Pattern Recognition Letters 68 (2015) 297–305

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Shape Context for soft biometrics in person re-identification and database retrieval☆

Athira Nambiar*, Alexandre Bernardino, Jacinto Nascimento

Institute for Systems and Robotics, Instituto Superior Técnico, Av. Rovisco Pais, 1, Lisbon, 1049-001, Portugal

ARTICLE INFO

Article history:
Available online 11 July 2015

Keywords:
Multimedia
Soft biometrics
Shape Context
Re-Identification
Silhouettes
Retrieval
Surveillance

ABSTRACT

We introduce a novel descriptor for the analysis of pedestrians and its applications to person re-identification and database retrieval. A Shape Context descriptor of the head-torso region of persons' silhouettes is shown to have a very good discrimination ability and application to re-identification. For database retrieval using human queries, we train a map from the Shape Context to interpretable soft biometric quantities that can be reasoned about by humans. We show that a good linear correlation exists between Shape Context descriptors and soft biometrics quantities in the upper human torso and illustrate its application to retrieval in databases from human queries. Shape Context to biometrics maps are learned from virtual avatars rendered by computer graphics engines, to circumvent the need for time-consuming manual labelling of data sets. We obtained promising results of Shape Context based person re-identification and database retrieval from human compliant description of biometric traits, in both synthetic data and real imagery.

1. Introduction

Soft biometrics are the physical, behavioral or adhered human characteristics, classifiable in predefined human compliant categories which are established and time proven by humans with the aim of differentiating individuals [7]. Different from hard biometrics, they lack the distinctiveness and permanence to identify an individual with high reliability. However, they have certain advantages over hard biometrics exclusively making them best suited to deploy in surveillance applications viz. non obtrusiveness, acquisition from distance, non-requirement for the cooperation of the subject, computational and time efficiency and human compliance.

Soft biometric features leverage human characteristic traits such as height, body size and gait. These characteristics are more coherent and reliable for long term re-identification than the commonly used temporary appearance cues such as dress color and texture info [5]. Recently, the arrival of sophisticated systems such as motion capturing devices, 3D sensors (Kinect) and high definition cameras accelerated the exploitation of soft biometrics in wide range. As a result, unprecedented real time applications were reported in person re-identification and other video surveillance applications.

However the direct computation of soft biometric features from video images is not trivial and existing methods rely on human manual measurements made on individual images [17]. Instead, automated computer vision analysis methods have been more successful with features that are not interpretable by humans, like SIFT [13], HOG [6], Shape Context [4] and others. These features, though useful in automated methods, are hard to reason about by humans and thus not suited for formulating verbal descriptions of search queries in databases. For instance, we would like to be able to search on a database for persons with large torso, thin neck, long head, etc. Thus, we propose a methodology to infer soft biometric person characteristics from their computer vision based descriptors, using regression analysis.

Obtaining a predictive model of soft biometric features from computer vision features involves several challenges and difficulties: (i) which computer vision features are more adequate; (ii) how to obtain the ground truth biometric features to train the model and; (iii) which regression model is more suitable.

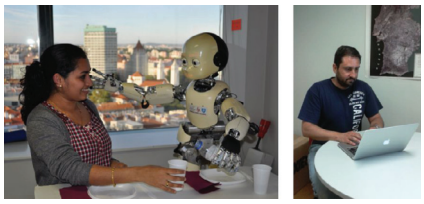
With respect to the first point, we propose the use of Shape Context features computed in the upper-torso part of the frontal human silhouettes, where we capture the human images from their video clips walking towards the camera. The upper torso region of the body presents less temporal variance with respect to arms and legs motions, thus producing more stable features. In addition to that, since person re-identification is carried out in an uncontrolled environment, there are chances for clutters and other interacting objects making the lower body part occluded. In many indoor surveillance systems cameras are placed along corridors at high positions and tilted down, which makes the legs and lower torso occluded when persons are close to the camera. However, the head to chest

☆ This paper has been recommended for acceptance by Paulo Lobato Correia and Thomas Moeslund.

* Corresponding author. Tel.: +351 218418050.
E-mail address: anambiar@isr.ist.utl.pt, nambiar.athira@gmail.com (A. Nambiar).

http://dx.doi.org/10.1016/j.patrec.2015.07.001
0167-8655/

(a)



(b)

Fig. 1. Images showing the relevance of head-to-chest region for person re-identification tasks in various situations. (a) A forward walking sequence captured in our HDA data set [14] highlights the visibility of head-to-chest region in most of the frames, while the other parts are occluded. (b) Relevance of Upper body region in Human-Robot/Computer interaction. Head to chest region are clearly visible in both cases and could be used for re-identification task.

region, unlike the waist and legs, maintain a relatively consistent shape through a broader range of walking frames. A real world example (video sequence in the HDA Data set[1]) where this effect is clear is shown in Fig. 1(a). In addition to that, there are scenarios in Human-Robot Interaction (Fig. 1(b) left) and Human-Computer Interaction (Fig. 1(b) right) that highlight the relevance of upper body part selection for person re-identification.

The use of a silhouette based feature is motivated by the fact that it is less sensitive to the color and texture of the inner region of person's images and thus making itself a better candidate towards long term based person re-identification. Furthermore, the Shape Context feature computes the density of boundary points at various distances and angles. As such, it more directly encodes soft biometric traits such as lengths, curvatures and size ratios in the human body. In this paper, we explore this idea with the goal of recovering the soft biometric features encrypted in the Shape Context descriptors of body silhouettes using regression methods.

The second challenge is the availability of ground truth biometric features to train the model in the regression analysis. It is not easy to model this in a real environment due to the necessity of a range of variations of discriminative biometric features in relatively large population. Also, it is laborious to annotate the human biometrics manually on real data. In order to tackle this issue, we used synthetic avatars in a virtual reality platform. In contrast to [17], where the training set was generated by manual annotations done on real imagery by a large number of human annotators, here we avoid such a troublesome training phase by generating the ground truth with the help of modern computer graphics technology.

In this work we leverage on the ability to simulate thousands of variations in biometrics on avatars according to our choice for two purposes. First, we conduct a baseline study to verify the impact of our descriptor for re-identification (Re-ID), since the simulated avatars provide flawless silhouette images. Second, we model the regression between computer vision based features (Shape Context) and human interpretable features (Biometrics) and thus bridge the gap between the human and machine interpretations of human body shape. Thus we present a novel automatic person retrieval system which could work in dual mode (viz., multimedia mode or human query mode) depending on the test data.

