Human Skeleton Detection from Semi-constrained Environment Video

Palwasha Afsar, Paulo Cortez and Henrique Santos
ALGORITMI Centre, Department of Information Systems, University of Minho, 4804-533 Guimaraes, Portugal
palo_afsar77@yahoo.com, {pcortez, hsantos}@dsi.uminho.pt

Keywords: Human Action, Action Recognition, Video Data, Skeleton Detection.

Abstract: The correct classification of human skeleton from video is a key issue for the recognition of human actions and behavior. In this paper, we present a computational system for a passive detection of human star skeleton from raw video. The overall system is based on two main modules: segmentation and star skeleton detection. For each module, several computer vision methods were adjusted and tested under a comparative analysis that used a challenging video dataset (e.g., different daylight and weather conditions). The obtained results show that our system is capable of detecting human skeletons in most situations.

1 INTRODUCTION

Due to the widespread use of digital cameras, human activity recognition from video is becoming a trendy topic in the area of computer vision (Poppe, 2007; Turaga et al., 2008; Enzweiler and Gavrila, 2009; Geronimo et al., 2010; Afsar et al., 2015a). The passive and automatic recognition of human actions (e.g., walking, sitting, exhibiting interest, aggression behavior), from monocular images and videos is a valuable tool in several computer applications, such as human computer interaction, video content retrieval, virtual reality, analysis of sports events, video indexing and surveillance. For instance, in (Cortez et al., 2016) a computer vision system was used to monitor and forecast the entrance of customers into commercial store, allowing to enhance the store management. Unlike other well defined computer vision tasks (e.g., motion blur, edge detection), human behavior recognition (e.g., exhibiting interest or aggression behavior) does not have a clear algorithmic definition and thus this task can be challenging.

Human star skeleton recognition from video is a crucial element of several human action recognition systems (Orrite-Urunuela et al., 2004; Fujiyoshi et al., 2004; Chen et al., 2006; Yang and Tian, 2014; Vemulapalli et al., 2016). However, most of these systems were proposed to work in controlled nonrestrictive environments (e.g., with humans close to the camera and wearing distinctive clothing colors when compared with the background). In this position paper, we present a computational system for a passive detection of human star skeleton from raw video and that was built to work in a semi-constrained but more realistic environment. The final and future goal of our system is to use the detected human skeletons to create human movement features (e.g., speed or acceleration) that will fed a machine learning classifier, which will be trained to detect interesting human behaviors (e.g., walking, sitting, making a cellular call).

The proposed computer vision system is composed of two main modules (Figure 1): segmentation (including background subtraction and shadow and highlight removal) and star skeleton detection. For each module, we experimented, adjusted and compared several computer vision methods (e.g., adaptive background mixture model based on Gaussian mixture model versus simpler background subtraction, thinning algorithm versus simpler zero-crossing). We report here the results so far achieved using video data collected in (Afsar et al., 2015b) and that corresponds to a semi-constrained but realistic university campus environment. The recorded digital videos are related with two cameras that were used to capture two particular examples of interior and exterior human walking areas from our campus. Since we adopted a real environment, the recorded data includes several restrictions that pose challenges: the cameras were set in front of a glass window (thus some reflection is captured) and far away from the human walking environment (some humans are captured with a low pixel definition); there are different weather conditions in the exterior campus area (e.g., rain and wind) and varying illumination in both interior and exterior areas due to different daytime recordings); there are clutter scenes in both areas due to the presence of

Afsar P., Cortez P. and Santos H.
Human Skeleton Detection from Semi-constrained Environment Video.
DOI: 10.5220/0006245803840389
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 384-389
ISBN: 978-989-758-226-4

trees and bushes; often, the human clothing includes colors that are very similar when compared with the background; and other uncontrolled conditions. After tuning the two system modules, interesting human star skeleton results were achieved (shown in Section 3).

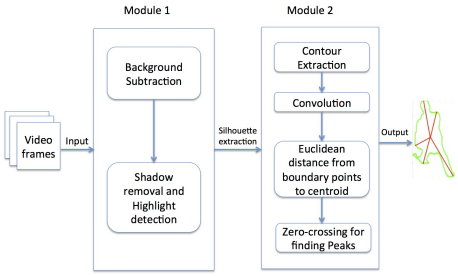

Figure 1: The overall framework of the system.

