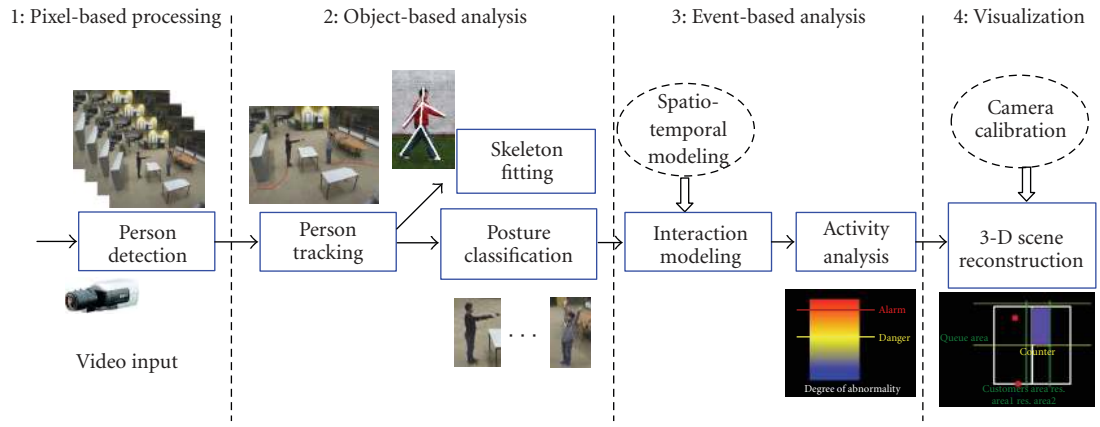


Figure 1: Block diagram of our four-level human motion analysis system.

the above systems generally suffer from the problem that they lack reliable continuous observation of people displacements. The monitoring performances of the above systems mainly rely on the detected trajectories of the concerned objects. Furthermore, the results are not sufficient for event analysis in some cases. As the local properties of the detected persons are missing, the developed systems lack the semantic recognition result of dynamic human activities. In this paper, we explore the *combination* of using trajectory, posture recognition, skeleton fitting, and 3D scene reconstruction in order to improve the semantic analysis of the human behavior. Furthermore, we apply the above techniques to a dual-camera system to improve the accuracy of event recognition.

1.2. 3-D Reconstruction. Scene reconstruction in 3D is a useful tool in semantic-event analysis, which is generally utilized in multimedia applications [5]. The accurate and realistic reconstruction in a virtual space can significantly contribute to the scene understanding, like crime-evidence collection and tactical analysis. Therefore, it is interesting to extend scene-reconstruction functionality in advanced surveillance applications, such as home-care monitoring and robbery-detection surveillance. The 3D scene reconstruction can be conducted to visualize the scene for further analysis. The 3D reconstruction is in fact a mapping of the 2D image data into a 3D real-world model. After the mapping from image to real world is performed, we can estimate the position and calculate the real speed of the persons involved in the video scene. Furthermore, the scene can be reconstructed by realistic 3D models by advanced modeling software to improve the reality. The above postprocessing step extends the framework with better visual presentation. In the application of a bank-robbery detection, for example, this extended processing plays a useful role in the crime-scene analysis, data retrieval, and evidence collection.

The principle of the mapping is basically a homography equation that describes the conversion of 2D point locations into 3D positions. For this purpose, we aim at finding a set of reference points that can be reliably detected. These reference points are used as an input for a multiparameter homography estimation. If sufficient reference points are available, the parameters of the homography can be computed. After this calibration, each input image can be mapped onto the 3D space using the computed homography. If we extend the system to a dual-camera setup, each of the cameras needs to be calibrated as described above. Both camera views are mapped onto the same 3D space. If one person is occluded in one camera view, his position can still be mostly determined by the second camera, so that a reliable 3D scene analysis can be conducted.

1.3. Research Objectives. To address the challenging problem of accurately analyzing human motion and summarizing events at high semantic level, we contribute in three aspects.

(i) A flexible framework is proposed to enable hierarchical human motion analysis. It can be utilized in surveillance applications with four-level analysis results using single or multiple cameras.

