Defined Human Pose Detection for Video Surveillance *

Xu He, Zhengping Wang, Bok-Suk Shin, and Reinhard Klette
The University of Auckland, Computer Science
Auckland, New Zealand
xhe585@aucklanduni.ac.nz

ABSTRACT

This paper presents a system for real-time detection of defined human poses (i.e. raising of hands) in surveillance video. A single (non-calibrated) video camera is used to record data in an indoor environment. There are two main steps in our proposed system, the extraction of human silhouettes in video data and pose classification. Silhouette extraction is refined by paying attention to the removal of shadow artefacts close to occlusion borders. For pose classification, we combined, adjusted, and implemented two existing methods (star skeleton calculation and its evaluation). We demonstrate that the proposed two-step technique is solving the given task for a large percentage of input data when recording an individual person only.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis

Keywords

Video surveillance, Silhouette extraction, Human pose, Pose detection

1. INTRODUCTION

The application of video surveillance systems is today a common feature at public places. A frequent purpose of using a video surveillance system is to ensure security, such as detecting unusual behaviour. A few years ago, commercial and public places used to rely fully on security personal, being increasingly replaced in recent days by surveillance cameras and automated, or semi-automated video analysis. Obviously it would not be efficient to have security personal now sitting in front of a screen for detecting unusual behaviour in recorded data. Semi-automatic video security analysis applications are developed in recent years. Our project aims at understanding sudden changes in human poses in a public area, with a focus on sudden raising of hands in this paper (e.g. in the costumer area of a bank).

Traditional video surveillance systems record all the information for potential later use; this consumes a huge amount of storage space, and data are only used if suspicious events have been indicated by other means. Modern commercial video surveillance systems are capable of recording video only when there is any movement (e.g. human motion) detected, which is a first step towards reducing the bandwidth of data transmission and video storage.

Motion detection and pose analysis [9] is commonly based on generic human body models or detecting changing pixels, not on interpreting the scene shown in a video, such as understanding the behaviour of people. Motion detection results are still inaccurate to some degree, which may lead to false detections. We aim at providing a real-time surveillance system which classifies detected human silhouettes with respect to particular classes of human motions. In this paper we discuss the detection of hand raising.

We provide a brief and selective review of examples of related work; a representative review is out of the scope for this brief conference paper. W4 [6] is a proposed solution for multi-person tracking and activity recognition when using an outdoor video surveillance system. W4 combines various methods of silhouette shape analysis and tracking to determine whether people are carrying objects, or whether people move coordinated in a group.

Fujiyoshi and Lipton proposed a method called *star skeletonization* [5] which is a silhouette-based method for human modelling, subsequently for the analysis of human poses. This method was proposed for classifying human poses into running or walking by using some kind of cycle detection in the Fourier domain.

Star skeletonization is computationally efficient, it is non-iterative and thus very suitable for real-time processing. Following the original proposal of the method, [3] estimate a human skeleton by connecting five *crucial points* (head, left hand, right hand, left foot, and right foot) to the center of gravity of a person. Star skeletonization is used in [10], based on a hidden Markov model, for action recognition.

*LATER

The paper [1] presents another human action recognition method by using three human silhouettes. To create a silhouette descriptor, the method maps the human silhouette into three polar coordinate systems that represent either the whole human body, the upper, or the lower body part, respectively. An action classifier can then be trained based on the derived descriptor.

Paper [2] proposed a method for recognising the raising of a hand in a meeting or class room. The method locates arms in the geometric structure of detected edges, and then compares the angle of a detected hand with the x-axis. Paper [4] also discusses the detection of a raised hand. A *candidate region* (CR) region is identified from silhouettes by locating positions of body parts.

In the the proposed system, we start with extracting a "fairly" accurate human silhouette by further refining an approximate median filtering technique initially suggested in [11]. This is followed by using star skeletonization. We show how star skeletons can be used to identify human poses characterised by raised hands. Figure 1 shows an outline of the basic steps of our system.

This paper is structured as follows: Section 2 outlines the used silhouette-detection process, and Section 3 specifies our applied pose classification method. Experimental results are provided in Section 4, and Section 5 concludes.

