An approach to the detection of bank robbery acts employing thermal image analysis

Conference Paper · January 2013

CITATION
1

READS
228

2 authors:

Maciej Szczodrak
Gdansk University of Technology
44 PUBLICATIONS **214** CITATIONS
SEE PROFILE

Grzegorz Szwoch
Gdansk University of Technology
48 PUBLICATIONS **219** CITATIONS
SEE PROFILE

An Approach to the Detection of Bank Robbery Acts Employing Thermal Image Analysis

Maciej Szczodrak, Grzegorz Szwoch
Gdansk University of Technology
Multimedia Systems Department
80-233 Gdansk, Poland, Narutowicza 11/12
szczodry@sound.eti.pg.gda.pl

Abstract— **A novel approach to the detection of selected security-related events in bank monitoring systems is presented. Thermal camera images are used for the detection of people in difficult lighting conditions. Next, the algorithm analyses movement of objects detected in thermal or standard monitoring cameras using a method evolved from the motion history images algorithm. At the same time, thermal images are analyzed in order to detect regions of concentrated energy. The detection results are combined in order to detect unusual situations such as people gathered in a small area during the assault. Performance of the proposed system was verified employing a set of recordings of a simulated bank robbery, prepared by the authors in a real bank branch site.**

Keywords-monitoring systems; thermal images; object detection; event detection

I. INTRODUCTION

Intelligent and automated surveillance systems for detection of important security threats in online mode are a current topic in scientific research and practical security applications [1]. Modern systems installed in banks and other indoor protected areas are able to detect simple events such as unauthorized access to restricted zones, they are able to count people in the queue, etc. However, security service officers responsible for introducing similar systems to bank offices informed the authors of this paper that there is a need for detection of more complex situations, mostly related to bank robbery, e.g. when customers inside the building are forced to raise their hands and gather (laying, kneeling, sitting) on the floor. Moreover, currently used systems are mainly based on standard video cameras only, which limits the accuracy of object and event detection due to low light, obstacles, etc. Therefore, the authors aimed to develop a system based on analysis of both video and thermal images for automatic detection of a bank robbery. It is expected that thermal cameras will provide complementary data for improving the detection accuracy. Because of the complexity of the described problem, this paper presents a first iteration of the proposed system only, as well as some preliminary evaluation results.

The idea of using thermal and video cameras for event detection has been utilized by several researchers. Davis and Sharma described a method of combining contours extracted from thermal images and color information from visible image within these contours for person detection [2]. Contours and blobs are combined using the watershed and A* search algorithms. Han and Bhanu presented the technique of merging thermal and visible image based on correspondence, employing genetic algorithms for human silhouette detection in indoor scenarios [3]. Putz et al described an approach to fusion of visible and thermal images, after acquisition by sensors, using Laplacian pyramids or Fast and Adaptive Bidimensional Empirical Mode Decomposition (FABEMD) algorithm implemented on FPGA [4] In work by Torresan et al, detection and tracking of objects are done in each modality and then correspondences between objects are found with master-slave approach. The criterion of choosing master is a confidence of object detection compared between modalities [5]. The authors of this paper also presented a method of fusion of the background subtraction results calculated separately for each modality [6]. The results of our experiments for outdoor scenes with various lighting conditions indicated an increased accuracy of object detection when both modalities were applied.

None of the existing solutions were found suitable for the specific task of bank robbery detection. Hence, a novel solution to this problem is presented in the paper. The details of the proposed algorithm are presented in Section II and some results of the preliminary tests are shown in Section III.

II. THE EVENT DETECTION SYSTEM

The system is organized as follows. First, moving objects (e.g. persons in the bank office) are detected in camera images. In order to improve object detection accuracy, thermal cameras are used for this task. Contours of moving objects are extracted from the object detection results. Next, movement of the detected objects is analyzed by means of detection of local movement directions and object splitting or merging. At the same time, regions of concentrated energy are detected in thermal images. Finally, the results are analyzed by the decision module which detects whether a specific event took place. This paper focuses on the object detection and tracking algorithms. The decision system was intentionally simplified in experiments described here in order to verify the validity of the algorithms described in this paper. Development of a more complex decision system lies outside of scope of this paper and will be performed in the future experiments.