(ii) A 3D reconstruction scheme is introduced for scene understanding based on automatic camera calibration. The location and posture of persons are visualized in a 3D space after context knowledge is integrated. More specifically, the 2D-3D mapping provides a platform for a normalized motion configuration (i.e., location and speed) and scene visualization/analysis in the real world.

(iii) A fast scheme is proposed to detect different body parts in human motion. More specifically, for every individual person, features of body ratio, silhouette, and appearance are integrated into a hybrid model to detect body parts. The conventional assumption of upright body posture is not required.

In the sequel, we first present a system overview in Section 2 and then describe in detail the techniques for each level in Section 3. The experimental results on surveillance video are provided in Section 4. Finally, Section 5 draws conclusions.

2. Our Proposed System Framework

Our work aims at the object/scene analysis and behavior modeling of deformable objects. The framework captures the human motion, analyzes and demonstrates its gesture/activity, infers the semantic event exploiting interaction modeling, and performs the 3D scene reconstruction. The previous analysis is only possible if the scene and its objects are analyzed at various levels (e.g., background modeling, moving objects, event recognition, etc.). The block diagram of our multilevel event-analysis system is shown in Figure 1. The term multilevel refers to the four different conceptual levels: *pixel-based level, object-based level* (including trajectory estimation, posture classification, and skeleton fitting), *event-based level,* and *visualization level*.

(i) *Pixel-based level*. The background modeling and object detection are implemented. Each image within the video covering an individual human body is segmented to extract the "blobs" representing foreground objects. These detected blobs are refined afterwards to produce the human silhouette.

(ii) *Object-based level*. It performs trajectory estimation, posture classification, and skeleton fitting. We first track every moving person. Afterwards, a shape-based analysis is conducted to classify different posture types. Finally, a skeleton model is adaptively produced for every object.

(iii) *Event-based level*. Interaction relationships are modeled to infer a multiple-person event. This semantic analysis is thus responsible for the human activity recognition.

(iv) *Visualization level*. With the aim of 2D-3D mapping calibration, the 3D scene reconstruction is conducted to visualize the scene for further analysis. This level can be simple for home use, but advanced for professional use (e.g., after crime analysis in 3D).

The purpose of the framework is that it should be powerful and robust enough to facilitate a few different surveillance applications. To fulfill this objective, the semantic-level analysis should be of sufficiently high performance. In the sequel, home-care monitoring and the detection of a robbery for security surveillance are our key applications.

3. Techniques for Human Behavior Analysis

3.1. Trajectory Estimation. At the pixel-based level, the human silhouette is detected based on background subtraction. This general method can be used to segment moving objects in a scene, assuming that the camera is stationary and the lighting condition is fixed. To improve the blob segmentation, a shadow-removing approach [6] is used in our scheme. The false segmentation caused by shadows can be minimized by computing differences in a color space (RGB) that is less sensitive to intensity changes.

At the object-based level, person tracking (trajectory estimation) and posture classification are performed. In the trajectory-estimation step, we employ the broadly accepted mean-shift algorithm for tracking persons, based on their individual appearance model represented as a color histogram. When the mean-shift tracker is applied, we extract every new person entering the scene and calculate the corresponding histogram model in the image domain. In subsequent frames for tracking that person, we shift the person object to the location whose histogram is the closest to the previous frame. Afterwards, from our previous work [7], we have adopted the Double Exponential Smoothing (DES) operator to track moving persons. This filter runs approximately 135 times faster than the popular Kalman filter-based predictive tracking algorithm, with equivalent prediction performance. When the trajectory is obtained, we can estimate the position of the persons involved in the video scene. Therefore, we can conduct a body-based analysis at the location of the person in every frame.

Based on the results of trajectory estimation, the person action is classified into three types: running, walking, and standing. In the case of standing, the speed of the moving person is below a predefined threshold. Only in that case, posture classification is performed, which will be addressed in the next subsection.