For obtaining some geometric features of the head to chest region, distinguishable from person to person, some measurable metrics which vary significantly within the population should be chosen. The measurement and study of such features and their variation is the domain of anthropometry. An anthropometric survey (ANSUR) was conducted by the U.S. military in 1988 upon more than 150 anthropometric dimensions, measured from 9,000 soldiers. A statistical summary of those standard biometric features related to the upper torso regions is provided in Table 1. In our study, we consider some of those key biometric features described here.

Table 1
A summary of anthropometric data taken from [9] relevant in the upper torso region. These statistical summaries reveals significant variation in the head and chest measures (All measurements are in centimeters.) The features shown in bold letters are some of the soft biometric cues used in our study.

Measurement name	Mean	Standard deviation	Min	Max
Biacromial breadth	39.70	1.80	33.0	45.10
Bideltoid breadth	49.18	2.59	41.0	59.3
Head width	15.51	0.60	13.6	17.7
Head circumference	56.77	1.54	51.4	62.7
Head length	20.02	0.72	17.6	22.6
Chest breadth	32.15	2.55	25.70	42.20
Neck-bustpoint length	27.24	1.81	22.2	34.2
Neck circumference	37.96	1.97	31.6	47.0
Shoulder circumference	117.52	6.04	96.6	142.4
Shoulder-elbow length	36.9	1.79	29.7	44.6
Shoulder length	15.05	1.10	11.4	18.5

Concerning the regression model, several choices are possible to be used, e.g., a non-linear mapping based on Gaussian process regression [16]. However, this kind of approaches usually require a significant amount of data. In our case, we will adopt a linear approach as a baseline, to perform a mapping between the Shape Context and soft biometrics. We will experimentally demonstrate that with the available amount of data such linear approach will suffice for obtaining high accuracy in the results.

The paper is organized as follows. Section 2 describes related literature on soft biometric based person retrieval as well as former applications of Shape Context in surveillance scenarios. The system architecture is explained in detail in Section 3. In Section 4, the main methodologies used in our framework viz. Shape Context feature extraction and regression are described. Then, Section 5 explains in detail our Re-ID experiments conducted in both real and virtual platforms, by acquiring Re-ID data set and by simulating avatars respectively. The regression analysis carried out is also explained here. Section 6 presents a set of promising results accentuating the reliability of the proposed architecture in person re-identification, as well as demonstrating the human retrieval performance with human compliant queries. Finally in Section 7, we summarize our work and enumerate some future work plans.

2. Related work

2.1. Shape Context on surveillance applications

The application of Shape Context *(SC)* in human video surveillance systems are reported in the state-of-the-art. Some works are found in pedestrian detection by [12], highlighting that *SC* descriptor trained on real edge images exhibited high performance, particularly on difficult images and backgrounds. Some application of *SC* have also been employed in gait recognition [22] where *SC* is used to compute the similarity between two procrustes mean shape, which is a compact representation of gait sequence. A similar application of *SC* is found in human pose estimation [1]. However, the literature is scarce concerning the use of *SC* in re-identification applications. One exception is [21] that created shape labelled images by means of shape and appearance models which was inspired from the idea of Shape Context.

[1] http://vislab.isr.ist.utl.pt/hda-dataset/ .

Another work [11], used Shape Context descriptors to represent the intra distribution of colors for person re-identification. In this work, we propose Shape Context features computed on the contour of the silhouette of frontal images of persons.

2.2. Silhouette based person re-identification

Most of the state-of-the-art methods leverage either color information or local feature descriptors inside the human body after segmenting the silhouettes. In [20], a robust classification procedure exploited the discriminative nature of sparse representation to perform people re-identification. [2] presented a person Re-ID method based on appearance classification and silhouette part segmentation using various descriptors such as SIFT, SURF and SPIN. In this work, instead of appearance cues, we exclusively depend upon contour information and propose a new way of long-term person re-identification using silhouettes. To the best of our knowledge no complete work for person re-identification, leveraging solely the edge information of the silhouette is reported in the literature.

2.3. Soft biometrics based person retrieval

The last decade witnessed pioneering research in video surveillance applications using soft biometry which was enhanced with the introduction of advanced motion capturing devices and high definition cameras. [3] presented a set of 3D soft biometric cues related to anthropometric measurements, obtained from KINECT RGB-D sensors and employed in person re-identification. Many studies on gait based person recognition and re-identification were also reported [8,15,18]. In [7] a bag of soft biometric traits (e.g., facial and body soft biometrics) was presented for person re-identification. They proposed a general framework by integrating both the primary biometrics (i.e. face, iris) and soft biometric system (i.e. height, gender). The retrieval using soft biometrics is also addressed in [17]. In that work, a novel method of comparative human descriptions for soft biometrics was introduced, in which manual annotations of comparative biometric measurements were collected. Out of these measurements, relative biometric measurements were inferred and exploited for retrieval. However, our retrieval system using soft biometrics neither requires very high image resolution as required for facial features in [7], nor laborious manual annotations over real world data set as in [17]. Here we propose a novel automatic person identification system exploiting machine learning technique and modern computer graphics technology.

3. System architecture

Our scheme is designed to work in two different modes depending on application scenarios. In the first scenario (***Scenario#1***), the test images/ videos of the subject to identify are provided, whereas in the second mode (***Scenario#2***) the probe is solely a query or description of the subject provided by a human operator. The former scenario mainly concerns the use of multimedia content for re-identification of a suspect in a video surveillance network by extracting his feature descriptors and matching with gallery database. The latter scenario instead does not require any multimedia content but exploits the eyewitness description of the suspect related to biometrics cues such as short neck, large chest etc. We note here that these human descriptions are analogous to *human compliant labeling* referred by [7] or *semantic annotation* referred by [19]. Rather than re-identifying, this mode is more pertinent towards categorizing the population based on their respective human compliant traits and thus retrieving their identifier(*#ID*). Since many people could have similar semantic labels resulting in subject interference, grouping them into classes with similar traits could be the best technique to tackle this issue. This is a kind of pruning method, which normally the security people do manually on receiving the human queries; we do it here automatically.

The general framework of the system is presented in Fig. 2. In the training phase, human video footage is acquired in a video surveillance system and stored in gallery. The camera network is connected to the *SC* descriptor module, where the acquired persons' *SC* features are extracted. These extracted *SC* features are stored in a gallery database for later use. The database is accessible by the feature matching module, which has the purpose to compare the feature descriptor of the person we want to identify (i.e., the probe) with the ones stored in the database. In ***Scenario#1***, when a new image frame of the person is acquired, his *SC* descriptor is extracted and compared with those in the gallery set. In the decision module, based on the matching similarity measurements, the most similar person *ID* in the train set is retrieved thus facilitating the system for person re-identification.