Research is subsidized by the European Commission within FP7 project INDECT, Grant Agreement No. 218086

A. Object detection in thermal images

Analysis of camera images begins with the detection of moving objects. As a result, a binary mask is created, marking the foreground/background pixels with values of one and zero, respectively. This mask is cleaned using morphological processing. In video cameras, the mask is usually obtained by means of background subtraction [1]. In thermal cameras, two approaches are possible. A thermal image may be thresholded in order to separate high temperature components (humans) from the static, low temperature background. Due to variations in ambient temperature and thermal sensor settings, the threshold has to be selected dynamically, e.g. on the basis of histogram evaluation [7]. The second approach is to apply the background subtraction algorithm to a single-channel thermal image and construct a dynamic background model, eliminating the need for threshold calculation. The latter approach was used in the experiments described in this paper.

Background subtraction is performed with the Gaussian Mixture Model algorithm, as proposed by Stauffer and Grimson [8]. Each pixel in the grayscale (white-hot) thermal map is represented with a weighted sum of Gaussians:

$$P(x_t) = \sum_{k=1}^{K} w_{k,t} \cdot \eta\left(x_t, \mu_{k,t}, \sigma_{k,t}\right) \tag{1}$$

where x_t is the current pixel value, μ is the mean background pixel value, σ is its standard deviation, w is the weight, K is the number of Gaussians (typically, K = 3). Image pixels are assigned to the foreground if a Gaussian is found for which:

$$\left|x_t - \mu_{k,t}\right| \le T \cdot \sigma_{k,t} \tag{2}$$

where T is the detection threshold (usually T = 2.5). If a matched Gaussian is found, it is updated as follows:

$$\begin{aligned} w_t &= w_{t-1} + \alpha\left(1 - w_{t-1}\right) \\ \mu_t &= \mu_{t-1} + \rho\left(x_t - \mu_{t-1}\right) \\ \sigma_t^2 &= \sigma_{t-1}^2 + \rho\left[\left(x_t - \mu_t\right)^2 - \sigma_{t-1}^2\right] \end{aligned} \tag{3}$$

where α is the weight update factor and ρ is the learning factor. The remaining Gaussians have their weights decreased:

$$w_t = \left(1 - \alpha\right) w_{t-1}. \tag{4}$$

If no match was found, a new Gaussian is created and it replaces the one with the lowest weight.

After the binary mask of foreground/background pixels is constructed, it is post-processed using the morphological opening and closing operations in order to clean the mask from noise and small gaps. Finally, contours of the moving objects (groups of foreground pixels) are extracted from the mask using the border following algorithm [9].

The advantage of using the approach based on background subtraction instead of a simple binarization is that the same procedure may be applied to video cameras in case the thermal cameras are not available. Moreover, a dual setup of a thermal and a video camera mounted close to each other may be used to improve the object detection accuracy in difficult conditions [6].

B. Analysis of object movement

Contours of moving objects detected in the previous stage are usually tracked with Kalman filters [10] or similar solutions. However, in the described setup, Kalman filters did not provide satisfactory accuracy due to frequent occlusions of persons moving close to the camera. Therefore, we propose a different approach inspired by the motion history images (MHI) algorithm designed by Davis [11]. The original method stores a number of timestamped contours of moving objects in the MHI. Detection of gradients in the MHI allows for determining directions of local movement. For the purpose of our system, we have modified this method in a following way. Let $\mathbf{M}_t$ denote the mask (background subtraction result) from the image frame t. Instead of using $\mathbf{M}_t$ for updating the MHI directly, two difference mask images are created. The foreground transition mask (TM) $\mathbf{F}_t$ marks the pixels that belonged to the background in the previous frame and now are assigned to the foreground:

$$\mathbf{F}_t = \mathbf{M}_t \times \sim\mathbf{M}_{t-1} \tag{5}$$

where ~ denotes the negative of a binary image. Similarly, the background TM $\mathbf{B}_t$ indicates the pixels that changed the state from the foreground to the background:

$$\mathbf{B}_t = \sim\mathbf{M}_t \times \mathbf{M}_{t-1} \tag{6}$$

Object movement is modeled using two MHIs (the foreground and the background one) which are updated with **F** or **B** mask on a frame-by-frame basis. Detection of local movement is performed similarly to the original MHI method. First, each MHI is convoluted with the Sobel filter in two dimensions separately. For each pixel, spatial derivatives $F_x(x,y)$ and $F_y(x,y)$ are computed. Local gradient orientation for the pixel (x, y) is calculated as [12]:

$$\phi(x,y) = \arctan\left(\frac{F_y(x,y)}{F_x(x,y)}\right) \;\Big|\; g_{\min} \le \sqrt{\left(F_x(x,y)\right)^2 + \left(F_y(x,y)\right)^2} \le g_{\max} \tag{7}$$

where $g_{\min}$ and $g_{\max}$ are the minimum and maximum magnitude of motion gradient that is taken into account (too large and too small gradients have to be filtered out). Next, local movement is computed by averaging and normalizing gradient orientations for pixels belonging to contours extracted from $\mathbf{F}_t$ and $\mathbf{B}_t$ masks.

