Person re-identification from CCTV silhouettes using Generic Fourier Descriptors

Rawabi Alsedais
Engineering and Digital Arts
University of Kent
Canterbury, Kent, UK
raaa3@kent.ac.uk

Richard Guest
Engineering and Digital Arts
University of Kent
Canterbury, Kent, UK
R.M.Guest@kent.ac.uk

Abstract—Person re-identification in public areas (such as airports, train stations and shopping malls) has recently received increased attention from computer vision researchers due, in part, to the demand for enhanced levels of security. Re-identifying subjects within non-overlapped camera networks can be considered as a challenging task. Illumination changes in different scenes, variations in camera resolutions, field of view and human natural motion are the key obstacles to accurate implementation. This study assesses the use of Generic Fourier Shape Descriptor (GFD) on person silhouettes for re-identification and further established which sections of a subject's silhouette is able to deliver optimum performance. Human silhouettes of 90 subjects from the CASIA dataset walking 0° and 90° to a fixed CCTV camera were used for the purpose of re-identification. Each subject's video sequence comprised between 10 and 50 frames. For both views, silhouettes were segmented into eight algorithmically defined areas: head and neck, shoulders, upper 50%, lower 50%, upper 15%, middle 35%, lower 40% and whole body. A GFD was used independently on each segment at each angle. After extracting the GFD feature for each frame, a linear discriminant analysis (LDA) classifier was used to investigate re-identification accuracy rate, where 50% of each subject's frames were training and the other 50% were testing. The results show that 97% identification accuracy rate at the 10th rank is achieved by using GFD on the upper 50% segment of the human silhouette front (0°) side. From 90° images, using GFD on the upper 15% silhouette segment was almost 98% accuracy rate at the 10th rank. This study illustrates which segments

Keywords— Computer Vision, Generic Fourier Descriptor, Biometric re-identification, Silhouette.

I. INTRODUCTION

Person re-identification (re-id) in public areas (such as airports, train stations and shopping malls) is increasingly receiving attention from computer vision researchers towards enhancing the security levels. Installing camera networks (or CCTV) within non-overlapped field of view provides enhanced coverage for re-identifying subjects. These cameras provide enormous amount of records that are manually monitored by enforcement officers. As human monitoring is error prone, automated analysis of these videos might improve the surveillance quality.

Person re-id task is mainly manipulated by extracting features from subjects exist in the scene. There are various methods that extract features, however, the most common are by neural network methods. For instance, [1] used convolutional neural network (CNN) technique to extract feature vectors from subjects frames and videos in public areas for the task of person re-id. Evaluating the approach on three public datasets showed that CNN method is effective for person re-id. The other features extraction approach is appearance-model based techniques, which is widely studied problem, most recently in [2][3]. Subject appearance includes soft biometrics such as clothing type and colour [4], age [5] and gender [6]. Furthermore, features extracted from the body shape are considered as appearance model features, such as height [7]. Shape descriptors are additionally used as appearance model techniques to represent the content of the shape within an image. These descriptors are manly categorized into two types: boundaries-based shape descriptors and region-based shape descriptors.

Boundary-based shape descriptors are solely rely on representing the information of the points existing in the contour of the shape, which means that the inner content is not captured. Common boundaries-based shape descriptors are Fourier descriptor [8], curvature scale space [9], and wavelet descriptor [10].

Region-based shape descriptors are able to describe the interior properties of the shape within image as well as the boundaries points information. Therefore, region-based techniques can be applied on further applications contrasting to boundaries-based shape descriptors. Region-based descriptors include Zernike moments [11], and Generic Fourier descriptor (GFD) [8] described in the methodology section. Such feature extraction methods can be applied on the entire body shape or on part of it. On this assumption, shape segmentation techniques are required.

A human body segmentation approach has been used in different modalities. For instance, [12] segmented human silhouette horizontally into three parts for gait recognition. Such segmentation allowed investigation the lower part of the body (leg movement). The results showed that significant gait recognition performance is achieved with this segmentation. Moreover, [13] segmented the body into head, torso and legs to extract the colour of each part for authentication purposes. This was performed by finding the colour difference at the neck and the waist. This approach was used to support traditional biometrics, and it showed improvements to the authentication procedure. In this paper, the human silhouette is segmented

into eight parts. The GFD is then implemented on each segment for features extraction for the task of person re-id.