Another major module of the system in the training phase is a regression block connected to the database of *SC* features. It divulges the relation of *SC* descriptors with soft biometrics, and it estimates biometric values (*BF*) corresponding to each sample. These estimated biometric values are stored in a gallery database of biometrics, which is connected to decision module. The decision module analyses these biometric data and carries out a statistical analysis among the population. In ***Scenario#2***, when the probe input in terms of human query

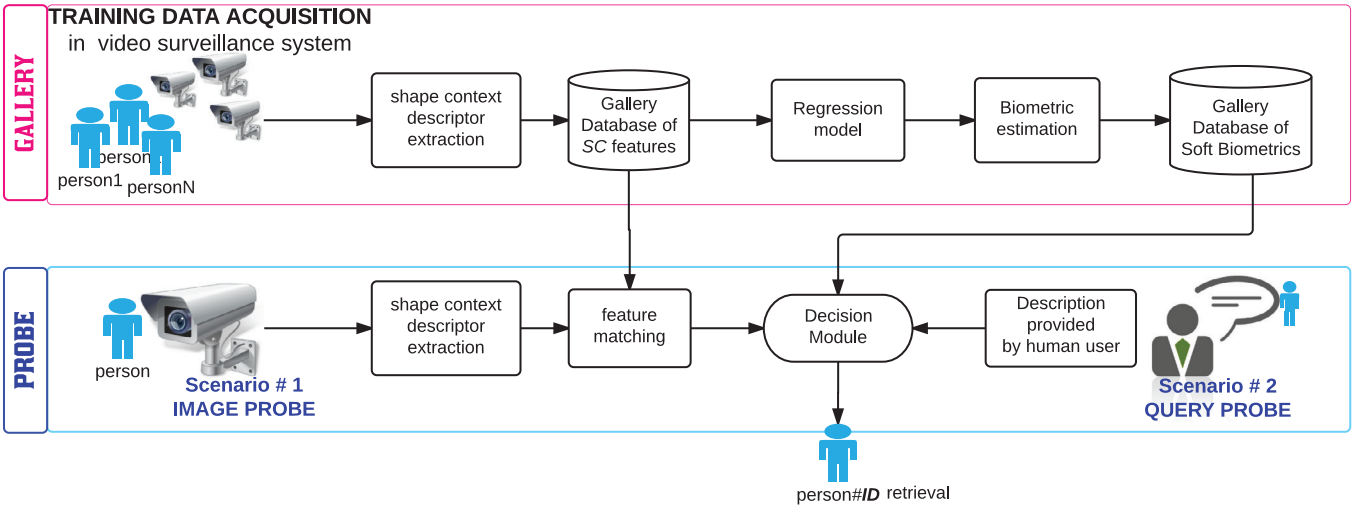


Fig. 2. The scheme presents the framework of our human identification system. The probe data can be either the images/videos of the subject to identify (*Scenario#1*), or a description of the subject provided by a human operator such as eyewitness statement in a criminal scene (*Scenario#2*).

enters in the decision module, it will examine the statistical profile of the population and retrieve the category of suspect. As a result, all the person *IDs*' in the suspected category as well as the tentative ranked list of suspect are published.

To learn the regression module, we require a vast and vivid benchmarking data set to substantiate the mapping between *SC* feature space and biometric space. For that purpose, we generate avatars in virtual reality and carry out regression analysis. When a new *SC* feature is received, the corresponding output biometric values are estimated based on this regression model. In addition to that, a virtual reality population is also employed to verify our methodology towards person re-identification and to compare with the counterpart experiment in real scenario. A more detailed description of simulation of avatars and their application in both ***Scenario#1*** and ***Scenario#2*** is given in section 5.1 and section 5.2, respectively.

4. Proposed methodology

This section describes the main ingredients of the proposed approach. Basically it comprises: (i) computation of the Shape Context features from the images containing the head and torso, (ii) matching the Shape Context between two head-torso silhouettes and (iii) the statistical regression analysis between Shape Context and the space of biometric features.

4.1. Shape Context

The original idea of Shape Context was described in the paper of [4]. In order to achieve the shape similarity or the shape distance, they introduced a new descriptor called Shape Context which measures the distribution of points in a shape relative to each point in that shape.

Fig. 3 depicts the method of obtaining Shape Context descriptors. The silhouette of an object is sampled at N discrete points along the contours, $P = (p_1, p_2, \ldots, p_N)$. For a point p_i, a coarse histogram h_i of the relative co-ordinates of the remaining $N-1$ points is identified and is termed as the Shape Context of p_i:

$$h_i(k) = \#\{q \neq p_i : (q - p_i) \in bin(k)\} \tag{1}$$

Thus, a compact and highly discriminative descriptor is computed as the distribution over these relative positions. A uniform binning scheme in log-polar space is adopted making the descriptor much more sensitive to nearby sample points than to those farther away. As shown in Fig. 3(c), we use 12 equally spaced angle bins and 5 equally spaced log-radius bins, altogether making the dimension of the *SC* as 60. In contrast to the closed object shapes proposed in the original work, we apply the re-sampling on the open shape of the silhouette, i.e., we do not consider the cropping line in the chest. To represent each silhouette (Fig. 3(a)) we used 40 points uniformly sampled from the Canny edges (Fig. 3(b)). Then we flattened and concatenated the complete set of 40 sample point *SC* each with 60 dimensions, thus producing a *SC* histogram of dimension 2400.

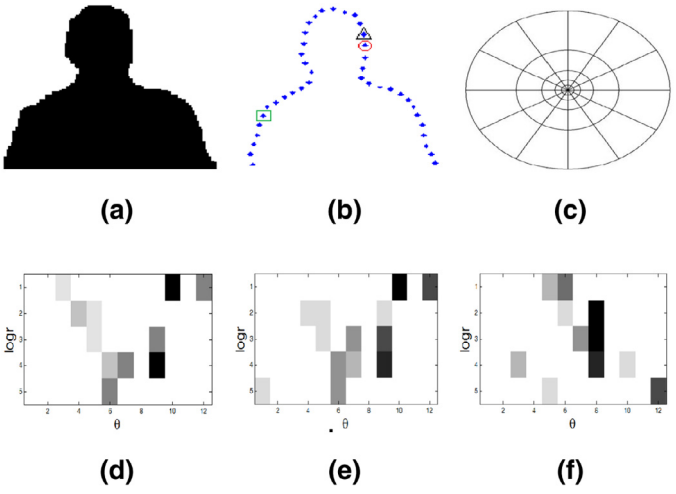


Fig. 3. Shape Context computation. (a) Silhouette of upper human body part. (b) Sampled edge points of the silhouette shape. (c) Diagram of log-polar histogram bins used in computing the Shape Contexts. We have used five bins for *logr* and 12 bins for θ. (d–f) correspond to the Shape Contexts for reference samples marked by △, ○ and □. Visual similarity of the Shape Context for nearby points △, ○ is pretty obvious whereas the Shape Context of the □ point, is quite different (note: dark = large value).