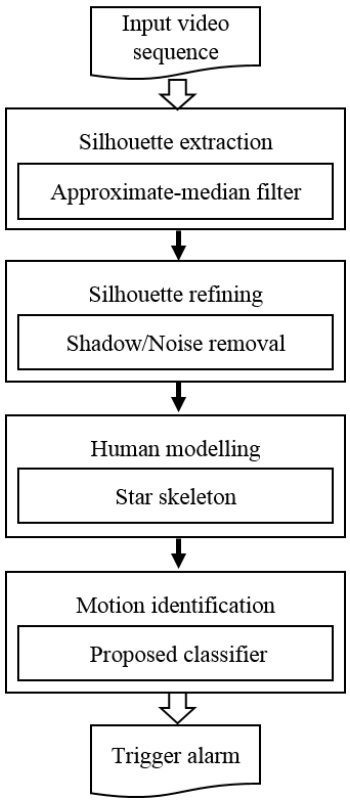

Figure 1: Steps of the proposed system.

2. SILHOUETTE EXTRACTION

Moving object detection has been a subject of research over more than 20 years, but it is still a challenging research topic due to the diversity of possible scenarios. For a monocular 2-dimensional video-surveillance system (i.e. not aiming at distance estimation from recorded data), the camera normally remains static for a long period of time. A common request is also that the computational complexity of all the performed calculations should be in an order to be capable of ensuring real-time processing ("real-time" is defined in an application scenario).

Background subtraction is an appropriate approach for obtaining the silhouette of a moving human in real-time, say up to 25 to 30 fps. In our pose analysis system, we modify the algorithm suggested in [11] (for extracting foreground silhouettes) for the particular purposes in this project. This algorithm uses an approximate median filter [8] together with a novel method for removing shadow artefacts connected to silhouettes.

The steps of our silhouette extraction method are as follows:

Step 1: The background of the video sequence is updated by using an *approximate median filter* [8], which is a statistical background modelling method complementary to simple *median filtering*. The following equation specifies how to update the background:

$$B(x,y,t) = \begin{cases} B(x,y,t-1)+1, \\ \qquad \text{if } I(x,y,t) > B(x,y,t-1) \\ B(x,y,t-1)-1, \\ \qquad \text{if } I(x,y,t) < B(x,y,t-1) \end{cases} \quad (1)$$

where $I(x,y,t)$ is the value of an image pixel at position (x,y) at time t [7], and $B(x,y,t)$ is the value of a background pixel at position (x,y) at time t. Let $I(x,y,0)$ (i.e. in the first frame) be the initial value of $B(x,y,0)$. Then, in each subsequent time frame, we update background pixel values by comparing with previously assigned background pixel values.

Step 2: We estimate the foreground by subtracting the current frame from the background image by using the following equation:

$$F(x,y,t) = \begin{cases} 1 \text{ if } |I(x,y,t) - B(x,y,t-1)| > \sigma_t \\ 0 \text{ otherwise} \end{cases} \quad (2)$$

where $F(x,y,t)$ is the foreground pixel value at position (x,y) at time t, with initial values $F(x,y,0) = 0$ (i.e. all the pixels are considered to be background pixels at the beginning).

A pixel at (x,y) in Frame t is a foreground pixel if the absolute difference between the current value of $I(x,y,t)$ and the background value $B(x,y,t-1)$ is larger than a chosen threshold σ_t. Parameter σ_t is taken as the standard deviation of all input pixels in Frame t, defined by the following