2 RELATED WORK

A key issue within the context of human action recognition is the automatic detection of human star skeleton, typically consisting of a 5-point star (with head, hands and legs – the human body extremes) and a body center of mass. In effect, several works have proposed and adopted methods based on human star skeletons for action recognition. Two dimensional images were used in:

- (Orrite-Urunuela et al., 2004) – where manual effort was used to define the skeleton points; and
- (Chen et al., 2006) and (Fujiyoshi et al., 2004) – where preprocessing (e.g., segmentation, dilatation) and zero-crossing methods were used to build the 5-point human stars.

More recently, three dimensional video has also been modeled to use similar human skeleton approaches (Yang and Tian, 2014; Vemulapalli et al., 2016). However, most of these human skeleton usage or detection works were proposed to work with clean, non-restrictive environments where: the camera is positioned close to the human subjects, the background is clear and colored with a distinctive color when compared with humans, most environments are interior and thus immune to weather conditions, etc. While such 2D controlled environments facilitate the development of the computer vision system, they are less realistic and thus are of less practical usage. Moreover, the usage of depth information (3D) does not have a good precision and/or requires very expensive equipment when human subjects are far away from the camera. In this position paper, we use a realistic 2D environment from our university campus and that contains several challenges (detailed in Section 3.1). As such, we tune and test several computer vision methods under this semi-constrained video environment, resulting in a computational system that is capable of detecting human skeletons.

3 PROPOSED APPROACH

The task of finding extreme points of a human silhouette in realistic environments is challenging due to self-occlusion, articulated human body, missing depth information, invariant appearance due to camera viewpoints, illuminations and loose clothing. In this paper, and similarly to (Chen et al., 2006), we assume that a human skeleton is developed by detecting extreme points such as head, hands and feet. These extreme points can be used for the recognition of many human actions (e.g., walking). The overall computational system is based on two main modules (Figure 1): segmentation (including background subtraction and shadow and highlight removal); and star skeleton detection. All experiments were conducted using the Matlab computational environment (https://www.mathworks.com). In each module, we experimented, adapted and compared several computer vision methods, as detailed in the next subsections.

3.1 Video data

For this research, we adopt the video dataset that was presented in (Afsar et al., 2015b). For capturing the data, two cameras HIK Vision and IR Network were installed. Both cameras were “hidden” behind a window glass (Figure 2) and set to capture real interior and exterior human walking campus areas (Figure 3). The whole dataset includes hundreds of small videos (with few seconds to few minutes each) that correspond to 32GB and that included a non controlled capturing of human (students, researchers and other staff) actions (e.g., walking, running, drinking coffee).

Figure 2: Cameras placement for recording the dataset.

The dataset is challenging, due to several uncontrolled factors, such as: there is some reflection in the video (since cameras were set behind a glass); wide scenes are captured and some humans are far away from the camera, thus captured with a low resolution; there are different weather conditions (e.g.,

Figure 3: Examples of the exterior (top frames) and interior (bottom frames) areas.

sun, rain and wind) and varying illumination due to different daytime recordings); there are clutter scenes due to the present of trees and bushes (in both interior and exterior environments); and often, human wear clothes that are very similar when compared with the background. Several of these elements can be seen in Figure 3.

3.2 Segmentation

3.2.1 Background Subtraction

The video is processed frame by frame and blob analysis is performed to look for any active blobs. In order to select only humans and to remove noise, we defined a minimum pixel area for blob selection, which was set to 2000 pixels (e.g., 45x45, 60x34) after some preliminary experiments.

For background subtraction, we first tested the gaussian mixture model proposed in (Kaewtrakulpong and Bowden, 2002). for separating the foreground pixels from the background. This detector works on data collected by a stationary camera and compares a color or gray scale video frame to a background model to figure out whether its part of the background or foreground. It then computes a foreground mask based on Gaussian Mixture Models (GMM). This algorithm has the ability to adapt itself slowly to the changing environment but for our dataset, the results obtained were not satisfactory. As shown in Figure 4, the algorithm was not able to detect legs, although there was a large space between the legs. Note that in Figures 4 and 5 we hide the human face with a red rectangle to preserve anonymity. Moreover, the GMM algorithm is computationally expensive.