4.2. Matching Shape Context

In order to compare two different shapes we must define a similarity metric. To mitigate problems of misalignments of the silhouettes' sampling points due to discretization, a previous alignment step is necessary. Two criteria are to be met while matching *SC* features: (1) corresponding points should have very similar descriptors, and (2) the correspondences should be unique.

First criteria is handled via cost matching technique. Let C_{ij} denote the cost of matching two sample points p_i and q_j in two different shapes, by means of χ^2 test statistics

$$C_{ij} = C(p_i, q_j) = \frac{1}{2}\sum_{k=1}^{K}\frac{[h_i(k) - h_j(k)]^2}{[h_i(k) + h_j(k)]} \tag{2}$$

where, $h_i(k)$ and $h_j(k)$ denote the K-bin normalized histogram at p_i and q_j. Given the set of costs C_{ij} between all pairs of points, the uniqueness criterion is addressed as follows. To match two shape contours say, P and Q, we minimize the total cost of matching

$$H(\pi) = \sum_i C(p_i, q_{\pi(i)}) \tag{3}$$

subject to the constraint that the matching is one-to-one, i.e., π is a permutation. This is an instance of the square assignment (or weighted bipartite matching) problem. In our experiments, we make use of the Hungarian algorithm [10].

4.3. Regression

Regression analysis is a statistical process for estimating the relationships among variables. The technique is widely used in machine learning for prediction and forecasting. In statistics, linear regression is an approach for modelling the relationship between a scalar dependent variable and one or more explanatory variables, in which data are modelled by linear functions and unknown model parameters are estimated from data.

Suppose a linear regression is carried out from an input space of dimension $\mathbb{R}^p$ to an output space of dimension $\mathbb{R}$. Each element in the input space is a feature vector of size $p \times 1$. i.e. $\mathbf{x} = \left[x^1, \ldots, x^p\right]^T$. We collect n such samples and represent it as a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ as follows:

$$\mathbf{X} = [\mathbf{x}_1 \mid \cdots \mid \mathbf{x}_n]^T \tag{4}$$

Each row in the $\mathbf{X}$ matrix represents a feature vector corresponding to the nth sample in the data set. We collect the response variable y_i corresponding to each input sample $\mathbf{x}$ and represent it as a vector $\mathbf{y}$ in the output space

$$\mathbf{y} = \left[y_1, y_2, y_3, \ldots, y_n\right]^T \in \mathbb{R}^{n \times 1} \tag{5}$$

The sample mean is $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}$ and the preprocessed centered data with zero mean is $\mathbf{X}' = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T$, where $\mathbf{1}$ is a vector of ones of dimension n. Similarly, their counterparts in output space are $\bar{y}$ and

$\mathbf{y}'$. Now, the least squares regression computes the weight vector $\mathbf{w}_{LS}$ that minimizes the error of fit:

$$\hat{\mathbf{w}}_{LS}(\mathbf{y}') = \arg\min_{\mathbf{w}\in\mathbb{R}^p} \sum_{i=1}^{n} (y'_i - \mathbf{w}^T\mathbf{x}'_i)^2 = \arg\min_{\mathbf{w}\in\mathbb{R}^p} \|\mathbf{y}' - \mathbf{X}'\mathbf{w}\|_2^2 \tag{6}$$

Solving (6) leads to the following linear regression solution:

$$\hat{\mathbf{w}}_{LS}(\mathbf{y}') = (\mathbf{X}'^T\mathbf{X}')^{-1}\mathbf{X}'^T\mathbf{y}' = \mathbf{X}'^{+}\mathbf{y}' \tag{7}$$

where $\mathbf{X}'^{+}$ is the pseudo inverse of $\mathbf{X}'$.

We apply Principal Component Analysis (*PCA*) for dimensionality reduction of the input space into $\mathbb{R}^d$. Eigenvalue decomposition of the covariance of $\boldsymbol{X}'$ produces the eigenvectors

$$\mathbf{V} = \left[\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_p\right] \in \mathbb{R}^{p\times p} \tag{8}$$

of which the reduced eigenvectors are the first d columns.

$$\mathbf{V}' = \left[\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_d\right] \in \mathbb{R}^{p\times d} \tag{9}$$

Then the reduced scores corresponding to the Principal Components are computed by the relation $\mathbf{Z}' = \mathbf{X}'\mathbf{V}'$. Thus, in the Principal Component egression method, the Eq. (7) turns out to be

$$\hat{\mathbf{w}}_{LS}(\mathbf{y}') = (\mathbf{Z}'^T\mathbf{Z}')^{-1}\mathbf{Z}'^T\mathbf{y}' \tag{10}$$

When a new input variable $\mathbf{x}_{new}$ is available, the regression model established above is used to estimate the predicted output variable, using the equation:

$$\hat{y}_{new} = (\mathbf{x}_{new} - \bar{\mathbf{x}})^T.\mathbf{V}'.\hat{\mathbf{w}}_{LS}(\mathbf{y}') + \bar{y} \tag{11}$$

Statistics such as estimate of *error variance* R^2, also known as the *coefficient of determination*, acts as the metric to check the performance of regression modelling. It is formally defined as

$$R^2 \equiv 1 - \frac{SS^{res}}{SS^{tot}}, \tag{12}$$

where SS^{res} is the residual sum of squares, measuring the discrepancy between the data and an estimation model and SS_{tot} is the total sum of squares, i.e., the sum of the squares of the difference of the dependent variable and its mean. It is a very important indicator to state if the regression is efficient while it informs the goodness of fit of a model. R^2 represents the percent of the data that is the closest to the line of best fit.

5. Experiments

We conducted the experiments in two modes, as mentioned in Section 3. First, we carry out experiments in ***Scenario#1*** to study the feasibility of upper torso Shape Context for person re-identification. Initially we conduct a study with an existing person Re-ID data set. However, the real world scenario is prone to segmentation noise. Thus, in order to validate our system in a noise free environment, we conducted our second experiment in a simulator platform, using virtual reality avatars. We simulated custom avatars corresponding to the humans in the real world, and conducted our experiments on them as well.