II. METHEDOLOGY

A. Dataset

The CASIA dataset [14], Dataset B, was used in this experiment, which was captured by National Laboratory of Pattern Recognition, Institute of Automation, The Chinese Academy of Sciences 2006. It includes video recording of 124 subjects that was captured in an indoor environment. Subject's movement is captured from 11 angles. In this experiment, angle 0° and 90° were chosen for analysing, because they differentially represent the views of a walking person. This aids investigating the reliable parts of the body for the task of person re-id in public areas. In 0° frames, the subjects were walking towards a fixed camera. While in 90° frames the videos showed a walking motion 90° to the camera, where the subject walk right to left.

The CASIA dataset provides the corresponding foreground records for every video sequence. Silhouettes are in the PNG format with size of 320x240. We, in addition, pick 82 subjects from the first view 0° and 88 subjects from the second view 90°, each of which has between 10 and 50 silhouette frames. The reason behind this selection is to nominate subjects where their corresponding silhouette frames are clear from noise and corruption.

B. Automated Segmentation

The CASIA dataset silhouettes from two different views 0° and 90° were algorithmically segmented. The procedure of the segmentation was as follow: we surrounded the silhouette with the smallest bounding box, where the horizontal and vertical sides of the bounding box represent x and y axis of the image respectively. That means the size of the bounding boxes are different depending on the silhouettes size. Then, we segmented the silhouettes in eight different ways. These segments were chosen to represent different part of a human body in motion, so that we can investigate which of them deliver reliable re-id data.

Segments height measured based on the vertical axis of the bounding box y. Segments are head & neck, shoulders, upper half, lower half, upper quarter, middle 35%, lower 40% and whole silhouette.

To automate the segmentation over the dataset frames, we define these segments using the language of mathematical morphology [15]. The top segments (i.e. head & neck, upper half, upper quarter) are segmented as:

$$I(x,y) = \begin{Bmatrix} (x,y)|x_c - \frac{w}{2} \le x \le \frac{w}{2}, \\ y_c + \frac{h}{2} \le y \le y_c + \frac{h}{2} - h\epsilon_1 \end{Bmatrix} \quad (1)$$

Where $I(x,y)$ is the bounding box frame with width w and height h. Its centre is (x_c, y_c) and it can be computed as:

$$(x_c, y_c) = \left(\frac{w}{2}, \frac{h}{2}\right) \quad (2)$$

The middle segments, which include shoulders and middle 35%, are computed as follow:

$$I(x,y) = \begin{Bmatrix} (x,y)|x_c - \frac{w}{2} \le x \le \frac{w}{2}, \\ y_c + \frac{h}{2} - h\epsilon_1 < y \le y_c + \frac{h}{2} - h(\epsilon_1 + \epsilon_2) \end{Bmatrix} \quad (3)$$

The bottom segments are lower half and lower 40% and they can be extracted as follow:

$$I(x,y) = \begin{Bmatrix} (x,y)|x_c - \frac{w}{2} \le x \le \frac{w}{2}, \\ y_c + \frac{h}{2} - h(\epsilon_1 + \epsilon_2) < y \le y_c - \frac{h}{2} \end{Bmatrix} \quad (4)$$

ϵ_1 and ϵ_2 are constants that can determine the height of each segment to the entire bounding box.

Table 1 shows samples of each segment from both angles (0° and 90°). Furthermore, it describes the location of the segments in respect to the bounding box. For each segment, there is a fixed percentage (or height). The reason behind these specification associated to each segment is to precisely cover certain areas stability of the body in motion. For example, in the upper half (and lower half), the hand is included but, in the upper quarter (and lower 40%), the hand is excluded. In these cases, we can investigate how hand movement influence the re-id task.

C. Feature Extraction

Figure 1 shows the framework followed in this experiment. After obtaining the eight segments, a feature vector for each segment are extracted by applying GFD [8]. This descriptor extracts the properties on the contour and the region of the silhouette. GFD is used in this experiment because it outperforms common shape descriptors [8]. It also generates a fixed length features vector. This helps towards avoiding preprocessing in the classification stage, where a fixed length features vector is required. This section describes how GFD was calculated.

