Invariant Classification of Gait Types

Fihl, Preben; Moeslund, Thomas B.

Published in:
Canadian Conference on Computer and Robot Vision, 2008. CRV '08

DOI (link to publication from Publisher):
10.1109/CRV.2008.24

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Fihl, P., & Moeslund, T. B. (2008). Invariant Classification of Gait Types. In *Canadian Conference on Computer and Robot Vision, 2008. CRV '08* (pp. 179-185). IEEE. https://doi.org/10.1109/CRV.2008.24

Invariant Classification of Gait Types

Preben Fihl and Thomas B. Moeslund
Laboratory of Computer Vision and Media Technology
Aalborg University, Denmark
{pfa,tbm}@cvmt.dk

Abstract

This paper presents a method of classifying human gait in an invariant manner based on silhouette comparison. A database of artificially generated silhouettes is created representing the three main types of gait, i.e. walking, jogging, and running. Silhouettes generated from different camera angles are included in the database to make the method invariant to camera viewpoint and to changing directions of movement. The extraction of silhouettes are done using the Codebook method and silhouettes are represented in a scale- and translation-invariant manner by using shape contexts and tangent orientations. Input silhouettes are matched to the database using the Hungarian method. A classifier is defined based on the dissimilarity between the input silhouettes and the gait actions of the database. The overall recognition rate is 88.2% on a large and diverse test set. The recognition rate is better than that achieved by other approaches applied to similar data.

1. Introduction

The human gait[1] has received much attention from the computer vision research community because of the large amounts of information that can be extracted from a person's gait. The human gait has successfully been used to detect the presence of people, e.g. in surveillance video [13, 19], based on its rather distinct cyclic nature. The ability to extract human gait from a distance has also motivated the use of gait as a non-intrusive biometric [9, 20, 18]. Finally, there has been considerable interest in the computer vision community in the classification of gait types or, more generally, of different types of human action [2, 4, 15]. In this paper we consider gait in the context of action recognition rather than the use of gait in personal identification.

The automatic classification of human actions has many applications in advanced user interfaces, annotation of video data, intelligent vehicles, and surveillance. The vast and increasing use of video surveillance and the fact that one of the main human activities that surveillance cameras observe is that of gait, i.e. walking, jogging, or running, is a strong motivation for gait type recognition. A natural part of automatic surveillance systems will be the ability to distinguish between different gait types. Surveillance cameras are often applied in unconstrained environments, so a gait type classification method is needed that is *invariant* to camera frame rate and calibration, viewpoint, moving speed, scale change, and non-linear paths of motion. This paper presents such a method.

Other papers have presented systems invariant to one or more of these factors within the area of classification of gait types, but so far none has considered all these factors simultaneously. [10] presents good results in classifying different types of human motion, but the system is limited to motion parallel to the image plane. [14] describes a method for understanding behavior by combining different human actions. This method can deal with fairly unconstrained scenes but classifies the action performed by speed of locomotion, which is not a reliable way to distinguish between the very similar actions of walking, jogging, and running. Furthermore, estimation of speed would require contextual knowledge that is not always accessible in surveillance video. [2] uses space-time shapes to recognize actions independently of speed. The method is robust to different viewpoints but cannot cope with non-linear paths created by changes in direction of movement. Other state-of-the-art approaches are mentioned in section 7 along with a comparison of results.

1.1. Our Approach

A current trend in the area of human motion analysis based on computer vision is to acquire large amounts of data that express the needed variability and then use this data to train the system. This training process often means spending a lot of time on extracting and annotating the data and subsequently aligning the different training sequences tem-

[1] By gait is meant bipedal locomotion: walking, jogging and running.