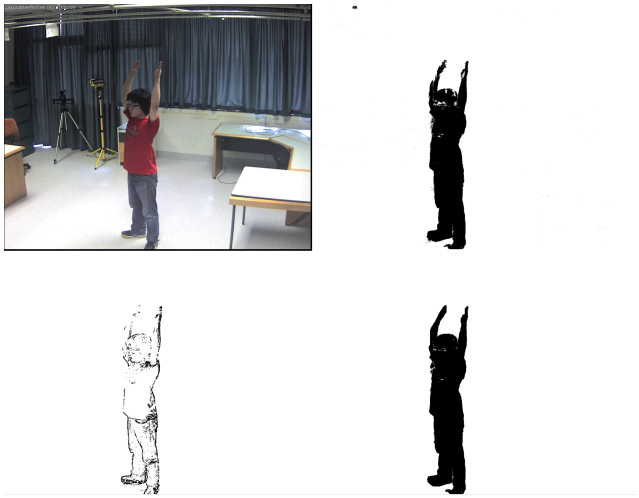

Figure 2: A screen shot of a side-view silhouette of a person. *Top-left:* **Current frame.** *Top-right:* **Foreground detection result.** *Bottom-left:* **Detected true occlusion border.** *Bottom-right:* **Resulting silhouette, the foreground mask.**

equation:

$$\sigma_t = \sqrt{\left(\sum_{x=1}^{N_{cols}} \sum_{y=1}^{N_{rows}} (I(x,y,t) - \mu_t)^2 \right) / |\Omega|} \quad (3)$$

$$\text{with } \mu_t = \left(\sum_{x=1}^{N_{cols}} \sum_{y=1}^{N_{rows}} I(x,y,t) \right) / |\Omega| \quad (4)$$

where μ_t is the mean of all input pixels in Frame t, N_{cols} and N_{rows} are width and height of the frame, respectively, and $|\Omega|$ is the total number of pixels in the frame defined on carrier Ω [7].

Step 3: We use the Sobel operator as a simple and robust edge estimator on the subtracted background image for obtaining raw occlusion boundaries of a person.

Step 4: We subtract the background boundaries from the raw occlusion boundaries (of a person) in order to extract the *true occlusion border* of a person. See Fig. 2, lower left.

5. Finally, we fill the true occlusion border to obtain the foreground mask, also called the *silhouette*. Figure 2, lower right, shows such a mask. An ideal silhouette is typically a simply-connected region.

The extracted silhouette is not "perfect", it still contains noisy artefacts (e.g. holes). Small noisy defects can be removed by applying morphological operations such as closing. "Large blob noise" can be removed by counting total numbers of pixels; we set a threshold to remove large blobs. In our system, we use 800 as a fixed threshold; for the used image resolution and expected distance of people to the camera, this is approximately one third of the total number of pixels of a single human silhouette.

Also, some body parts can appear as separate regions not connected with the main body part due to erroneous background detection or shadows. We apply a set of morphological operators to remove the described noise and holes, and for connecting isolated body regions with the main region of a silhouette.

For example, the star skeletonization paper [5] suggested to apply morphological dilation twice followed by an erosion (i.e. a kind of a closing operator), and we estimated numbers of applications based on sizes of regions. This can effectively generate a silhouette which also represents fairly accurate smaller or thinner body parts.

3. POSE UNDERSTANDING METHODS

We aim at a classification of obtained silhouettes (of a single standing person) into two classes, either "normal body pose" or "hands raised".

3.1 Star Skeletonization

A skeleton is a common way for matching a detected region in an image with a model of a human. In our system, we use star skeletonization for modelling a human skeleton by following [5].

The proposed method is defined by a robust approach for detecting extrema points on the border of a detected silhouette. The algorithm for this subprocess consists of the following steps:

Step 1: Trace the border of the extracted silhouette.

Step 2: Find the centre of gravity of the target border. Suppose that there are N border pixels, and the centroid of the border is denoted by (x_c, y_c), being defined as follows:

$$x_c = \frac{1}{N} \sum_{i=1}^{N} x_i, \quad y_c = \frac{1}{N} \sum_{i=1}^{N} y_i \quad (5)$$

for the N positions (x_i, y_i) of border pixels.

Step 3: In a second scan of the border, we define a distance function $D(i)$ which is the distance between the centre of the gravity (x_c, y_c) and each border pixel point:

$$D(i) = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \quad (6)$$

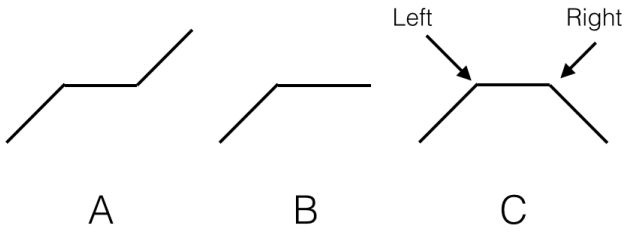

Figure 3: Three point sequences. A and B are not a maximal situation. C is a maximal situation.