In the next step, a global movement direction is calculated for each object (or a group of objects) represented by a contour extracted from the background subtraction mask. These contours are related with these obtained from the combined

($\mathbf{M}_t + \mathbf{B}_t$) mask. For each contour, N_F and N_B local orientations φ_F and φ_B are found from the analysis of masks $\mathbf{F}_t$ and $\mathbf{B}_t$, respectively. The global contour orientation is calculated by averaging the local orientations:

$$\varphi_G = \operatorname{atan}2\left(\frac{\sum_{i=1}^{N_F} a_{F,i}\sin\varphi_{F,i} + \sum_{j=1}^{N_B} a_{B,j}\sin\varphi_{B,j}}{\sum_{i=1}^{N_F} a_{F,i} + \sum_{j=1}^{N_B} a_{B,j}}, \frac{\sum_{i=1}^{N_F} a_{F,i}\cos\varphi_{F,i} + \sum_{j=1}^{N_B} a_{B,j}\cos\varphi_{B,j}}{\sum_{i=1}^{N_F} a_{F,i} + \sum_{j=1}^{N_B} a_{B,j}}\right) \quad (8)$$

where a is the pixel area of a given contour fragment. Moreover, global orientations calculated for each tracked contour are time-averaged in order to obtain a short-term trend of movement direction:

$$\overline{\varphi}_{G,t} = \operatorname{atan}2\left(\gamma\sin\overline{\varphi}_{G,t-1} + (1-\gamma)\sin\varphi_{G,t}, \gamma\cos\overline{\varphi}_{G,t-1} + (1-\gamma)\cos\varphi_{G,t}\right). \quad (9)$$

During object tracking, contours of objects constantly merge and split due to occlusions and fragmentation. However, we are not interested in tracking every individual object, so we simply record the number of merges c and splits s that a given contour took part of. Therefore, we describe each tracked contour with a vector:

$$\mathbf{v}_t = \left[x_t, y_t, w_t, h_t, m_t, \varphi_{G,t}, \overline{\varphi}_{G,t}, c_t, s_t\right] \quad (10)$$

where (x, y) is the position of mass center of the contour, w and h are the width and height of the contour's bounding box, m is the contour mass – a number of foreground pixels and the remaining symbols were described before.

C. Analysis of energy concentration

Another part of the algorithm realizes the detection of people grouping on the floor using an approach similar to the energy method described by Zhong et al [13]. In the original method, energy is considered as a kinetic one and calculated as a pixel intensity difference between image frames in visible spectrum. Since information about temperature is available directly in thermal camera images, the authors propose different approach. Thermal energy is used, since the direct measure of the thermal energy is the object temperature. The heat radiation is described by the Stefan-Boltzmann law: $\Phi = \varepsilon\sigma T^4$, where ε is emissivity, σ – Stefan-Boltzmann constant and T – temperature in Kelvins. An assumption that temperature of humans is different from the ambient temperature is justified in the described scenario, since the operation hall in bank is a closed space, likely equipped with an air conditioning system.

In normal human pose, thermal energy is distributed in a different way than in pose of kneeling on the floor. The energy in the thermal image is represented by the pixel intensity. Considering a $M \times N$ image with intensity of each pixel $I(x,y)$, using the white-hop mapping (higher pixel intensity means higher temperature), the following functions are calculated:

$$\begin{aligned} f_{MAX}(x) &= \max(I(x,y_0), I(x,y_1), \ldots, I(x,y_N)) \\ f_{MAX}(y) &= \max(I(x_0,y), I(x_1,y), \ldots, I(x_M,y)) \end{aligned} \quad (11)$$

$$\begin{aligned} f_{MIN}(x) &= \min(I(x,y_0), I(x,y_1), \ldots, I(x,y_N)) \\ f_{MIN}(y) &= \min(I(x_0,y), I(x_1,y), \ldots, I(x_M,y)) \end{aligned}. \quad (12)$$

In each frame, the T-percent width ranges are calculated for functions $f_{MAX}(x)$ and $f_{MAX}(y)$ (11). Typically, $T = 66$ and the value depends on image contrast. The outcome of this operation is a set of boundary coordinates. Boundaries of each T-percent width are denoted as $x_{1,L}, x_{1,H}, x_{2,L}, x_{2,H}, \cdots$ and $y_{1,L}, y_{1,H}, y_{2,L}, y_{2,H}, \cdots$. L index means lower, H – higher boundary.