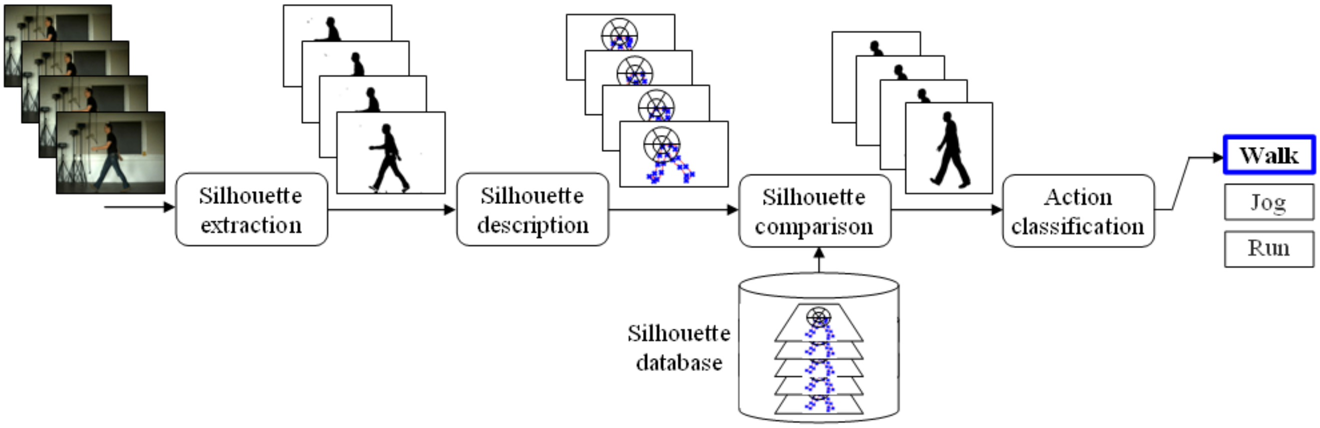

Figure 1. An overview of the approach. See text for details.

porally. Our approach circumvents these problems by use of computer generated training data. A computer graphics model of a human is animated to perform the required actions. Classification of gait types does not require the extraction of the exact pose of a person and silhouettes are therefore sufficient as inputs. Gait analysis methods based on silhouettes have succeeded in extracting detailed information about gait from silhouettes [3, 9] and the use of silhouettes makes considerations regarding the appearance of people's clothes superfluous. A completely realistic-looking computer graphics model is not needed to generate silhouettes as long as the shape is correct and the 3D-rendering software Poser [17], which has a built-in Walk Designer, can be used to animate human gaits. This approach gives us the advantages of very fast training plus the ability to easily generate training data from new viewpoints by changing the camera angle.

Our computer-graphics-based approach offers two main contributions. Firstly, the methods applied are chosen and developed to allow for classification in an unconstrained environment. This results in a system that is invariant to more factors than other approaches, i.e. invariant in regard to camera frame rate and calibration, viewpoint, moving speeds, scale change, and non-linear paths of motion. Secondly, the use of the computer graphics model decouples the training set completely from the test set. Usually methods are tested on data similar to the training set, whereas we train on computer-generated images and test on video data from several different data sets. This is a more challenging task but it makes the system more independent of the type of input data and therefore increases the applicability of the system.

The framework described in this paper is shown in figure 1. A database of reference poses is first created (section 2). We extract the human silhouette from input video (section 3) and represent it efficiently (section 4). We then compare the silhouette with computer graphics silhouettes from the database (section 5). The results of the comparison are calculated for an entire sequence and the action in that sequence is then classified (section 6).

2. Silhouette database

We create a database of human silhouettes performing one cycle of each of the main gait types: walking, jogging, and running. To make our method robust to changes in viewpoint we generate database silhouettes from three different camera angles. With 3D-rendering software this is an easy and very rapid process that does not require us to capture new real life data for statistical analysis. The database contains silhouettes of the human model seen from a side view and from cameras rotated 30 degrees to both sides. The three camera angles allow us to match database silhouettes with silhouettes of people moving at angles of at least ± 45 degrees with respect to the viewing direction. People moving around in open spaces will often change direction while in the camera's field of view (creating non-linear paths of motion), thus we cannot make assumptions about the direction of movement. To handle this variability each new input silhouette is matched to database silhouettes taken from all camera angles. Figure 7, row 1 shows a sequence with a non-linear motion path where the first frames will match database silhouettes from a viewpoint of -30 degrees and the last frames will match database silhouettes from a viewpoint of 30 degrees.

The silhouettes generated are represented as described in section 4. We generate T silhouettes of each gait type from the three viewpoints, i.e. $T \cdot 3 \cdot 3$ silhouettes in total. Figure 2 shows example poses of the database silhouettes.