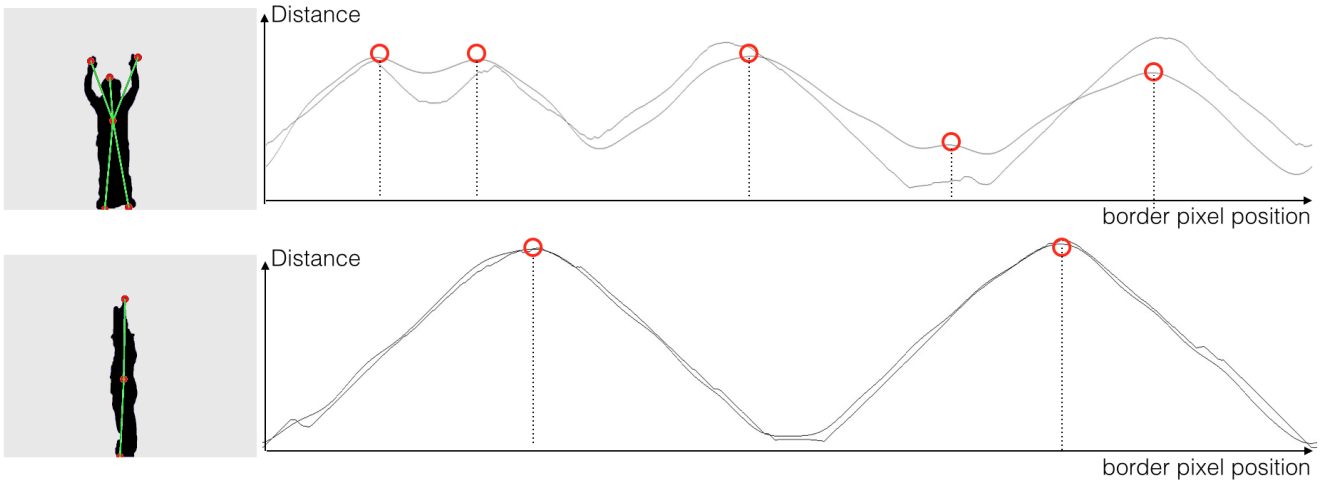

Figure 4: *Right:* **Original distance signal D and smoothed distance signal $\hat{D}$ defined by a low pass in the Fourier domain.** *Left:* **Calculated skeletons.**

Border tracing can be in clockwise or anti-clockwise order. In the experiments we used the Euclidean distance as shown in Equ. (6). However, we could also use the squared Euclidean distance or, for example, the L_1 (for saving computation time) without any significant impact on final results.

Step 4: The values $D(i)$ obtained in Step 3 are noisy (i.e. irregular). We smooth this distance signal $D(i)$ by applying a low-pass filter in the frequency domain. The low pass filter has a cutoff-threshold a for filtering out the high-frequency components. Let $\mathbf{D}$ be the Fourier transform of D. We set

$$\mathbf{D}(u) = 0 \text{ if } |u| \geq a \cdot N \tag{7}$$

where N is the total number of border pixels and u the frequency coordinate. The larger the threshold the less local maxima (e.g. five local maxima for $a = 0.0025$, and three local maxima for $a = 0.01$) can be detected in the distance signal $\hat{D}$ (obtained after the inverse Fourier transform). Paper [5] used $a = 0.015$ as the threshold for their data and purposes, but we have higher image resolution and the goal of detecting raised hands, and we used $a = 0.0004$ in our experiments.

Step 5: We take all the detected local maxima in the filtered signal $\hat{D}$ as extrema points. Figure 3 shows three different situations of $\hat{D}$ value sequences. Values may be constant within some intervals. A local maximum is calculated for situation C by taking the mean

$$i_{\max} = \frac{i_L + i_R}{2} \tag{8}$$

of the shown left and right endpoint of the interval (having indices i_L and i_R in $\hat{D}$) of constant values.