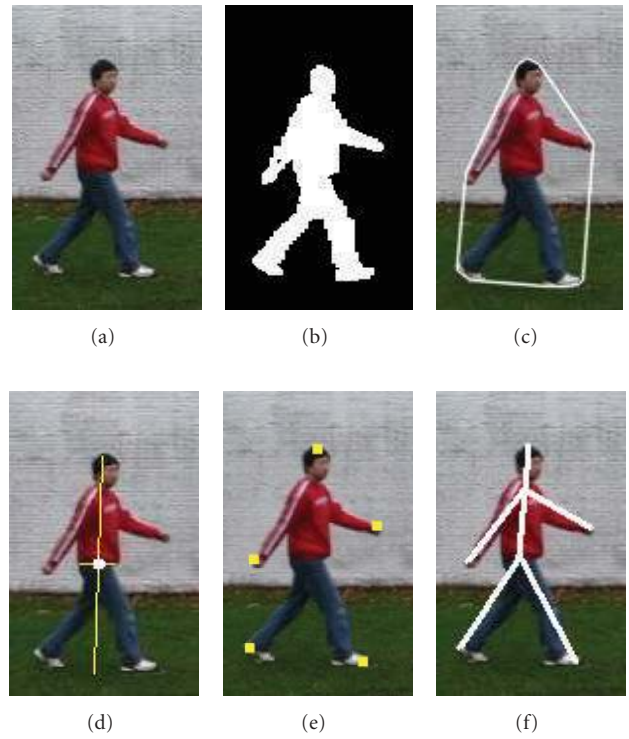


Figure 2: Procedure of skeleton-fitting processing: (a) original frame, (b) foreground segmentation (after shadow removal), (c) body modeling based on convex hull, (d) center-point estimation, (e) body-part location and, (f) skeleton construction in single-person motion.

3.2. Individual Posture Recognition with CHMM. We adopt a simple but effective shape descriptor to analyze the human

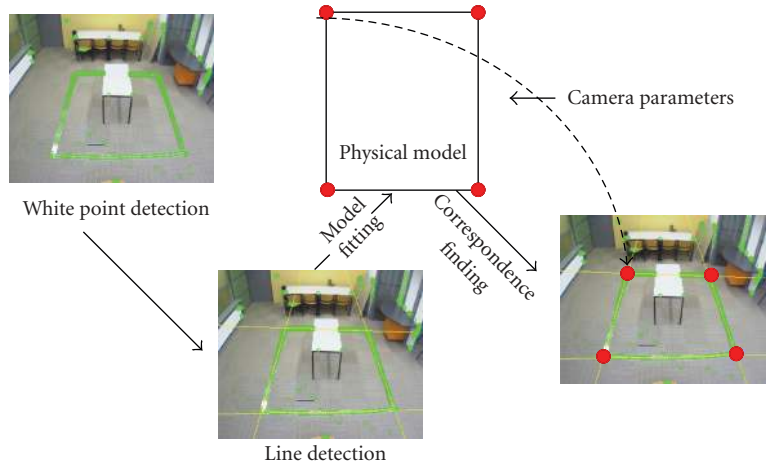


Figure 3: Example image of the corresponding homography based on camera calibration.

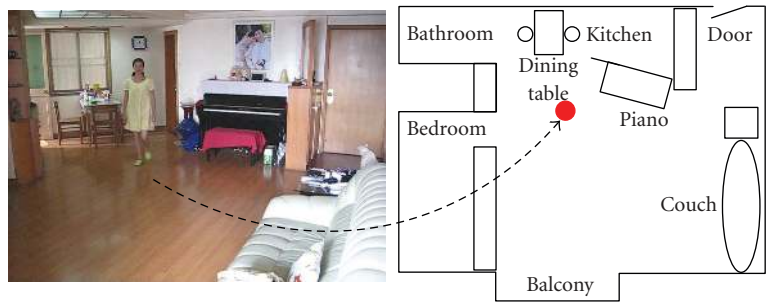


Figure 4: Example image of corresponding 2D/3D mapping in home-care monitoring.

silhouette prior to conducting the temporal modeling scheme of a Continuous Hidden Markov Model (CHMM) to recognize the posture type.

Individual posture classification is important for human-activity recognition. Firstly, we adopt a shape-based descriptor to analyze the human silhouette. Our posture classifier utilizes two features commonly used for object classification: area and the ratio of the bounding box attached to each detected object. This approach is simple but efficient, and it contributes significantly to the tracking and avoids a complex procedure for training data. The nonperson objects and image noise can be effectively removed. The disturbance generated from different person heights is also considered. We perform a training step regarding different heights in the scene before an adaptive threshold is applied. Finally, we can obtain the observed 2D feature vector of the silhouette.