The second mode of experiments is done in ***Scenario#2***. Here, we explore the relationship between Shape Context descriptors and soft biometrics by means of regression. Thus, we bridge the gap between human and the machine definition of biometrics with aid of computer vision and machine learning techniques. One noteworthy aspect is that in both experimental modes, the system does not require the co-operation of the subject as in hard-biometric data acquisition, thus making this soft biometric system very suitable for surveillance applications, where such cooperation is hard to achieve.

Fig. 4. Sample images and corresponding silhouettes in our real-world experiment.

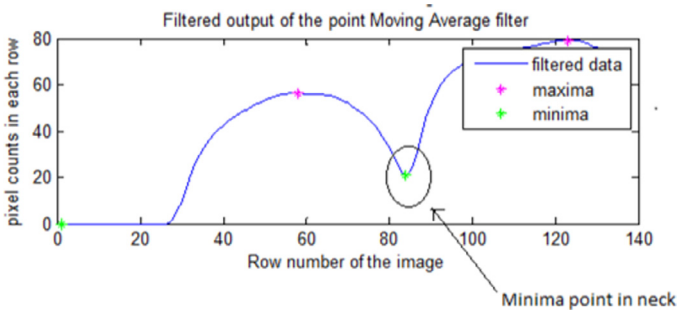


Fig. 5. Pixel count value curve facilitating the automatic cropping of the upper body region by observing the minima point in the neck.

5.1. *Scenario#1*

5.1.1. *Real data set for re-identification*

We conducted a pilot study in the real world, where we incorporate the human silhouettes captured using KINECT camera in RGB-D person re-identification Dataset[2] [3]. Along with each human image, corresponding human silhouette information is also provided. An example of the data set is seen in Fig. 4. For our studies, we made use of their 'walking1' and 'walking2' categories, where we can obtain frontal appearance of the walking people. As a case study we only used 20 people, each one with 4 samples. There are images with different backgrounds, and the same person in different dressing, thus making the data set very suitable to study the impact of our methodology in long-term person identification based on shape of the silhouette.

Since we are interested only in the upper torso region, we carry out some pre-processing in order to get the cropped images. Initially, we split the body into 2 parts and select only the region of interest, which is the upper part. Then we try to localize the neck location, which could be acted as key point. As seen in Fig. 5, the pixel count along the row of the image is plotted against the row number, which depicts the variation of the silhouette's thickness. We apply moving average filter to smooth out the fluctuations in the data curve. A key point corresponding to the neck is found by searching the minimum in the curve. Next, a standard amount of height equal to head to neck, is added towards bottom onto the chest region from the neck point in order to define the crop line in the chest. Afterwards, we normalize the height and rescale the width of the cropped region, maintaining the aspect ratio.

Prior to conducting the experiment of person re-identification, we had to apply some initial pre-processing steps to address the problem of silhouette imperfection mostly occurring due to segmentation errors and pixel noise. To get rid of the void spaces in silhouettes and to attain data quality, we applied morphological operations such as dilation followed by erosion. Afterwards, while the silhouettes are ready for our experiment, we equally split the data set into half. Former set is the training set and the latter is the test set. In real world scenario, out of 80 sample images, we have 40 samples in both gallery and probe, i.e., 2 samples per 20 different persons are made available

[2] http://www.iit.it/en/datasets-and-code/datasets/rgbdid.html .

Fig. 6. Sample instances of custom virtual avatars simulated corresponding to the real world data set.

Fig. 7. Six standard avatars used in the synthetic platform for the generation of large data set by changing the biometric features. We make use of only the upper-torso region including head, shoulder and chest.

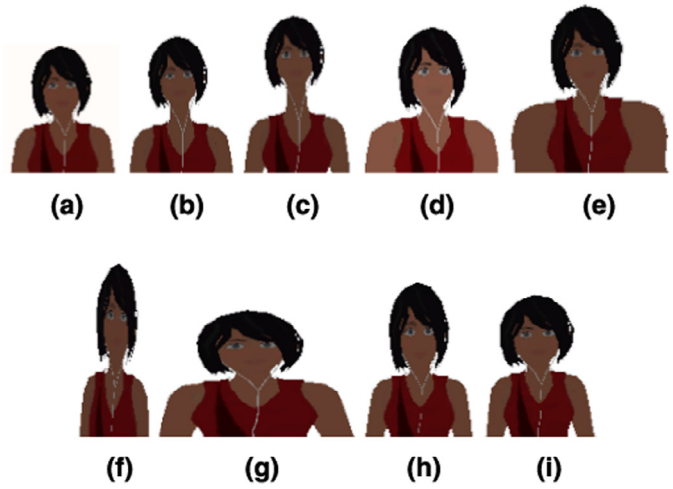


Fig. 8. The nine variations of biometrics simulated in the generic avatars. Only the upper torso region is shown since it is the region of our interest. Please refer to the Table 2 for measurement details.

in both training and test set. Afterwards, the *SC* descriptor for each silhouette in the gallery is calculated. When the test set is provided, its matching cost towards each of the 40 gallery samples is found using the Hungarian method. Then, each test sample will search for the minimal cost between itself and the gallery descriptors. The gallery sample with minimal cost corresponds to maximum similarity and is selected as the best matching.

5.1.2. Custom avatars for re-identification

In the previous section we discussed about the experiments conducted in real database. In this section we evaluate the influence of the noise of the segmentation while extracting the head-to-torso region. To perform this study, we replicate the real data set with virtual reality avatars leveraging computer graphics tool (the game engine *Unity3D*®) that allow us to render and manipulate the shape of synthetic humans. We used some standard avatar packages viz. male character pack and female character pack from Mixamo 3D[3] character animation service and Character Pack 02 from Animation arts Creative GmbH.[4] We modelled the custom avatars as close as the corresponding human instances by matching their shape traits and incorporating the posture and inclination of shoulders. Samples of the real human instances and their corresponding custom avatar models are illustrated in Fig. 6. After generating the custom avatars, we executed walking animations of these avatars and captured random 4 frames for each person which resembled the video surveillance image acquisition. Thus our virtual reality data set also consisted of 80 synthetic samples corresponding to the 20 human instances in real world experiment. Then, we split them into gallery and probe and conduct descriptor matching in the same way conducted for real world data set.

5.2. *Scenario#2*

5.2.1. Generic avatars for regression

Albeit we simulated *custom* avatars in our previous experimental setup, the data set was limited in terms of variability of biometric features since only 20 human instances were generated in the simulator. In order to compute the regression model between Shape Context features and soft biometrics, this was not enough to represent variation range of the real human population. Thus, we introduced a more global avatar set called as *generic* avatars, by imposing larger variabilities as observed in the human population. Such a generic population is preferred over custom avatars for modelling the regression map, since it covers wider ranges of features. By incorporating extremal shapes, the generic data set provides a higher Signal to Noise ratio[5] available for regression analysis.