In the next step, the modified centroid of each region embraced by coordinates $x_{i,L}, x_{i,H}, y_{i,L}, y_{i,H}$ is calculated according to equation:

$$c_i(x,y) = \left(x_{i,L} + \frac{x_{i,H} - x_{i,L}}{2}, y_{i,L} + \frac{y_{i,H} - y_{i,L}}{2}\right). \quad (13)$$

The value of y coordinate of each obtained modified centroid (13) is compared to the threshold value T_G. The threshold is dependent on the camera setup and is found experimentally. Number of regions with centroid below T_G is calculated with the equation:

$$N = \sum_i c_i^G \quad (14)$$

where:

$$c_i^G = \begin{cases} 1 & \text{if } c_i(y) < T_G \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

If N is equal to total number of obtained centroids, the thermal energy of objects is located in the area adjacent to the floor.

D. Detection of important events

In order to detect the act of forcing a group of people to kneel on the floor, both approaches are combined for improving the detection accuracy. First, regions of concentrated energy are detected, as described in Subsection C. In the next stage, these candidate regions are verified using the contour tracking results. The event is detected if the following

conditions are met: (i) a contour has a sufficiently large number of merges *c*, indicating that a number of persons formed a group, (ii) the mass of the contour is sufficiently large, (iii) the merged contour is situated inside the region of concentrated energy, (iv) the averaged direction of movement is directed towards the base point of the high energy region, and (v) the region of the concentrated energy remains in the same area. These conditions are tested using the defined threshold values. As it was previously mentioned, this decision system is considerably simplified for the purpose of the initial tests. For example, in real system, conditions (i-ii) should not be applied in some cases, e.g. if a number of persons is very small. However, such a simple decision module was sufficient for the described experiments.

III. EXPERIMENTS

In order to perform a quantitative analysis of the system performance, benchmark data are required. Since no standard dataset for scenario presented in this paper was found, and real recordings from the bank monitoring systems were not available due to security constraints, self-made recordings were used in tests (a more elaborated performance evaluation will be made in the future). Four scenes of the simulated bank robbery were played by actors in the real bank scenery. Data were recorded using a thermal image camera (AXIS Q1910, 384×288 px, 8.33 fps), and additionally with a video camera (AXIS P1346, 1920×1200 px, 15 fps). The size of the room was 20 × 15 × 3.5 m. In the played scenarios, a robber entered the bank and forced all persons to gather and kneel down with their hands put on their heads.

Fig. 1 presents the example results of energy distribution analysis at the moment the persons were kneeling on the floor. The scene depicted there was recorded for the preliminary tests, with a small number of actors. It can be observed that the thermal energy is concentrated in the area close to the ground. Moreover, functions described in Subsection II-C are shown in Fig. 1. In this simple case, the position of each person can be estimated. The detected regions of concentrated energy were treated as candidates for the event detection.

Verification was made using the object detection and tracking modules. An example of the obtained results is presented in Fig. 2. The robber orders people to gather, causing the contours of the individual objects to merge into a single, large contour with a high value of merges *c*. At the same time, averaged local movement gradients indicate that the objects move towards the same area, and then all individuals lower their position (reflected by a downward local movements in the upper part of the object contour). Since the position of the merged contour and its direction of movement is consistent with the previously detected region of high energy, the decision about a potentially dangerous event is made.

Fig. 3 presents the example results of energy distribution analysis at the moment the customers in the bank were already grouped on the floor. Function in the right part of the picture (black line) show maxima of thermal energy. Center of gravity of high energy region is located in lower part of image, below the threshold TG. In this case N is equal to total number of centroids. Moreover it should be stressed that thermal image is