Due to noise from segmentation errors, a single-frame recognition is not sufficiently accurate when we require general motion classification. The temporal consistency is required for a good posture recognition. Therefore, we adopt the HMM as our posture classifier, since it has proven to be an effective tool for sequential data processing. We use the Continuous Hidden Markov Model (CHMM) with left-right topology [8]. Suppose a CHMM has E states $Q = \{q_1, q_2, \ldots, q_E\}$ and F output symbols $V = \{v_1, v_2, \ldots, v_F\}$. It is fully specified by the triplet $\lambda = \{A, B, \pi\}$. Let the state at time step t be s_t; then the $E \times E$-state transition matrix $\mathbf{A}$ can be defined by

$$\mathbf{A} = \left\{a_{ij} \mid a_{ij} = P\left(s_{t+1} = q_j \mid s_t = q_i\right)\right\}, \quad 1 \le i,\ j \le E. \tag{1}$$

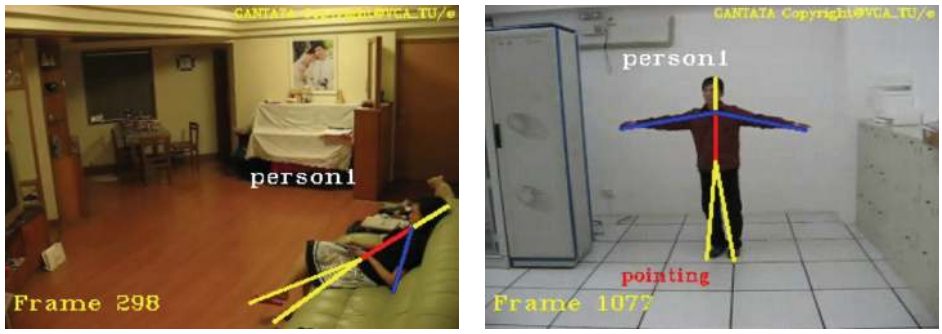
The $E \times F$-state output probability matrix $\mathbf{B}$ is defined as

$$\mathbf{B} = \left\{b_j(k) \mid b_j(k) = P\left(v_k \mid s_t = q_j\right)\right\}, \quad 1 \le j \le E, \quad 1 \le k \le F. \tag{2}$$

The initial state distribution vector π is specified as

$$\pi = \{\pi_i \mid \pi_i = P(s_1 = q_i)\}, \quad 1 \le i \le E. \tag{3}$$

We assign a CHMM model to each of the predefined posture types for the observed human body. Each CHMM is trained based on the Baum-Welch algorithm [8]. The learning process can calculate all parameters of the model using the training data. In other words, the triplet λ is obtained for each model. After having the models for each posture, we can proceed to implement the online testing. Given an observation sequence $Obv = \{Obv_1, Obv_2, \ldots, Obv_T\}$, we can calculate $P(Obv \mid \lambda_i)$, which is the probability of the observation sequence Obv given model i with λ_i. The probability $P(Obv \mid \lambda_i)$ can be obtained by using the forward algorithm [8]. After each model's output probability



(a) (b)

Figure 5: Example images of the skeleton-fitting result of human activity in two indoors events.

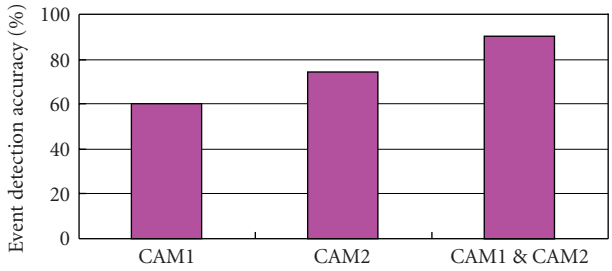


Figure 6: Event detection classification results based on single/multiple cameras.