Again we exploit the graphics engine *Unity3D*® to simulate the multiple avatars in virtual reality. Here we used six standard avatars viz. male character pack and female character pack (shown in Fig. 7) from Mixamo 3D character animation service, as the baseline avatars. The default avatars models available in the package were considered as standard models, in which we assumed a unitary scale factor of each biometric measurement(see Fig. 8(a)). Afterwards, we generated the other avatars by imposing variations to the biometric features with respect to this standard model in the *Unity3D*® platform. The scale parameters of the avatar examples are defined by analysing the variability in real world human population. Here are the biometric cues we employ in our experiment:

- Neckness (N) : length of the neck
- Chestsize (C) : horizontal distance between the lateral margins of the upper torso
- Bodysize (B) : overall body size
- Headlength (HL) : maximum vertical length of the head
- Headwidth (HW) : maximum horizontal width of the head

Table 2 shows the soft biometric parametrization imposed for simulating generic avatar population. Each value in the table corresponds to the scale applied to the standard model counterpart of that anthropometric measurement. We alter the biometric values one at a time by keeping other features intact. Thus, as per mentioned in the table, we can have 8 different modified avatar models generated out of the standard avatar, by altering each biometric feature individually. Fig. 8 shows an example of the different virtual avatar samples generated out of a single basic standard avatar.

Fig. 8 (a) is a standard avatar where all the parameters are normalized (100%). Fig. 8(b) and (c) correspond to 200% and 300% Neckness, which intuitively means those models' neck length is twice and thrice longer compared to the standard one. Fig. 8(d) and (e) illustrate the 200% and 300% chest sized avatars respectively. Thin body

[3] https://www.assetstore.unity3d.com/en/#!/publisher/150.
[4] https://www.assetstore.unity3d.com/en/#!/publisher/6659.
[5] Considering the noise in the *SC* features as constant (discretization noise), a higher variability in the range of the features (signal) will result in a better signal-to-noise ratio that will improve the quality of the regression model.

Table 2
Chart showing the soft biometric scale factors for the simulated avatar versions in Fig. 8. Values highlighted in bold characters in each row represents the modification imposed for that particular avatar.

Avatar Index	Neckness (N)	Chestsize (C)	Bodysize (B)	Headlength (HL)	Headwidth (HW)	Human description label
(a)	100%	100%	100%	100%	100%	Standard
(b)	**200%**	100%	100%	100%	100%	Large neck
(c)	**300%**	100%	100%	100%	100%	Very large neck
(d)	100%	**200%**	100%	100%	100%	Large chest
(e)	100%	**300%**	100%	100%	100%	Very large chest
(f)	100%	100%	**50%**	100%	100%	Thin body
(g)	100%	100%	**200%**	100%	100%	Fat body
(h)	100%	100%	100%	**125%**	100%	Long head
(i)	100%	100%	100%	100%	**125%**	Wide head

size and fat body are generated in Fig. 8(f) and (g) by setting scaling the body size parameter by 50% and 200% respectively. The last two avatars concentrate on the geometric parameters of head, by increasing 25% horizontally (head width) and 25% vertically (head length). Thus, we managed to generate an approximate variation of biometric features in synthetic population as observed in the human population.The idea was to be able to cover the range of variability as much as possible with the least number of examples. This way we could enhance the signal-to-noise ratio of the regression analysis.

Altogether nine variations were generated out of each of 6 standard avatar. Then, we executed walking animations and captured random 4 frames for each person which resembled the video surveillance image acquisition. Thus our *generic* avatar data set consisted of 216 images.

5.2.2. Regression model

Putting Section 4.3 in practice, we have Shape context (*SC*) descriptors correspond to input vectors $\mathbf{x}$ and Biometric features (*BF*) values correspond to output variables y. Since we have $n = 216$ sample avatars in the generic data set, the input matrix $\mathbf{X}$ is of dimension $\mathbf{216 \times 2400}$. Next, we perform *PCA* to reduce the dimension of space leading to a reduced input matrix of size $\mathbf{216 \times 60}$. Considering five biometric features exploited in our study viz. $BF = (N, C, B, HL, HW)$, we perform linear regression as described in Section 4.3 individually for each of the biometric in the set *BF*. More specifically, $\boldsymbol{y}$ in Eq. (5) will be a vector of dimension 216 containing a given biometric feature for all the avatars. Based on the Eqs. (8)–(10), the regression analysis is carried out further. Afterwards, when a new sample *SC* descriptor is provided, our model will estimate the corresponding response variables, $\hat{BF}_{new}$ using the Eq. (11).

6. Results

6.1. Person re-identification using Shape Context

Regarding both the experiments in ***Scenario#1***, the goal is to retrieve the most similar person in the gallery set for a given test person, by matching its Shape Context descriptor with those in the gallery. Or in other words, when the probe imagery of the suspect is provided, its shape similarity with all the other training images in the gallery is measured by bipartite graph matching technique on *SC* features and the person re-identification is carried out.

We depict the result of re-identification with the help of confusion matrix. Confusion matrix is a specific table layout that allows visualization of the performance of an algorithm, where each column of the matrix represents the instances in a predicted class (predicted person ID), and each row represents the instances in an actual class (actual person ID).

Our results of person Re-ID is illustrated in Fig. 9. The first result in Fig. 9(a) is the *confusion matrix* corresponding to our study with 20 real world instances and showing a re-identification accuracy of 92.5%. Fig. 9(b) is the counterpart *confusion matrix* in virtual setup with 20 custom avatars, and it achieved 95% accuracy in Re-ID. In both cases, we could observe high performance of our proposed *SC* algorithm to re-identify people. This accentuates the feasibility of utilizing shape as an effective soft-biometric cue in re-identification scenarios. Moreover, by conducting the comparative study in virtual setup, we could observe the influence of segmentation noise in reducing the Re-ID rate in the real world scenario. Real data segmentation is more irregular due to sensor measurement noise, whereas segmentation in the avatars is perfect, apart from pixel discretization errors. Thus, better segmentation methods should be sought for achieving higher accuracy in the real world.