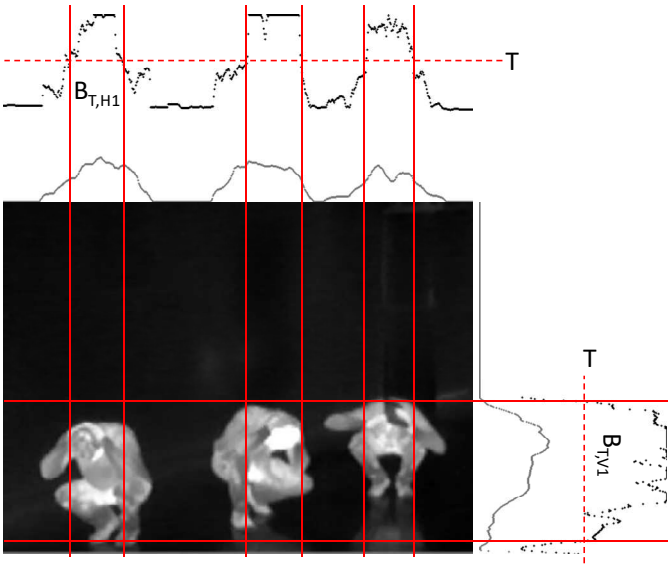

Figure 1. Example of detection of people kneeling (preliminary experiment). *T*-percent ranges are shown

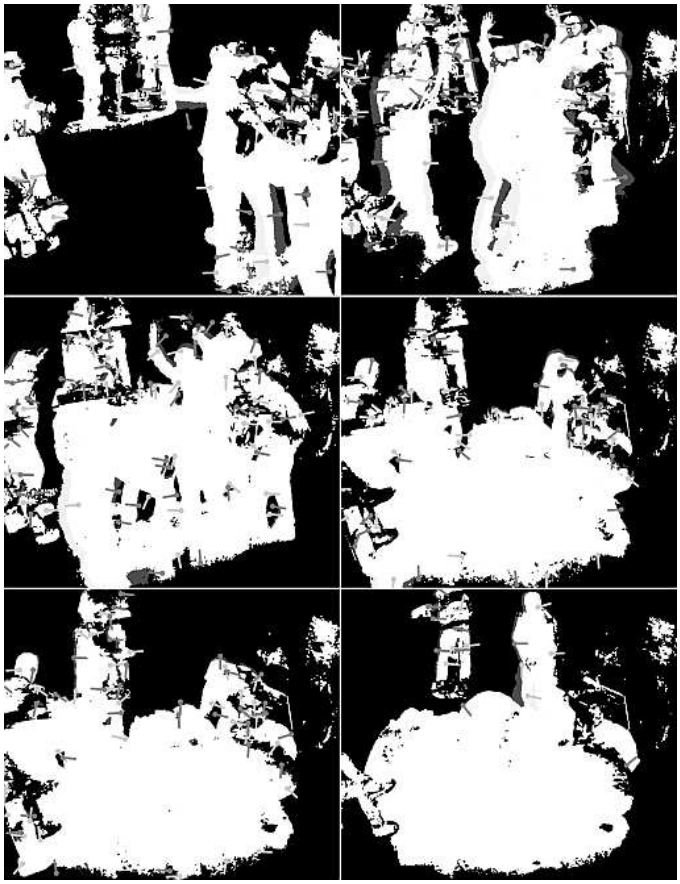

Figure 2. Example of the scene analysis using the contour tracking procedure in a scene when persons are forced to kneel on the ground in a group. White pixels denote contours of moving objects, with parts belonging to F and B maked with light and dark gray color, respectively. Lines denote the local directions of movement (lines begin with dots). The large contour visible in the bottom right image is related to the detected region of concentrated energy.

distorted by bank queue display system which can be seen in the foreground of the scene.

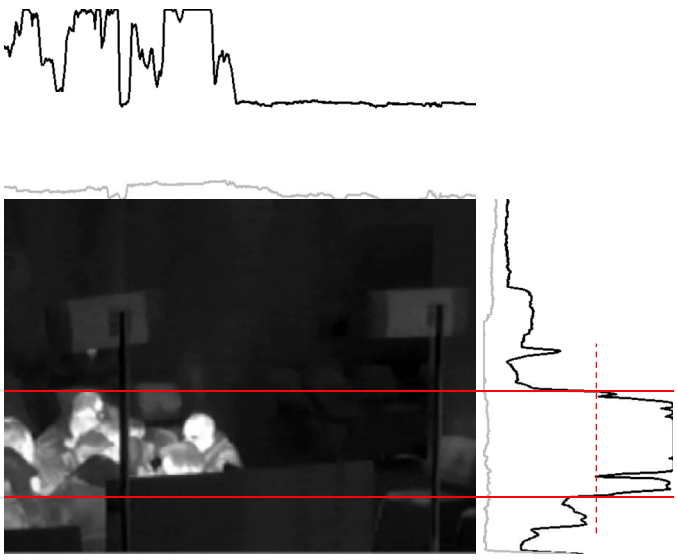

Figure 3. Example of detection of people kneeling while being grouped on the floor. Energy is concentrated in the lower part of frame

At this stage of the experiments, it is not possible to obtain direct performance measures of the proposed algorithm. However, in these preliminary tests it could be observed that the idea presented in this paper works as expected. Detection of the dangerous event was possible in each recorded case. However, the thresholds had to be carefully tuned in order to avoid false detection results. The energy algorithm alone produced a significant number of false positive decisions, but when it was combined with the contour tracking procedure, these errors could be avoided. However, a more detailed, quantitative analysis of the algorithm performance has to be left for a later time.