is calculated, the model with maximum probability is chosen as the recognition result. We can therefore recognize the posture class C_T as being the one that is represented by the maximum probable model among K types, which is specified by

$$C_T = \arg\max_i P(Obv \mid \lambda_i), \quad 1 \le i \le K. \tag{4}$$

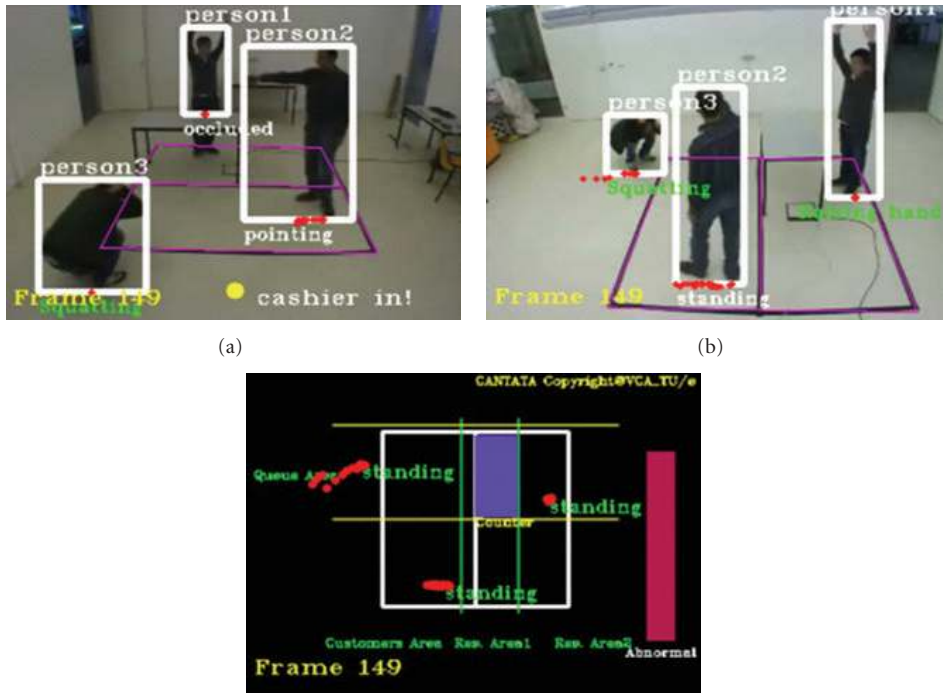
In our investigated case ($T = 20$, $K = 4$), every given posture is finally classified into one of the following types: *pointing, squatting, raising hands overhead,* and *lying*.

3.3. Skeleton Fitting. The purpose of skeleton fitting is to visualize the behavior of the person. To this end, we need to detect individual body parts at each frame. Since this has roots in earlier scientific research, we first briefly present an overview of this work below.

Accurate detection and efficient tracking of various body parts play an important role in presenting the human behavior. Existing fast techniques can be classified into two categories: appearance-based and silhouette-based methods. *Appearance-based* approaches [9, 10] utilize the intensity or color configuration within the whole body to infer specific body parts. They can simplify the estimation and collection of training data. However, they are significantly affected by the variances of body postures and clothing. For the *silhouette-based* approach [11–13], different body parts are located employing the external points detected along the contour, or internal points estimated from the shape analysis. The geometric configuration of each body part is modeled prior to performing the pose estimation of the whole human body. However, the highly accurate detection of body parts remains a difficult problem, due to the effectiveness of segmentation. Human limbs are often inaccurately detected because of the self-occlusion or occlusion by other objects/persons. Summarizing, both silhouette and appearance-based techniques do not offer a sufficiently high overall accuracy of body-part detection. Also, the assumption of upright posture is generally required. In our work, we do not need the assumption of an upright posture. We had to design a new algorithm because in the desired applications, persons are not always in an upright position. In the following, we summarize our algorithm that was reported first in [14].

We develop a fast scheme to detect different body parts in human motion. More specifically, for every individual person, features of body ratio, silhouette, and appearance are integrated into a hybrid model to detect body parts. The conventional assumption of upright body posture is not required. The skeleton-fitting processing step models the human motion by a skeleton model. The detailed procedure is illustrated in Figure 2. In the example of single-person motion, the input frame (Figure 2(a)) is first subject to shadow removal, and then segmented to produce a foreground blob (Figure 2(b)). Afterwards, the convex hull is implemented for the whole blob (Figure 2(c)). The dominant points along the convex hull are strong clues, in the case of single-person body-part detection. They infer the possible locations of the ending points of body parts, like head, hands, and feet. Here we employ a *content-aware* scheme to estimate the center point (Figure 2(d)), which is fundamentally used to position the human skeleton model. Meanwhile, dominant points along the convex hull are selected and refined to locate the head, hands and feet (Figure 2(e)). Finally, different body parts are connected to a predefined skeleton model involving a center point, where the skeleton is adapted to the actual situation of the person in the scene (Figure 2(f)).