We then tested a much simpler and lightweight approach, based on the absolute subtraction of input and

Figure 4: Segmentation results (first column, a), denotes the original input frame; second column, b), the Gaussian mixture model result; and third column, c), the background subtraction result).

background image. Background was updated by using the same position of the bounding box obtained through blob detection as both frames are of the same size. As an initial step, we subtracted the input Image I from the background Image B, both in RGB color space.

$$M = |B - I| \tag{1}$$

The obtained mask M was converted to grayscale and finally to binary using Otsu's thresholding (Otsu, 1975). This method has the drawback that when the clothing of the individual's were light colored, the results were not good even after apply morphological operations. There were some information loss in the human silhouette, which generated several smaller blobs rather than one blob for an individual (Figure 5).

Figure 5: a) Original input image, b) result of absolute subtraction, c) binary image and d) result with shadow removal and highlight detection.

Due to light color clothing, illuminations, and distance from the camera, there is some information loss.

In the resultant image (in RGB space), we tried to:

$$M'(x,y) = \begin{cases} 1 & \text{if } M(x,y) \geq \tau. \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

where M' is the final mask, $M(x,y)$ is the pixel value of the mask, and τ is a threshold. We tested several τ values (e.g., $\tau \in \{10,20,30\}$). The best results were achieved with $\tau = 30$ (as shown in Figure 6) and thus this was the selected threshold value.

Figure 6: a) Original image, b) mask for τ=20, c) mask for τ=30 and d) grayscale mask for τ=30.

Then, we experimented to compute M' using the HSV space, instead of RGB, since this is the same color space used by the shadow and highlight removal method (Section 3.2.2). The results obtained are shown in Figure 7. The human silhouette still has some information loss.

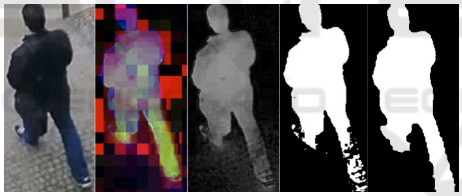

Figure 7: a) Original image, b) HSV image, c) *Value* component, d) binary image and e) result after shadow and highlight removal.

3.2.2 Shadow Removal and Highlight Detection

We tested the method proposed in (Duque et al., 2005), where both images, background B and input image I are transformed into the HSV color space. A shadow mask SM and highlight mask LM are generated from these images, defining the areas where shadows or highlights are present. In these masks each pixel will have value "1" if it is marked as shadow or highlight, and value "0" otherwise. The following equations define the process of computing the shadow and highlight masks:

$$SM(x,y) = \begin{cases} 1 \text{ if } & \alpha \leq \frac{I^V(x,y)}{B^V(x,y)} \leq \beta \\ & \wedge \mid I^S(x,y) - B^S(x,y) \mid \leq \tau_S \\ & \wedge I^S(x,y) - B^S(x,y) \mid \leq \tau_H \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$SM(x,y) = \begin{cases} 1 \text{ if } & \frac{1}{\beta} \leq \frac{I^V(x,y)}{B^V(x,y)} \leq \frac{1}{\alpha} \\ & \wedge \mid I^S(x,y) - B^S(x,y) \mid \leq \tau_S \\ & \wedge I^S(x,y) - B^S(x,y) \mid \leq \tau_H \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

The $I^H(x,y)$, $I^S(x,y)$ and $I^V(x,y)$ represent respectively the hue, saturation and value components at coordinate *(x,y)* of the input image *I*. The same notation is applied to the background image *B*.

The α is the main parameter and its value depends on the light source, radiance and reflectance properties of the objects in the scene. High reflective and high intensive light sources or irradiant objects can have low α values. For our dataset, α varies from 0.60 to 0.90. If we decrease the value below 0.60, there will be information loss. Similarly, if the value goes above 0.90, the final image will contain noise. We tested the algorithm with different alpha values, within the range {0.65,0.66,67,...,0.90} and the best results were achieved for α=0.70. After setting α, we experimented distinct values for β, which prevents misclassification and varies in the data from 0.75 to 0.79. Since this is a less relevant parameter, we experimented distinct values but achieved the same results, thus this parameter was fixed to $\beta = 0.75$. The parameters τ_S and τ_H are the maximal variation allowed for the saturation and hue components. We define τ_S as 15% of the digitizers saturation range. The variation of hue should not pass the 60 degrees. This value is obtained through the division of the hue range (360°) by the six colors (red, yellow, green, cyan, blue and magenta). The results obtained were satisfactory, as shown in Figures 5 d), 7 e) and 8.