A human skeleton is now constructed by connecting the centre of gravity with the detected local maxima for the given silhouette. Figure 4 illustrates in the top row the resulting skeleton for a front-view of a human silhouette. Parameter a was chosen in a way such that the smoothed signal $\hat{D}$ of such a front-view human silhouette typically contains five maximal points. Each of those maximal points matches then typically a distinctive feature of the human body (i.e. feet, head, and possibly hands or shoulders). The bottom row in Fig. 4 illustrates a side view for raised hands.

3.2 Detection of Raised Hands

A person which is raising the hands can be recorded by a camera in various *poses* (i.e. positions and directions). Figure 6 illustrates three different *main poses* of a person raising hands as usually appearing in a surveillance camera. The figure illustrates that the front-view skeleton normally contains five maximal points, and raising hands are defining two of those.

The top row also shows a variation in front-view skeletons. For the first frame, only head and feet are detected. The shoulders do not lead to maxima in the filtered distance signal. The second frame misses the detection of the left foot, and there are two maxima detected for the left hand. The third frame is missing the right hand. All the other frames lead to five maximal points.

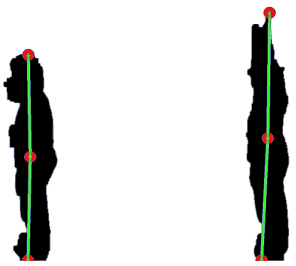

Figure 5: The two extreme cases denoted as *I-shape*. A temporal change in height indicates the raised hands. *Left:* **A side view of a standing person.** *Right:* **A side view of a standing person having the hands up.**

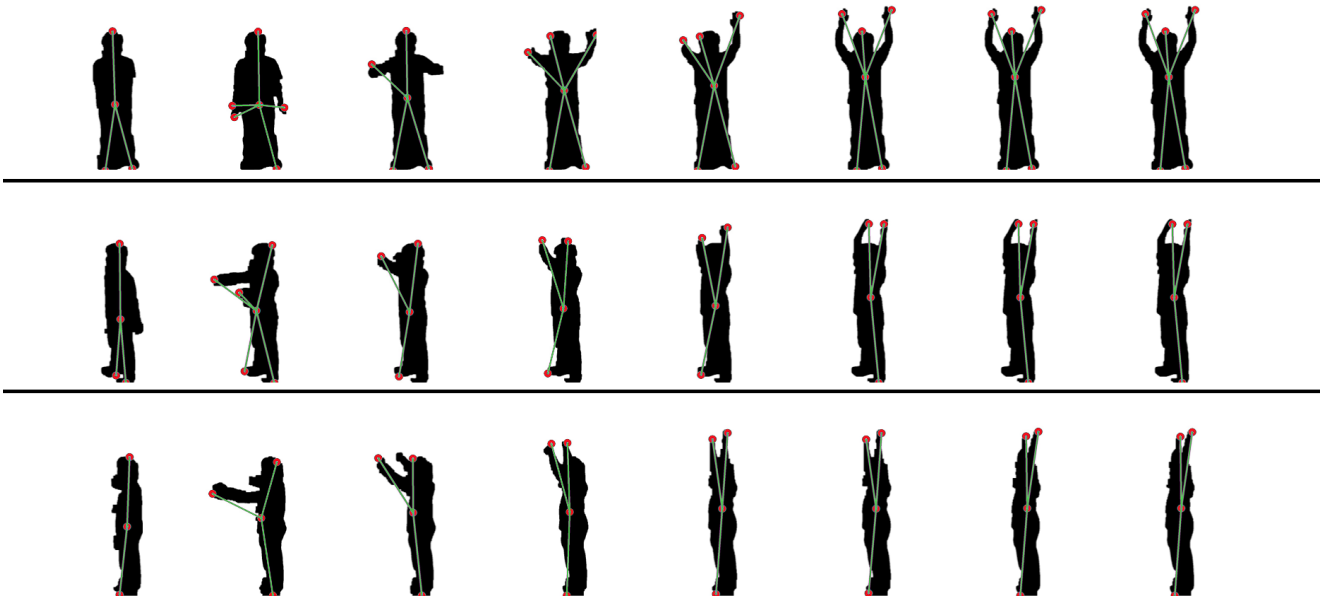

Figure 6: Different states of a person which raises the hands. *Top Row:* **Front view.** *Second Row:* **Half-side view.** *Bottom Row:* **Side view.**

The second row shows states of a half-side pose. The first frame does not detect hands since they are barely seen for this pose. The second frame contains all the five maximal points. The next six frames all miss to detect one of the feet.