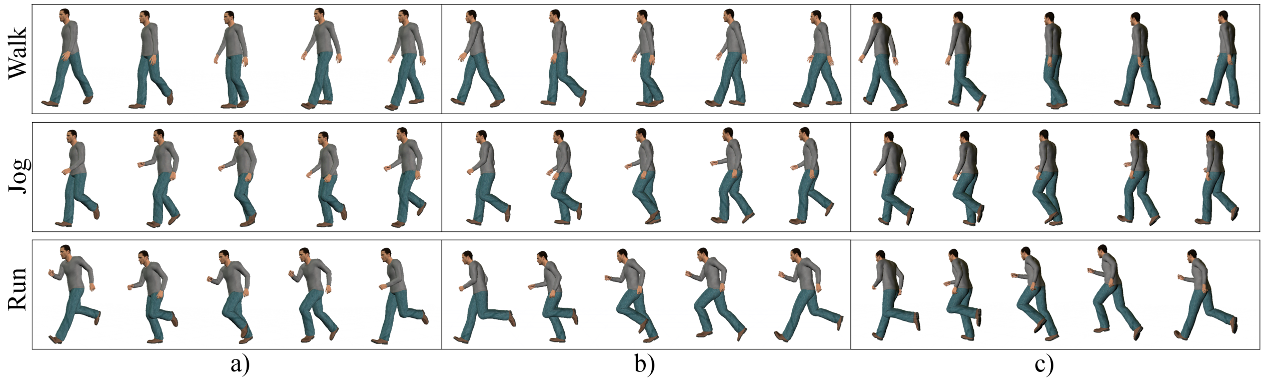

Figure 2. Example of database silhouettes generated by 3D-rendering software. Silhouettes are generated from three viewpoints. a) and c) illustrate renderings from cameras rotated 30 degrees to each side. b) illustrates renderings from a direct side view.

3. Silhouette extraction

We extract silhouettes from video sequences by foreground segmentation using the Codebook background subtraction method as described in [5] and [7]. The method has been shown to be robust in handling both foreground camouflage and shadows by separating intensity and chromaticity in the background model. Each pixel in the background model is represented as a vector **u** in the RGB-cube. The distance, in terms of chromaticity, **r**, from a new pixel, **x**, to the background model is measured as the perpendicular distance from the vector. The difference in intensity is measured along the vector and denoted, **h**, see figure 3. A codeword for a background pixel defines the maximum allowable chromaticity distortion and intensity variation at that pixel resulting in a cylindrical decision boundary D.

Each background pixel can be represented by multiple codewords making the background model multi modal. By allowing creation of new codewords at runtime the background model can furthermore handle multiple layers of background. These two properties allow the method to model moving backgrounds like tree branches and objects that become part of the background after staying stationary for a period of time. To maintain good background subtraction quality over time it is essential to update the background model and [5] describes two different update mechanisms that handle rapid and gradual changes, respectively. This robust background subtraction method allows us to use quite diverse video sequences from both indoor and outdoor scenarios as input.

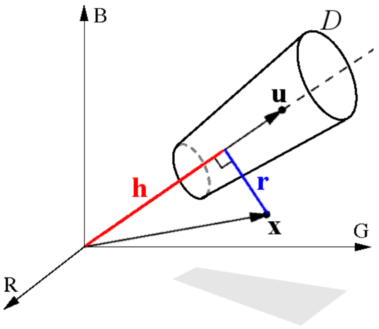

Figure 3. Illustration of the representation used in the background subtraction. See text for details.

4. Silhouette description

When a person is moving around in a typical surveillance setup his or her arms will not necessarily swing in a typical ”walking” manner; the person may be making other gestures, such as waving, or he/she might be carrying an object. To circumvent the variability and complexity of such scenarios we chose to classify the gait solely on the silhouette of the legs. Furthermore, [8] shows that identification of people on the basis of gait, using the silhouette of legs alone, works just as well as identification based on the silhouette of the whole person. The silhouette of the legs is defined as the bottom half of the whole silhouette.

To allow recognition of gait types across different scales we use shape contexts [1] to describe the leg silhouettes. n points are sampled from the contour of the leg silhouette and for each point we determine the shape context and the tangent orientation at that point. Scale invariance is achieved with shape contexts by normalizing the radial distances of the histogram by the mean distance between all point pairs on the contour. The shape context description is not sensitive to small amounts of noise in the foreground mask and it also makes the method robust to inaccuracies in the division of the foreground mask into the leg silhouette.

5. Silhouette comparison

To find the best match between an input silhouette and database silhouettes we follow the method of [1]. We calculate the cost of matching a sampled point on the input silhouette with a sampled point on a database silhouette using the χ^2 test statistics. The cost of matching the shape contexts of point p_i on one silhouette and point p_j on the other silhouette is denoted $c_{i,j}$. The normalized shape contexts at points p_i and p_j are denoted $h_i(k)$ and $h_j(k)$ respectively with k as the bin number, $k = 1, 2, ..., K$. The χ^2 test statistics is given as:

$$c_{i,j} = \frac{1}{2} \sum_{k=1}^{K} \frac{[h_i(k) - h_j(k)]^2}{h_i(k) + h_j(k)} \tag{1}$$

The normalized shape contexts gives $c_{i,j} \in [0; 1]$.