3.4. Interaction Modeling. In multiperson events, the event analysis is achieved by understanding the interactions among



(a) (b)

(c)

Figure 7: Example of our simulated multicamera robbery-detection result.

people involved in the scene. The temporal constraints of two-person interactions are defined by two events in terms of causal and coinciding relations of the two persons' posture changes. The events are seldom instantaneous and often significantly rely on the temporal order and relationship of their subevents (the individual posture). We introduce appropriate spatial and temporal constraints for each of the various two-person interaction patterns as domain knowledge. The satisfaction of specific spatial/temporal constraints contributes to the semantic recognition of the interaction. Therefore, the event-level recognition is characterized by the integration of domain-specific knowledge, whereas the object-level recognition is more closely related to the pure motion of a human body.

In order to represent temporal relationships of subevents, we apply the temporal logic based on interval algebra, as used in [15]. Seven temporal relationships are indicated from the set of $TR = \{after, meets, during, finishes, overlaps, equal, starts\}$. These keywords can link different sub-events after the individual event is analyzed. In this way, the scene becomes a chain of sub-events which are linked by the previously mentioned key words. To describe the semantic meaning of the scene, we apply heuristic rules. For example, in the application of a bank-robbery detection, the heuristic rules are based on expert knowledge. In our investigated case of robbery detection [14], the posture "pointing" is a key reference posture. It can significantly infer the robbery event. Other postures are also estimated to improve the recognition accuracy based on specific temporal constraints. Suppose person A has an action labeled as "pointing", person B is detected to be "raising both hands" and person C is "squatting" during the sub-event from person A, we can infer that a robbery actually occurs. After performing the interaction modeling, we are able to calculate the degree of abnormality. If the degree value is above a predefined threshold, the surveillance system will trigger the alarm signal (e.g., when a detected robbery event happens). The advantage of using such a metric is that an abnormal situation can raise the degree value to alert security people without signaling an alarm. The degree value can thus be used as a preventive measure, rather than alarming when the actual robbery takes place.

3.5. 3D Scene Reconstruction. Based on the automatic or manual camera calibration, we can implement the 2D-3D mapping. In other words, the image contents can be described in a 3D world domain. Furthermore, the real scene is reconstructed in a virtual space. This 3D scene reconstruction is useful for after-crime analysis and it contributes to the crime-evidence collection.

The objective of the camera calibration is to provide a geometric transformation that maps the points in the image domain to the real-world coordinates. An example of the scene reconstruction is visualized in Figure 3. In our system, we analyze the human behavior based on the person's trajectory and/or speed on the ground, so that the height information of the human is not required. Since both the ground and the displayed images are planar, the mapping between them is a homography, which can be written as

Table 1: The detection results for human activity recognition in home-care monitoring.

	In kitchen	Sitting at dining table	Sitting on couch	Playing piano	To balcony	In bedroom	In bathroom	Enter/Leave by door
In kitchen	14/16	2/16	0	0	0	0	0	0
Sitting at dining table	0	8/8	0	0	0	0	0	0
Sitting on couch	0	0	10/10	0	0	0	0	0
Playing piano	0	1/8	0	7/8	0	0	0	0
To balcony	0	0	0	0	10/10	0	0	0
In bedroom	0	0	0	0	2/12	10/12	0	0
In bathroom	0	1/7	0	0	0	0	6/7	0
Enter/Leave by door	0	0	0	0	0	0	0	5/5

a 3×3 transformation matrix H, transforming a point $p = (x, y, z)^T$ in image coordinates to the real-world coordinates $p' = (X, Y, Z)^T$ with $p = Hp'$, which is equivalent to

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}. \quad (5)$$

The transformation matrix H can be calculated from four points whose positions are known both in the real world and in the image. In our previous work [7], we have developed an automatic algorithm to establish the homography mapping for analyzing a tennis video, where the court lines and their interaction points are identified in the image. Such lines and points are related to the lines and points in a standard tennis court. Therefore, the homography mapping described in (5) can be established after the correspondences are found. This approach has been adopted in our surveillance system. The basic idea is to manually put four white lines forming a rectangular on the ground (see Figure 3). We have measured the length of each line in the real world, thereby defining their coordinates in the real-world domain. Afterwards, the algorithm proposed in our previous work can be applied for calculating parameters of the homography mapping. The complete algorithm comprises four steps, which are white-pixel detection, line detection, finding intersection points and calculating the parameters. For more details, we refer to an earlier publication [7].