The regional-based technique GFD extracts a spectral feature in the frequency domain to produce the feature vectors. First, the image is converted from Cartesian space to 2D rectangular image in polar space. The polar coordinates (r, θ) can be extracted from Cartesian co-ordinates (x, y) as:

$$r = \sqrt{(x - x_c)^2 + (y - y_c)^2} \quad (5)$$

$$\theta = arctan\left(\frac{y - y_c}{x - y_c}\right) \quad (6)$$

Where (x_c, y_c) is the centroid of the foreground image. A 2D Discrete Fourier transform (DFT) is then applied on the polar images to extract Fourier coefficients which are used to create the feature vector that measures the similarity and represent the shape.

Table 1: Silhouette Segmentation

Segment Name	Segment Height	Segment Description	Angle 0°	Angle 90°
Head and neck	15%	Head and neck motion is contained in this segment. It represents the top 15% of the bounding box.		
Shoulders	14%	It only comprises the shoulders motion of the body, which is 14% of the body starting after the head and neck segment.		
Upper half	50%	Represents the arms and major hands movement of body. It locates at the top 50% of the bounding box.		
Lower half	50%	Shows the motion in the lower half of the body, which represent the lowest 50% of the bounding box.		
Upper quarter	25%	Represents the combination of two segments, head and neck segment and shoulders segment, which takes place at the top 25% of the bounding box.		
Middle	35%	Represents the movement in the middle part of the body, excluding the head, neck, shoulders and legs motion. This segment is 35% of the bounding box that starts after the shoulders, head and neck segments.		
Lower	40%	Represents legs movement of the body excluding hands and arms movements. It locates at the lowest 40% of the bounding box.		
Body	100%	The complete bounding box covers the entire body motion.		

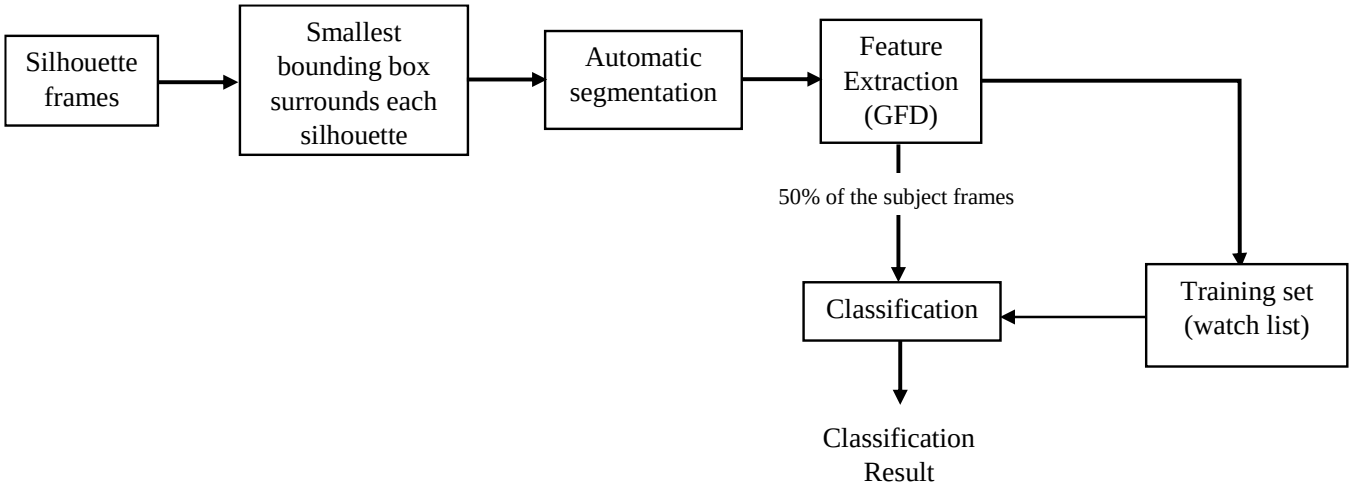

Figure 1: Experimental Framework

$$PF(\rho,\tau) = \sum_r \sum_\theta f(r,\theta) exp^{-j2\pi\left(\frac{r}{R}\rho + \frac{\theta}{T}\tau\right)} \quad (7)$$

Where R is the radial resolution and T is the angular frequency. Then, for each segment, the generated feature vector is represented as followed:

$$GFD = \left(\frac{PF(0,1)}{PF(0,0)}, \dots, \frac{PF(0,T-1)}{PF(0,0)}, \dots, \frac{PF(0,1)}{PF(0,0)}, \dots, \frac{PF(R-1,T-1)}{PF(0,0)}\right)$$

D. Classification

A linear discriminant analysis (LDA) classifier [16] was applied to each image segment type individually. A separate GFD feature vector is extracted for each frame. For a particular segment type, the vectors are split into two groups: an input set (test) and watch list (training). Each set is randomly assigned a 50% of the features vectors of each subject. After training the LDA on the input set, the error rate is then computed by testing the input set against the watch list.