The difference in tangent orientation $\phi_{i,j}$ between points p_i and p_j is added to $c_{i,j}$ ($\phi_{i,j} \in [0; \pi]$). We choose not to normalize $\phi_{i,j}$ before the addition. Experiments have shown that $\phi_{i,j}$ effectively discriminate dissimilar points whereas $c_{i,j}$ express more detailed differences which should have a high impact only when tangent orientations are alike.

The final cost $C_{i,j}$ of matching the two points are:

$$C_{i,j} = c_{i,j} + \phi_{i,j} \tag{2}$$

The costs of matching all point pairs between the two silhouettes are expressed as the cost matrix C, see figure 4. The Hungarian method [11] is used to solve the square assignment problem of identifying which one-to-one mapping between the two point sets that minimizes the total cost. The costs involved in this one-to-one mapping are added up and express the dissimilarity or distance between the two silhouettes. An input silhouette is matched to all silhouettes of the database and we can now identify the best match in the whole database by taking the database silhouette with the shortest distance to the input. While the silhouette comparison is done on a frame-by-frame basis the action classification incorporates the temporal information from the whole input sequence.

6. Action classification

To get a robust classification of actions we combine three different types of information, as described below. We first calculate an *action error* E for each action and then two associated weights: *action likelihood* α and *temporal consistency* β.

6.1. Action Error

The output of the silhouette comparison is a set of distances between the input silhouette and each of the database silhouettes. These distances express the difference or error between two silhouettes. Figure 5 illustrates the output of the silhouette comparison. The database silhouettes are divided into three groups corresponding to walking, jogging, and running, respectively. We accumulate the errors in the best matches within each group of database silhouettes to find the difference between the action being performed in the input video and each of the three actions in the database, see figure 6. These accumulated errors constitute the *action error* E.

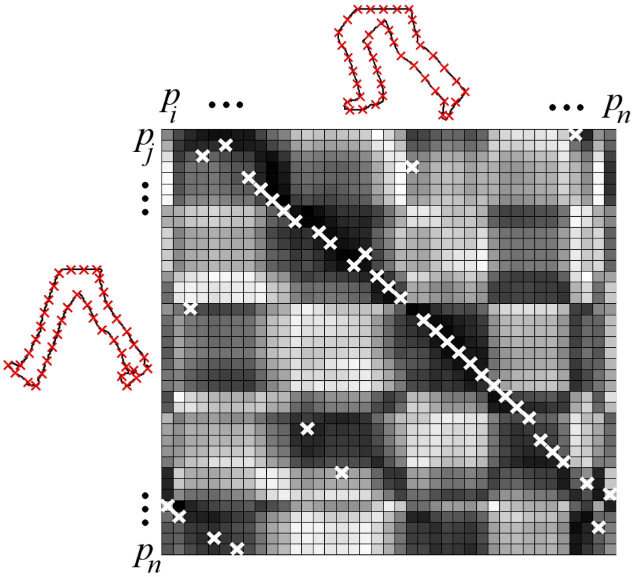

Figure 4. Illustration of the cost matrix of matching each point pair of the two silhouettes (dark colors correspond to low costs). White crosses mark the best one-to-one matching between the two silhouettes.

6.2. Action Likelihood

When silhouettes of people are extracted in difficult scenarios and at low resolutions the silhouettes can be noisy. This may result in large errors between the input silhouette and a database silhouette, even though the actual pose of the person is very similar to that of the database silhouette. To handle such inaccuracies we weight the action error by the likelihood of that action. Since we use the minimum action error the actual weight applied is one minus the action likelihood:

$$\alpha_a = 1 - \frac{n_a}{N} \tag{3}$$

where n_a is the number of input silhouettes in a sequence with the best overall match to a silhouette from action a, and N is the total number of input silhouettes in that video