Figure 8: Example of human silhouettes using background subtraction combined with shadow and highlight removal.

3.3 Star Skeleton

The main idea of a star skeleton is to connect the extreme points (head, legs, hands) with the centroid (the body of mass). In this paper, the human contour is used as the main feature for the construction of a 5 point star skeleton. Depending on the posture of the human, the points in the skeleton can be either 5 or less than 5. These points can represent a human posture effectively and make faster the execution of the final human action detection. As an initial step, the

contour is extracted from the human silhouettes. For the removal of noise from the human contour, convolution was applied, which removes noise and smooths a function. A threshold value of 12 was used in order to achieve the desired level of smoothing. To calculate the distance of individual boundary points from centroid, the euclidean function was used. The whole human contour was processed in a clock-wise order.

In a function, extreme points (or local maxima) are the high peaks points or the points where a zero-crossing is detected when analyzing smoothed distance differences. For the construction of the 5-star skeleton, when the number of points are greater than 5, a threshold value of 40 was used to find the boundary distance between those points, i.e., detect which boundary points are closer to each other. Then, the median of such points is used as a representative of the extreme human part. An example of this is shown in the third row of Figure 10, where two points are replaced by the median for the right hand. The last step is to connect all of the points to the centroid. Figure 9 depicts the overall procedure for the construction of the star skeleton. The points A, B, C, D, E and F represent the extreme points of the smooth distance function. Since both points A and F are closer in terms of the boundary space, we take the median of these two points to represent the left leg when defining the star skeleton. More examples of the obtained star skeletons are shown in Figure 10.

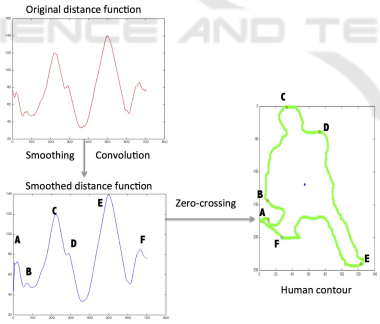

Figure 9: Process flow for the construction of a star skeleton.

The full steps for the construction of a star skeleton algorithm are:

1 Calculate the centroid of the contour of the input image (x_c, y_c).

$$x_c = \frac{1}{N_b}\sum_{i=1}^{N_b} x_i \qquad y_c = \frac{1}{N_b}\sum_{i=1}^{N_b} y_i \tag{5}$$

where B_n are the number of boundary points and (x_c, y_c) denotes the centroid of the input contour.

2 Determine the distance d_i from each boundary point (x_i, y_i) to centroid (x_c, y_c).

$$d_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \tag{6}$$

3 To remove the noise or unwanted peaks, the distance function is smoothed using convolution.

4 Find the local maximum by detecting zero crossing of the distance function differences $(d_{i+1} - d_i)$.

For a comparative analysis, we compared the simpler zero crossing method with a thinning method (based on the Matlab "bwmorph" function). Figure 10 shows some results obtained by our star skeleton method and "bwmorph". It is clear that our approach performs better for calculating the posture of human as compared with "bwmorph" function. Moreover, our approach requires less computation when compared with the thinning method.

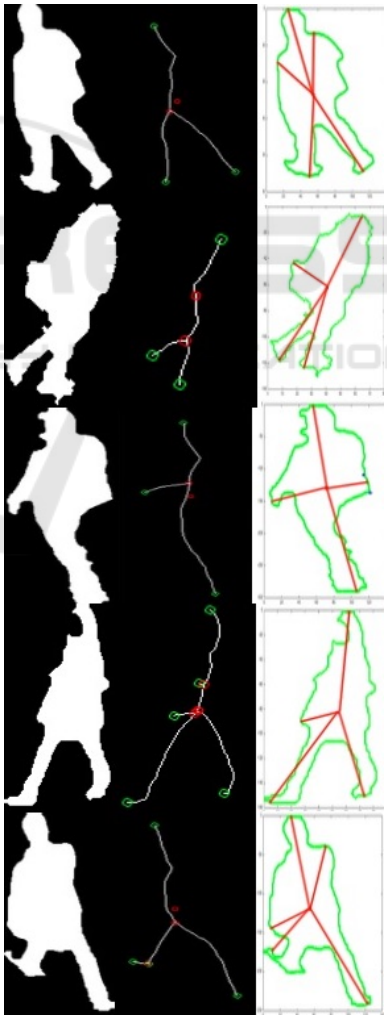

Figure 10: Comparative analysis of star skeleton with *"bwmorph"* function a) Binary Image b) Modified *bwmorph* result c) our skeleton algorithm