When raising the hands, the skeleton forms a *Y-shape* for front and half-side pose.

The third row shows a side-view skeleton. The skeleton is at first (i.e. first frame) just connecting two points (head and feet), and turns then into a Y-shape when the person is raising the hands.

Following above observations, the formed skeleton does not always contain exactly five maximal points due to differences in poses. That means, head, hands, or feet cannot be detected and tracked correctly as maximal points over a sequence of frames.

If a person is raising the hands, in all cases the common feature is that the positions of the two hands are on top of the head position at some stage. In consequence, we focus on the analysis of three maximal points defining the positions of three upward maxima.

We propose to detect raised hands by identifying one of five shapes, called *I*-, *Y*-, *X*-, ψ-, or *star-shape*. A single upper maxima in the *I*-shape, or two upper maxima in the *Y*-, *X*-, ψ-, or *star*-shape, are considered to be hands, and the bottom two maxima in the *X*- and *star*-shape are considered to be the feet.

For the *Y*-, *X*-, ψ-, or *star*-shaped skeletons, we specify two thresholds for comparing hand(s) and feet maxima with an adaptive estimate H for the height of the shown person (based on calculating heights of silhouettes over some frames by a sliding mean). Assume that the y-axis is pointing upward in the image. Then we use

$$y_{\text{hands}} > 0.8 \cdot H \;\&\; y_{\text{feet}} < 0.2 \cdot H \tag{9}$$

By using this method, we are able to identify *Y*-, *X*-, or *star*-shapes fairly accurate in a given frame.

The *I*-shape (see Fig. 5) requires a particular consideration. Here we compare the current height with the previously estimated height H.

4. RESULTS

In this paper we report results obtained for five different test persons. The first test person (called Person A) was repeatedly shown in previous figures. For test persons B, C, D, and E, see Fig. 8 for some illustrations. The heights of those five persons A-E are 165 cm, 175 cm, 166 cm, 175 cm, and 158 cm, respectively

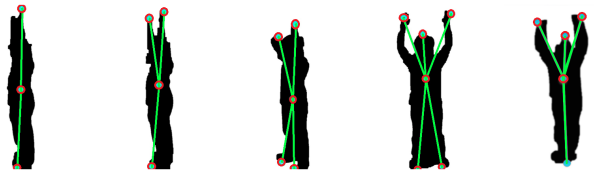

Figure 7: *Left to right:* **The *I*-, *Y*-, *X*-, *star*-, ψ-shape.**

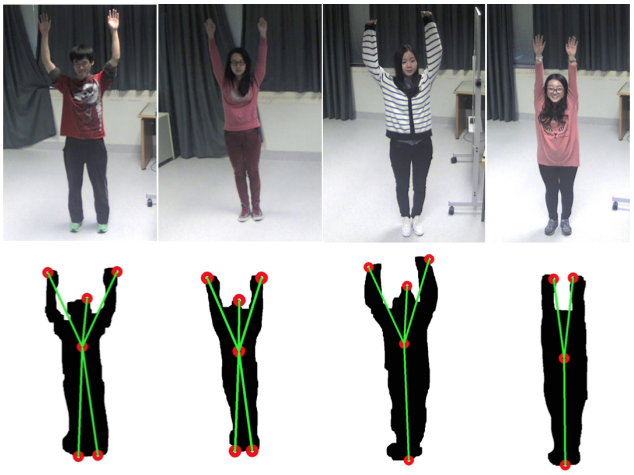

Figure 8: Samples of extracted silhouettes for test persons B to E, left to right.