The value of each rank was computed as follow: 1) a matrix of $n \times m$ (where n is the number of GFD samples in test set and m equals to the number of the samples in the training set) is obtained. 2) This matrix contains the similarity scores of each sample in the test set against each sample in the training set. 3) The values (scores) in each column of the matrix is then sorted in descending order.

The number of the GFD samples for each class varies between 10 and 50 based on the number of the frames available for each subject in the dataset. In order to obtain the ranks, each cumulative distance score per class in the training set is divided by number of training GFDs for that subject.

III. Experimental Results

The re-id accuracy rate results are stated using the Cumulative Match Characteristic curves (CMC). The CMC curves report the re-id accuracy rates achieved at every rank (the number of the ranks equal the number of candidates). Rank 1, for example, plots the percentage of the correct matches in the test set against the training test. Rank 10 refers to the percentage of the correct matches in the test set against the training set among the first 10 candidates.

Figure 2 and 3 show the resulting CMC curves of 0° and 90° angles respectively of the 8 human silhouette segments. The CMC curves revealed significant findings. Although the difference between the silhouettes was undistinguishable, GFD was able to extract discriminative features for each subject in the dataset based on certain segments.

A. 0° Angle

The most stable segment of a body in motion from the front is the upper half. Figure 2 shows that this segment outperforms all other parts with a 10th rank match 97.55%. The logical explanation of this result is that removing the segment with the most movement, which is the lower part of the body, increased the overall stability.

The whole body and upper quarter segments occupy the second and third position respectively. As shown in figure 2, these segments are close to the upper half. Leg movement prevent the whole-body segment increasing the shape's reliability to deliver stable data. The *upper quarter's* segment has a physically smaller size compared with the upper half and whole-body segments, which reduces the amount of the extracted features. Therefore, the upper quarter is only 95.7% in the 10th rank.

The middle 35%, with the head, neck, and shoulders segments attain 10th rank match with approximate 82%. Compared to the upper half, these three segments are much smaller in size, which cause fewer features to be extracted. Lower half and lower 40% segments achieve weak results due to the strong leg movements.

B. 90° Angle

The 90° angle showed that the most stable parts of the body are different from the 0° angle. The upper quarter segment outperforms other segments in the 90° side with a 98%

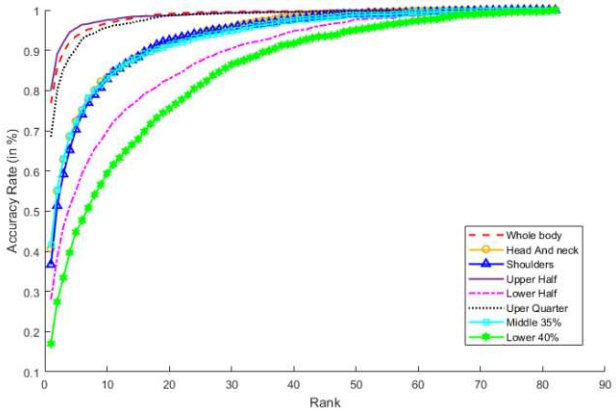

Figure 2: CMC curves showing the most stable segment of the human body from a 0° angle

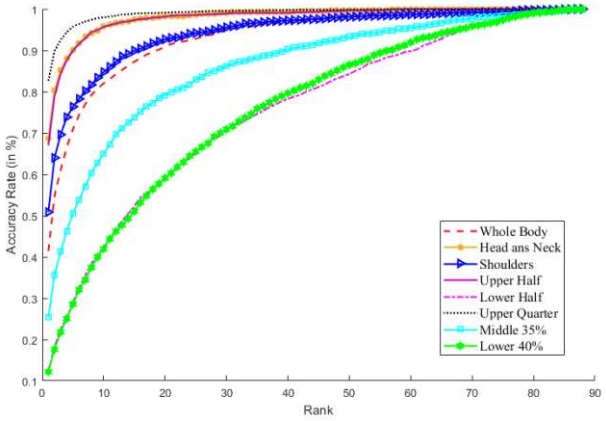

Figure 3: CMC curves showing the most stable segment of the human body from a 90° angle

accuracy rate, in the 10th rank. The upper half and head and neck overlap in the second position with a 95.7% accuracy rate. Regarding the middle 35%, the lower half and lower 40% parts represent weak results, as they involve comprehensive movements of the body.