As far as the computation time is concerned, background subtraction and calculation of local gradients are the most complex parts of the algorithm. However, thanks to low resolution and frame rate of the thermal camera used for the recordings, online analysis was possible. It should be noted that low frame rate limits the accuracy of contour tracking using the proposed procedure, since the frame-to-frame object displacements may become large. Therefore, a device with a larger frame rate is recommended.

IV. CONCLUSIONS

The approach to detection of complex events such as bank robbery involving grouping people on the floor in small area, was presented. The proposed method combines two methods: detection of regions of concentrated energy and detection of merged object contours and their direction of movement. A novel approach presented in this paper is that thermal images are used for analysis, which allow for accurate detection in difficult conditions (poorly lit spaces, obstacles, etc.). The results of preliminary experiments indicate that the solution described in this paper allows for detection of the described event. However, more thorough testing, including the performance evaluation and comparison with standard camera systems is necessary and these issues are planned for the next stage of the research, Further developments on this topic include a more sophisticated decision module and a method of selection of the threshold values. Moreover, we also intend to incorporate audio analysis to the detection procedure, which will lead to a multimodal system for efficient detection of dangerous events in banks – a need of the end users of the current security systems.

REFERENCES

[1] A. Czyżewski, G. Szwoch, P. Dalka, et al, “Multi-stage video analysis framework”, in Video surveillance, W. Lin, Ed,. Rijeka: InTech, 2011, pp. 147-172.

[2] J. W. Davis, and V. Sharma, “Fusion-based background-subtraction using contour saliency”, Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2005. San Diego, 2005.

[3] J. Han, and B. Bhanu, “Fusion of color and infrared video for moving human detection”, Pattern Recognition, vol. 40, pp. 1771-1784, 2007.

[4] B. Putz, M. Bartyś, A. Antoniewicz, et al, “Real-time image fusion monitoring system: problems and solutions”, in Image Processing and Communications Challenges 4, Advances in Intelligent Systems and Computing, R. S. Choraś, Ed., pp. 143-152, Berlin Heidelberg: Springer, 2013.

[5] P. Kumar, A. Mittal, and P. Kumar, “Fusion of thermal infrared and visible spectrum video for robust surveillance”, ICVGIP 2006, Lecture Notes in Computer Sscience, vol. 4338, pp. 528–539, Berlin Heidelberg: Springer, 2006.

[6] G. Szwoch, M. Szczodrak, “Detection of moving objects in images combined from video and thermal cameras”, in Communications in Computer and Information Science, A. Dziech, A. Czyżewski, Eds., vol. 368, pp. 262-272, Heidelberg: Springer, 2013.

[7] N. Otsu, “A threshold selection method from gray-level histogram”, IEEE Trans. Systems, Man, and Cybernetics, vol. 9, pp. 62–66. 1979.

[8] C. Stauffer, W.E.L. Grimson, “Adaptive background mixture models for real-time tracking”, Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 246–252, 1999.

[9] S. Suzuki, K. Abe, “Topological structural analysis of digitized binary images by border following”, Computer Vision, Graphics, and Image Processing, vol. 30, pp. 32-46, 1985.

[10] G. Welch, and G. Bishop, “An introduction to the Kalman filter”, Technical Report TR-95041, Dept. Computer Science, Univ. North Carolina, Chapel Hill ,2006.

[11] J. Davis, and A. Bobick, “The representation and recognition of human movement using temporal templates”, Proc. Comp. Vis. and Pattern Rec., pp. 928-934, 1997.

[12] G.R. Bradski, and J. Davis, “Motion segmentation and pose recognition with motion history gradients”, J. Machine Vision and Applications, vol. 13, pp. 174-184, 2002.

[13] Z. Zhong, M. Yang, S. Wang, W. Ye, Y. Xu, “Energy methods for crowd surveillance”, Proc. International Conference on Information Acquisition, pp.504-510, 2007.