We use a monocular (non-calibrated) IP camera (ACM-1511) for recording video samples. In order to simulate a standard indoor surveillance situation, the camera is mounted close to the ceiling. Recording is at 8 fps, with 1280 x 1024 image resolution.

We tested our algorithm on several video sequence for five different people. For each person, recorded sequences contained all the three poses illustrated in Fig. 6. Table 1 summarises the results of pose classification. FP (i.e. false-positive) specifies the number of detections of raised hands when there are actually no raised hands, whereas FN specifies the number of cases where we miss to detect the raised hands in a frame. We manually classify poses into "hands raised" if both hands are at the height level of the head, or above the head. The "Ratio" is finally the total number of frames minus (FP+FN), divided by the total number of frames, i.e. the percentage of correct decisions.

For example, the relatively large value FP = 22 for Person C is due to missing head silhouettes; Person C defined a "difficult case" for silhouette detection due to signal similarities between person and background.

The total sequential processing time of the system was on average 682 ms per frame, using an 1280 × 1024 image resolution on a four core (4.2 GHz) machine with 16GB RAM.

Table 1: Classification results for raised hands for five test persons

Sequence	Total frames	FP	FN	Ratio
#A	172	0	4	97.67%
#B	305	2	10	96.07%
#C	287	22	4	90.94%
#D	293	4	8	95.90%
#E	243	5	4	96.30%
Total	1300	33	30	95.15%

5. CONCLUSIONS

The proposed system detects raised hands in surveillance videos by using a refined silhouette extraction method followed by star-skeleton calculation and star-skeleton evaluation. The method was tested for observing a single person. A multiple-person case where people partially occlude eachother requires further considerations. However, the temporal change of estimated height of calculated skeletons can again be used as specifying parameter as long as the group of people is defining "one layer in depth" only. Having occlusions between people at different depths will probably require to analyse the actual activities directly in the image data (i.e. without generating silhouettes).

Acknowledgments. Authors thank TN Chan of Compucon (Auckland) for providing a camera and further support for the reported project.

6. REFERENCES

[1] C. Hsieh, P. S. Huang, and M. Tang. Human action recognition using silhouette histogram. In Proc. *Australasian Computer Science Conference*, pages 11–16, 2011.

[2] N. B. Bo, P. V. Hese, D. V. Cauwelaert, P. Veelaert, and W. Philips. Detection of a hand-raising gesture by locating the arm. In Proc. *IEEE Int. Conf. Robotics Biometrics*, pages 976–980, 2011.

[3] P. Correa, J. Czyz, T. Umeda, F. Marques, X. Marichal, and B. Macq. Silhouette-based probabilistic 2d human motion estimation for real-time applications. In Proc. *IEEE Int. Conf. Image Processing*, pages 836–839, 2005.

[4] X. Duan and H. Liu. Detection of hand-raising gestures based on body silhouette analysis. In Proc. *IEEE Int. Conf. Robotics Biometrics*, pages 1756–1761, 2009.

[5] H. Fujiyoshi and A. J. Lipton. Real-time human motion analysis by image skeletonization. In Proc. *IEEE Workshop Applications Computer Vision*, pages 15–21, 1998.

[6] I. Haritaoglu, D. Harwood, and L. S. David. W4: real-time surveillance of people and their activities. *IEEE Trans. Pattern Analysis Machine Intelligence*, **22**:809–830, 2000.

[7] R. Klette. *Concise Computer Vision*. Springer, London, 2014.

[8] N. J. B. McFarlane and C. P. Schofield. Segmentation and tracking of piglets in images. *Machine Vision Application*, **8**:187–193, 1995.

[9] B. Rosenhahn, U. Kersting, K. Powell, R. Klette, G. Klette, and H.P. Seidel. A system for articulated tracking incorporating a clothing model. *Machine Vision and Applications*, **18**:25–40, 2006.

[10] D. Singh, A. K. Yadav, and V. Kumar. Human activity tracking using star skeleton and activity recognition using HMMs and neural network. *IJSRP*, **4**, May 2014.

[11] Z. P. Wang, B.-S. Shin, and R. Klette. Accurate silhouette extraction of a person in video data by shadow evaluation. *J. Computer Theory Engineering*, **6**:476–483, 2014.