IV. Conclusion

Person re-identification in public areas within non-overlapped camera networks can be considered as a challenging task. The purpose beyond this study was to investigate the part of a walking person sequence that produces the most reliable re-identification data. GFD was able to extract discriminative features on certain body segments. These features were then tested using LDA classifier. The result shows that upper half segment outperformed other parts from front side of the subject (angle 0°). However, the upper quarter segment was the most stable part of human silhouette from angle 90°.

This experiment proves that exploiting the silhouettes in performing person re-id overcomes many obstacles that negatively affected the performance, such as illumination changes, variations in camera resolutions, field of view and human natural motion.

Further studies will explore these results from a wider set of view positions and will facilitate the construction of a unique signature of each person monitored by CCTV for the purposes of re-identification.

Acknowledgment

This work has been co-funded by the Ministry of Saudi Arabia, King Faisal University.

References

[1] D. Zhang, W. Wu, H. Cheng, R. Zhang, Z. Dong, and Z. Cai, "Image-to-Video Person Re-Identification with Temporally Memorized Similarity Learning," IEEE Trans. Circuits Syst. Video Technol., vol. XX, no. XX, pp. 1–1, 2017.

[2] M. S. Nixon, P. L. Correia, K. Nasrollahi, T. B. Moeslund, A. Hadid, and M. Tistarelli, "On soft biometrics," Pattern Recognit. Lett., vol. 68, pp. 218–230, 2015.

[3] A. Dantcheva, P. Elia, and A. Ross, "What else does your biometric data reveal? A survey on soft biometrics," IEEE Trans. Inf. Forensics Secur., vol. 6013, no. c, pp. 1–1, 2015.

[4] E. S. Jaha and M. S. Nixon, "From Clothing to Identity: Manual and Automatic Soft Biometrics," IEEE Trans. Inf. Forensics Secur., vol. 11, no. 10, pp. 2377–2390, 2016.

[5] Yongxin Ge, Jiwen Lu, Xin Feng, and Dan Yang, "Body-based human age estimation at a distance," in 2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2013, pp. 1–4.

[6] M. Collins, J. Zhang, P. Miller, and H. Wang, "Full body image feature representations for gender profiling," 2009 IEEE 12th Int. Conf. Comput. Vis. Work. ICCV Work. 2009, pp. 1235–1242, 2009.

[7] N. Ramstrand, S. Ramstrand, P. Brolund, K. Norell, and P. Bergstrom, "Relative effects of posture and activity on human height estimation from surveillance footage," Forensic Sci. Int., vol. 212, no. 1–3, pp. 27–31, 2011.

[8] D. Zhang and G. Lu, "Shape-based image retrieval using generic Fourier descriptor," Signal Process. Image Commun., vol. 17, no. 10, pp. 825–848, Nov. 2002.

[9] A. Mackworth, "Scale-Based Description and Recognition of Planar Curves and Two-Dimensional Shapes," no. 1, 1986.

[10] Q. M. Tieng and W. W. Boles, "Recognition of 2D object contours using the wavelet transform zero-crossing representation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 8, pp. 910–916, 1997.

[11] C.-H. Teh and R. T. Chin, "On image analysis by the

methods of moments," IEEE Trans. Pattern Anal. Mach. Intell., vol. 10, no. 4, pp. 496–513, 1988.

[12] I. Venkat and P. De Wilde, "Robust Gait Recognition by Learning and Exploiting Sub-gait Characteristics," pp. 7–23, 2011.

[13] S. Denman, C. Fookes, A. Bialkowski, and S. Sridharan, "Soft-biometrics: Unconstrained authentication in a surveillance environment," DICTA 2009 - Digit. Image Comput. Tech. Appl., pp. 196–203, 2009.

[14] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in Proceedings - International Conference on Pattern Recognition, 2006, vol. 4, pp. 441–444.

[15] R. Gonzalez and R. Woods, Digital image processing. 2002.

[16] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," Ann. Eugen., vol. 7, no. 2, pp. 179–188, Sep. 1936.