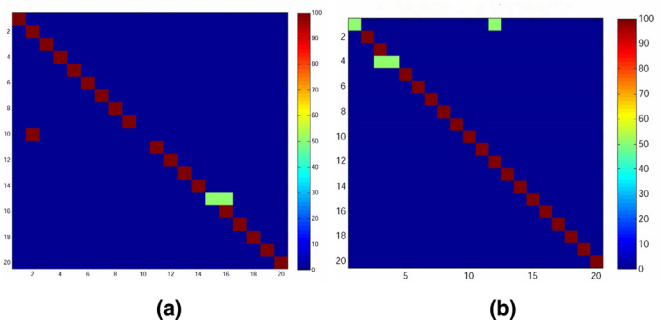


Fig. 9. (a) Confusion matrix showing a re-identification accuracy of 92.5% among the 20 humans in the real world scenario. (b) and 95% among the 20 custom avatars simulated corresponding to the instances in the real human data set.

6.2. Categorization from human queries

Refering to system architecture in Fig. 2, human query based categorization is related to ***Scenario#2***. Here, the input to the system is a human query specifying the biometric features of the probe, rather than an image query. With this query system, our system will not produce a unique *human#ID* as if working with a Re-ID ***Scenario#1***. Instead, the output will be a set of people belonging to that particular category according to the probe description.

6.2.1. Regression analysis

In order to facilitate this retrieval purely based on biometric query, we carried out linear regression analysis between the Shape Context descriptor and biometic features as explained in Section 5.2.2. For the diagnosis of the quality of regression modelling, we use the R^2 statistics. In all of our regression analysis, we could observe the value of R^2 around 0.9, implying that the regression function approximates well the true values.

As explained earlier, the input is a human query conveying some qualitative information regarding the biometric features of the person. The regression coefficients obtained from the *generic* avatar regression model is applied to the *SC* descriptors of the real human silhouettes in the gallery database and corresponding biometrics are estimated and stored in a gallery database of soft biometrics. Our

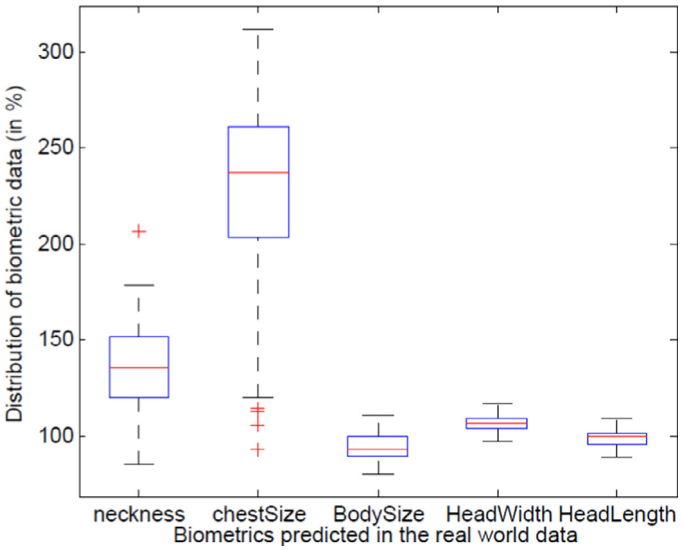


Fig. 10. Biometric data distribution predicted for the real human population, using the regression model learned using simulated avatars.

Fig. 11. A sample real world data set for the retrieval test based on human queries on biometric info.

Fig. 12. Human categorization based on biometric query: The results for large chest (*L*) query and short chest (*S*) query are presented in first and second row, respectively. The retrieved ranked list of human *#ID*s along with the predicted biometric data value are shown.

system collects this gallery data set of soft biometrics and analyse the distribution of the estimated biometric data in the training population. The most common semantic categories such as *Short (S), Medium (M) and Large (L)* are interpreted in terms of data ranges in this distribution profile. When a human query is available (eyewitness makes a statement regarding the characteristics of the suspect), it is compared against the aforementioned semantic categories, and the valid category of interest retrieved.

The statistical analysis on such estimated biometric values among real human dataset is presented in Fig. 10. We could observe a range of variances along the biometrics estimated among the data set. The distribution of Neckness ranges between 90% and 200% of the trained simulator models. Larger necks above that range (like Fig. 8(c)) are unexpected in real scenario. The parameter distribution of chest size ranges between 100% and 300%, with median close to 220%. This makes sense while checking with similar avatar models in Fig. 8(d), which is a common candidate in the real world. Body size, head width and head length are centered near the 100%, and have lower variances.

It is important to have certain biometrics with large variance in the population in order to avoid the problem of subject interference and to improve the distinctiveness among people. They acts as the most discriminative features. One interesting fact to notice is that, from the survey results in Table 1, we observed the variance in the chest width, viz., bideltoid breadth (2.59) is larger compared to the others. In our sample real world population in Fig. 10, we could also observe that chest size shows large variance and happens to be very good discriminative feature. At the same time, head length and width do not show the same level of variances. A very similar analysis was reported in the real human data set in Table 1, showing smaller variances for head length (0.72) and head width (0.60). These are very interesting observations highlighting the intuitive fact that, our regression model trained in virtual world, could generate similar test result statistics in the real world.

6.2.2. Person categorization

Here we conduct a sample test of person retrieval. Consider a sample real world human data set of 10 people shown in Fig. 11. We assume 3 categories among the population for each biometric viz. *Short* (***S***—less than lower quartile) ,*Medium* (***M***—lower quartile to upper quartile) and *Large* (***L***—above upper quartile). For example, in search of a person with large chest size, we try to retrieve the people whose *chestsize* $\geq$ 260%, which is more than the upper quartile of the distribution. Similarly, for the short chest, we can identify the people category *chestsize* $\leq$ 210%, which is less than the lower quartile in the data distribution profile. The result thus retrieves a ranked list of people trained in those respective category along with their *#ID*s. Retrieval based on chest query is depicted in Fig. 12. According to the results,

Table 3
Results of person retrieval based on biometric feature vectors estimated by regression. GT refers to the ground truth biometrics defined by manual inspection, and retrieval rate is the rate with which our retrieved category agrees with that of ground truth.

Person index	Neckness(N)		Chestsize (C)		Bodysize (B)		Headwidth (HW)		Headlength (HL)	
(*#ID*)	GT	Retrieval rate	GT	Retrieval rate	GT	Retrieval rate	GT	Retrieval rate	GT	Retrieval rate
#1	M	0.25	M	1	M	1	M	0.75	M	0.25
#2	L	1	S	0.5	S	0.5	S	0	L	1
#3	L	0	S	1	M	0.5	M	0.25	M	0.25
#4	M	1	M	1	M	0.5	M	1	M	0.75
#5	L	0	S	0.25	S	1	S	1	M	0.25
#6	S	0.5	L	1	L	0	L	0	S	0.75
#7	L	1	L	1	M	0.75	M	0.5	M	0.25
#8	S	0	L	0.25	L	0	L	0	S	0
#9	S	1	M	1	L	1	M	0	M	0.75
#10	M	1	M	0.75	M	0.5	S	0	L	0
Average retrieval accuracy	**57.5%**		**77.5%**		**47.5%**		**35%**		**42.5%**	

human index #6, #7 and #8 were classified with *large* chest, and #2, #3 and #5 found to have a *short* chest. The retrieval based on other parameters is analogous. Albeit, we considered the aforementioned three classes (*S,M and L*) as default in our case study, it could be reduced to two classes or increased to 4 or more classes (like *XXS, XS, XL, XXL*). The selection of the number of classes and range for each class are the choice of the operator. According to the requirement, he can either split or merge classes. However, with the increase in the number of classes, the retrieval performance decreases due to increase of noise influence in class assignment and inter-class ambiguity.