After performing the mapping from the image to real world, we can estimate the position and calculate the real speed of the persons involved in the video scene. The label of walking or standing can be therefore assigned to an individual person. Furthermore, the scene can be reconstructed in 3D space. The above post-processing step extends our framework with better visual presentation and scene understanding. For instance, the 3D location of every moving person infers his actual behavior in the application of home-care monitoring. In the application of a bank-robbery detection, this extended processing is useful in the crime-scene analysis, data retrieval and evidence collection.

4. Experimental Results

We have trained our framework using 10 video sequences of various single/multiperson motion (15 frames/s) in both home-care and robbery-event scenarios. Then we have used 15 similar sequences for testing. On the frame basis, we have obtained a 98% accuracy rate on person detection, 95% detection rate on person tracking (where the criterion is that at least 70% of the human body is included in the detection window), and 82% detection rate on posture classification (in the robbery-event scenario).

4.1. Single-Camera Experiment: Home-Care Monitoring. Based on the trajectory estimation, we can calculate the speed and estimate the location of each individual person. We have conducted the experiment in our first case study on home-care monitoring. The experimental videos were captured in an apartment involving 6 persons (with different gender, height, age and clothes). The length of video sequences is more than 2 hours. The layout of the apartment is demonstrated in Figure 4. Based on the detected location and speed, the human daily activity is classified into 8 types (in kitchen, sitting at dining table, sitting on couch, playing piano, to balcony, in bedroom, in bathroom, and enter/leave by door). The classification results of activity recognition (involving the detected sequence numbers and total test sequence numbers, zero means that no corresponding sequence is detected) are summarized in Table 1. In our experiments, the ground truth of body-part locations were manually obtained. The maximum tolerable errors in the evaluation is set to 20 pixels. The skeleton model is further reconstructed on the individual body. Two examples of such a modeled presentation are portrayed by Figure 5.

4.2. Dual-Camera Experiment: Robbery-Event Detection. To further combat the problem of occlusion, multiple cameras are employed for capturing the same scene from different angles. We have conducted a dual-camera experiment in our second case study on a robbery-event detection. We analyze both camera views and combined the semantic of both views into one degree of abnormality. Currently, an OR logic operator is applied to link the two viewpoints at the level of the abnormal-event detection. Two different event types (normal and abnormal) are defined based on

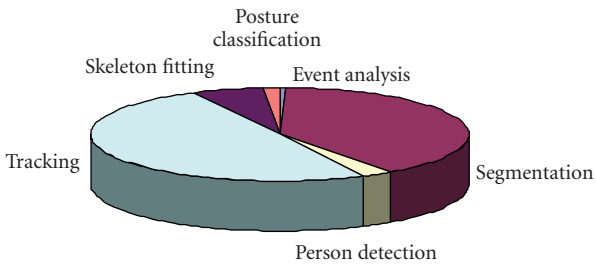


Figure 8: System performance: average time-consuming percentage of each module.

domain knowledge. The detection accuracy for each camera and the combination are shown in Figure 6. It proves that the dual-camera scheme effectively improves the event-based semantic analysis. Figure 7 shows a detection example of a simulated bank-robbery event. The position of every person is visualized after trajectory estimation. The postures are estimated and the semantic event is highlighted after interaction modeling from two different viewpoints (Figures 7(a) and 7(b)). The camera calibration is performed and the 2D-3D mapping is visualized. The degree of abnormality is also calculated and shown in Figure 7(c). Although the posture pointing is not recognized in one camera (see Figure 7(b)), it is correctly recognized in the other camera (see Figure 7(a)). The robbery event is successfully detected afterwards.