4 CONCLUSIONS

In this paper, we propose a computational system for a passive detection of a 5-star human skeleton based on raw video. The system includes two main modules, segmentation and star skeleton detection, and it was adjusted and evaluated using a semi-restricted realistic environment. The realistic videos are related with interior and exterior human walking areas with varying illumination, clutter, and other uncontrolled conditions (e.g., weather). Several computer vision methods were explored for the segmentation and star skeleton modules. The best results were achieved using simpler approaches: background subtraction and shadow and highlight removal using HSV color space; smoothed Euclidean distance to centroid and zero-crossing of distance differences to detect the human extremes. In future work, we intend to use motion and memory to estimate the position of human parts (e.g., hands) that might be temporarily hidden. Also, we plan to create motion skeleton features (e.g., velocity, acceleration) in order to train a machine learning classifier such that it can learn to detect human actions (e.g., walking, making a cellular call).

ACKNOWLEDGEMENTS

This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and *FCT - Fundação para a Ciência e Tecnologia* within the Project Scope: UID/CEC/00319/2013 and research grant FCT SFRH/BD/84939/2012.

REFERENCES

Afsar, P., Cortez, P., and Santos, H. (2015a). Automatic human action recognition from video using hidden markov model. In *Computational Science and Engineering (CSE), 2015 IEEE 18th International Conference on*, pages 105–109. IEEE.

Afsar, P., Cortez, P., and Santos, H. (2015b). Automatic visual detection of human behavior: A review from 2000 to 2014. *Expert Systems with Applications*, 42(20):6935–6956.

Chen, H.-S., Chen, H.-T., Chen, Y.-W., and Lee, S.-Y. (2006). Human action recognition using star skeleton. In *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*, pages 171–178. ACM.

Cortez, P., Matos, L. M., Pereira, P. J., Santos, N., and Duque, D. (2016). Forecasting store foot traffic using facial recognition, time series and support vector machines. In *Soft Computing Models in Industrial and Environmental Applications (SOCO), Advances in Intelligent and Soft Computing (AISC) Vol. 527*, pages 267–276, San Sebastian, Spain. Springer.

Duque, D., Santos, H., and Cortez, P. (2005). Moving object detection unaffected by cast shadows, highlights and ghosts. In *IEEE International Conference on Image Processing 2005*, volume 3, pages III–413. IEEE.

Enzweiler, M. and Gavrila, D. M. (2009). Monocular pedestrian detection: Survey and experiments. *IEEE transactions on pattern analysis and machine intelligence*, 31(12):2179–2195.

Fujiyoshi, H., Lipton, A. J., and Kanade, T. (2004). Real-time human motion analysis by image skeletonization. *IEICE TRANSACTIONS on Information and Systems*, 87(1):113–120.

Geronimo, D., Lopez, A. M., Sappa, A. D., and Graf, T. (2010). Survey of pedestrian detection for advanced driver assistance systems. *IEEE transactions on pattern analysis and machine intelligence*, 32(7):1239–1258.

Kaewtrakulpong, P. and Bowden, R. (2002). An improved adaptive background mixture model for real-time tracking with shadow detection. In *Video-based surveillance systems*, pages 135–144. Springer.

Orrite-Urunuela, C., del Rincon, J. M., Herrero-Jaraba, J. E., and Rogez, G. (2004). 2d silhouette and 3d skeletal models for human detection and tracking. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 244–247. IEEE.

Otsu, N. (1975). A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27.

Poppe, R. (2007). Vision-based human motion analysis: An overview. *Computer vision and image understanding*, 108(1):4–18.

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

Vemulapalli, R., Arrate, F., and Chellappa, R. (2016). R3dg features: Relative 3d geometry-based skeletal representations for human action recognition. *Computer Vision and Image Understanding*.

Yang, X. and Tian, Y. (2014). Effective 3d action recognition using eigenjoints. *Journal of Visual Communication and Image Representation*, 25(1):2–11.