Among 10 people sample test set each with 4 samples, our retrieval rate for each biometric feature is given in Table 3. Since there is no availability of the ground truth for the performance evaluation in the real human data set, we rely on visual inspection of the probe images and define our *ground truth(GT)*. The rate of correct category retrieval obtained for each person with respect to the ground truth is denoted in retrieval rate. Average retrieval accuracy is found to be the highest for chest size (77.5%), thus proving to be the best discriminative features among the biometrics.

7. Conclusion and future works

In this paper, we presented a novel proposal towards identifying people in a video surveillance system either through the multimedia data acquired via video cameras or solely by means of manual queries describing natural human compliant labels known as soft biometric traits. We introduced a novel feature descriptor, Shape Context descriptor extracted on the head-to-torso region on frontal human silhouettes and verified its practicality in both real and virtual reality data sets. A slightly higher level of re-identification performance was reported in our experiment in virtual environment compared to their counterpart experiment in real world. We assign this fact to the lack or silhouette imperfections in the virtual reality scenario, thus better image segmentation methods could impact positively in real world scenario.

Another innovative contribution of this paper was the exploitation of linear relationship between Shape Context descriptors and soft biometrics. Such a regression phase filled the gap between the manual and machine interpretation of human profile and equipped the system to retrieve the person merely by soft biometric description of the subject. In order to learn this mapping, we simulated *generic* virtual avatars rendered by *Unity3D®* graphics engine, thus bypassing the requirement for laborious manual annotation of data sets. We substantiated the performance of our system by carrying out person retrieval in a sample real surveillance database.

Our automatic dual mode system is found quite appropriate in the search of an incident happened in a video surveillance, where the security personnel could opt collecting either multimedia info from the camera or eyewitness description of the suspect, which are the common ways of person identification. In future work, we plan to extrapolate the feature extraction over full body and to exploit a large set of soft biometrics. Also, we will combine other modalities (e.g., color, texture, face, gait) along with soft biometric features using multimodal fusion techniques.

Acknowledgements

This work was supported by the FCT projects [UID/EEA/50009/2013], AHA CMUP-ERI/HCI/0046/2013, doctoral grant [SFRH/BD/97258/2013] and by European Commission project POETICON++ (FP7-ICT-288382).

References

[1] A. Agarwal, B. Triggs, 3D human pose from silhouettes by relevance vector regression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, 2004, pp. 882–888.
[2] K. Aziz, D. Merad, B. Fertil, People re-identification across multiple non-overlapping cameras system by appearance classification and silhouette part segmentation, in: Proceedings of the 8th IEEE International Conference AVSS, 2011.
[3] B.I. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, V. Murino, Re-identification with RGB-d sensors, in: Proceedings of the First International Workshop on Re-Identification, 2012.
[4] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 509–522.
[5] L. Chunxiao, G. Shaogang, C. Change Loy, X. Lin, Person re-identification: what features are important? in: Proceedings of the Computer Vision ECCV 2012., Florence, Italy, 2012, pp. 391–401.
[6] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, 2005, pp. 886–893.
[7] A. Dantcheva, C. Velardo, A. D'angelo, Dugelay, Jean-Luc, Bag of soft biometrics for person identification: new trends and challenges, Multimedia Tools Appl. 51 (2011) 739–777.
[8] M. Goffredo, I. Bouchrika, J. Carter, M. Nixon, Performance analysis for automated gait extraction and recognition in multi-camera surveillance, Multimedia Tools Appl. 50 (2010) 75–94.
[9] C. Gordon, T. Churchill, C. Clauser, 1988 Anthropometric Survey of U.S Army Personnel: Methods and Summary Statistics, United States Army Natick Research Tecnical Report, 1989.
[10] H.W. Kuhn, The Hungarian method for the assignment problem, Naval Res. Logistics Quart. 2 (1955) 83–97.
[11] I. Kviatkovsky, A. Adam, E. Rivlin, Color invariants for person reidentification, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1622–1634.
[12] B. Leibe, E. Seemann, B. Schiele, Pedestrian detection in crowded scenes, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2005.
[13] D. Lowe, Distinctive image features from scale invariant features, Int. J. Comput. Vis. 60 (2004) 91–110.
[14] A. Nambiar, M. Taiana, D. Figueira, J. Nascimento, A. Bernardino, A multi-camera video data set for research on high-definition surveillance, Int. J. Mach. Intell. Sens. Signal Proc. 1 (2014).
[15] A.M. Nambiar, P.L. Correia, L.D. Soares, Frontal gait recognition combining 2D and 3D data, in: Proceedings of the on Multimedia and Security, 2012, pp. 145–150.
[16] J.C. Nascimento, J. Silva, Manifold learning for object tracking with multiple motion dynamics, in: Proceedings of the European Conference on Computer Vision, 2010.
[17] D. Reid, M. Nixon, Using comparative human descriptions for soft biometrics, in: Proceedings of the First International Joint Conference on Biometrics, 2011, pp. 1–6.
[18] A. Roy, S. Sural, J. Mukherjee, Hierarchical method combining gait and phase of motion with spatiotemporal model for person re-identification, Pattern Recognit. Lett. 33 (2012) 1891–1901.
[19] S. Samangooei, M. Nixon, B. Guo, The use of semantic human description as a soft biometric, in: Proceedings of the BTAS, 2008, pp. 1–7.
[20] D. Truong Cong, C. Achard, L. Khoudour, People re-identification by classification of silhouettes based on sparse representation, in: Proceedings of the International Conference on Image Processing Theory Tools and Applications (IPTA), 2010.
[21] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, PeterTu, Shape and appearance context modeling, in: Proceedings of the IEEE 11th ICCV, 2007, pp. 1–8.
[22] Y. Zhang, X. Wu, Q. Ruan, Combining procrustes shape analysis and shape context descriptor for silhouette-based gait recognition, Electron. Lett. 45 (2009) 674–675.