4.3. System Performance. Our system performance was tested by video sequences at 640 × 480 resolution (VGA), with a P-IV 3-GHz PC. Results show that our system fulfils the real-time requirement, as 13–15 frames/second and 6–8 frames/second are obtained for monocular and two-view video sequences, respectively. Figure 8 presents the average cycle-consumption percentage of every module. We have assumed that camera calibration is an off-line process taking place first. As can be noticed, foreground/background segmentation and tracking modules consume most computing cycles.

5. Conclusion

We have proposed a layered framework that enables multilevel human motion analysis, featuring layers at pixel, object, event and visualization level. The framework captures the human motion, classifies its posture, generates a fitting skeleton model after body-part detection, infers the semantic event exploiting interaction modeling, and performs the 3D scene reconstruction. We have applied this framework in a single-camera setup and a dual-camera setup. In the last case, it is possible to benefit from the extra view in case of occlusion and it may also add to after-crime analysis. This extension of multiple-view fusion improves the event-based semantic analysis by 15%–30%.

The framework was evaluated for two applications, a home-care monitoring case and a robbery-detection case. The practical results have shown that our framework can be used for various surveillance cases. The choice of using single or multiple cameras is basically independent on the type of surveillance applications and it is more ruled by the quality requirements or the occurrence of occlusions. Performance evaluations have shown that the framework is efficient and achieves a fast performance (13–15 frames/second and 6–8 frames/second) for monocular and two-view video sequences. Therefore, it can be used in an embedded system implementation.

We are improving the effective modeling of multi-person interaction, in order to obtain a probability-based inference engine. The self-occlusion problem is not yet thoroughly tackled at the current stage. Thus we intend to integrate an effective occlusion-handling module, which was reported in [16] to improve the motion-analysis robustness.

References

[1] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man and Cybernetics Part C*, vol. 34, no. 3, pp. 334–352, 2004.

[2] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: realtime tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 780–785, 1997.

[3] R. T. Collins, A. J. Lipton, T. Kanade, et al., "A system for video surveillance and monitoring," Tech. Rep. CMU-RI-TR-00-12, CMU, Pittsburgh, Pa, USA, 2000.

[4] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809–830, 2000.

[5] J. Han, D. Farin, and P. H. N. de With, "A real-time augmented-reality system for sports broadcast video enhancement," in *Proceedings of the ACM International Multimedia Conference and Exhibition*, pp. 337–340, Augsburg, Germany, 2007.

[6] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773–780, 2006.

[7] J. Han, D. Farin, P. H. N. de With, and W. Lao, "Real-time video content analysis tool for consumer media storage system," *IEEE Transactions on Consumer Electronics*, vol. 52, no. 3, pp. 870–878, 2006.

[8] L. R. Rabiner, "Tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[9] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, pp. 734–741, 2003.

[10] S. Park and J. K. Aggarwal, "Simultaneous tracking of multiple body parts of interacting persons," *Computer Vision and Image Understanding*, vol. 102, no. 1, pp. 1–21, 2006.

[11] H. Fujiyoshi, A. J. Lipton, and T. Kanade, "Real-time human motion analysis by image skeletonization," *IEICE Transactions on Information and Systems*, vol. E87, no. 1, pp. 113–120, 2004.

[12] C.-C. Yu, J.-N. Hwang, G.-F. Ho, and C.-H. Hsieh, "Automatic human body tracking and modeling from monocular video sequences," in *Proceedings of the IEEE International Conference*

on Acoustics, Speech and Signal Processing (ICASSP '07), vol. 1, pp. I917–I920, Honolulu, Hawaii, USA, 2007.

[13] P. Peursum, H. H. Bui, S. Venkatesh, and G. West, "Robust recognition and segmentation of human actions using HMMs with missing observations," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 13, pp. 2110–2126, 2005.

[14] W. Lao, J. Han, and P. H. N. de With, "Fast detection and modeling of human-body parts from monocular video," in *Articulated Motion and Deformable Objects*, vol. 5098 of *Lecture Notes in Computer Science*, pp. 380–389, Springer, Berlin, Germany, 2008.

[15] J. F. Allen and G. Ferguson, "Actions and events in interval temporal logic," *Journal of Logic Computation*, vol. 4, pp. 531–579, 1994.

[16] J. Han, M. Feng, and P. H. N. de With, "A real-time video surveillance system with human occlusion handling using nonlinear regression," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '08)*, pp. 305–308, Hannover, Germany, 2008.