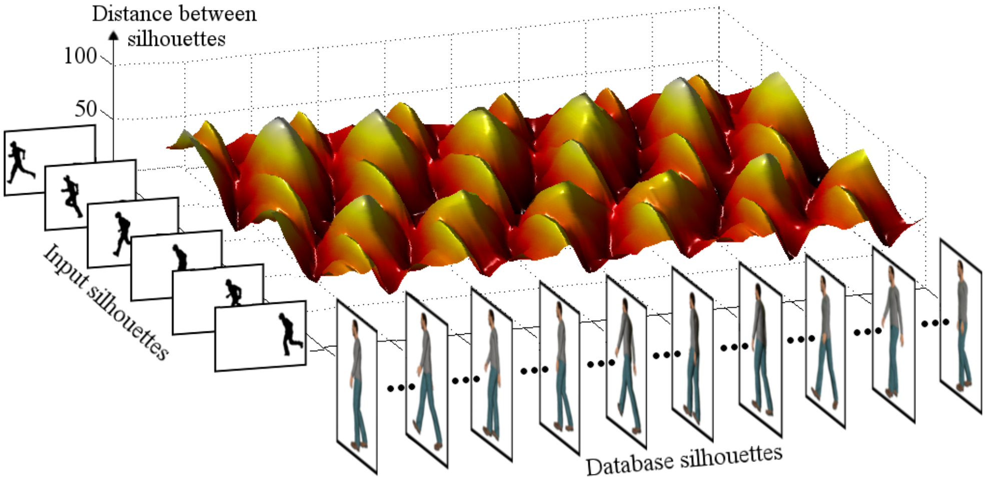

Figure 5. Illustration of the silhouette comparison output. The distances between each input silhouette and the database silhouettes of each gait type are found (shown for walking only). 90 database silhouettes are used per gait type, i.e. $T = 30$.

sequence. This weight will penalize actions that have only a few overall best matches, but ones with small errors, and will benefit actions that have many overall best matches, e.g. the running action in figure 6.

6.3. Temporal Consistency

When considering only the overall best matches we can find sub-sequences of the input video where all the best matches are of the same action *and* in the right order with respect to a gait cycle. This is illustrated in figure 6 where the running action has great temporal consistency (silhouette numbers 14-19). The database silhouettes are ordered in accordance with a gait cycle. Hence, the straight line between the overall best matches for input silhouettes 14 to 19 shows that each new input silhouette matches the database silhouette that corresponds to the next body configuration of the running gait cycle.

Sub-sequences with correct temporal ordering of the overall best matches increase our confidence that the action identified is the true action. The temporal consistency describes the length of these sub-sequences. Again, since we use the minimum action error we apply one minus the temporal consistency as the weight β_a:

$$\beta_a = 1 - \frac{m_a}{N} \tag{4}$$

where m_a is the number of input silhouettes in a sequence in which the best overall match has correct temporal ordering within action a, and N is the total number of input silhouettes in that video sequence.

Our definition of temporal consistency is rather strict when you consider the great variation in input silhouettes caused by the unconstrained nature of the input. A strict definition of temporal consistency allows us to weight it more highly than action likelihood, i.e. we apply a scaling factor w to β to increase the importance of temporal consistency in relation to action likelihood:

$$\beta_a = 1 - w \cdot \frac{m_a}{N} \tag{5}$$

Applying both action likelihood and temporal consistency as weights on the action error yields the final classifier:

$$Action = \arg\min_a (E_a \cdot \alpha_a \cdot w\beta_a) \tag{6}$$

7. Results

This method of classification has been tested on a large and diverse data set. We have compiled 136 video sequences from 4 different data sets. The data sets include indoor and outdoor video sequences, different directions of movement with respect to the camera (±45 degrees from the viewing direction), non-linear paths of motion, different camera elevations and tilt angles, different video resolutions, and varying silhouette heights (from 41 pixels to 454 pixels). Figure 7 shows examples from the input videos. For the silhouette description the number of sampled points was 100 and the number of bins in the shape contexts was 60. Thirty silhouettes were used for each action, i.e. $T = 30$. The temporal consistency was weighted by a factor of four, i.e. $w = 4$, determined through quantitative experiments. We achieve an overall recognition rate of 88.2%.

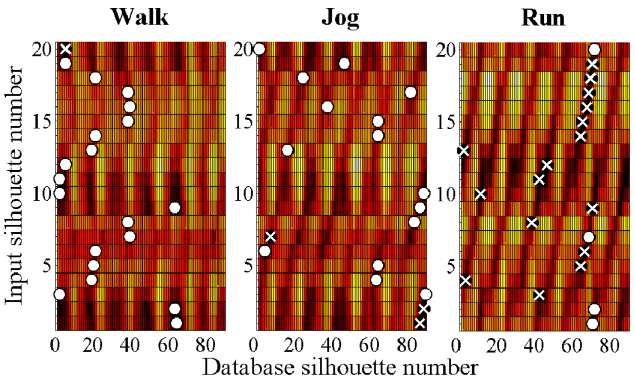

Figure 6. The output of the silhouette comparison of figure 5 is shown in 2D for all gait types (dark colors illustrate small errors and bright colors illustrate large errors). For each input silhouette the best match among silhouettes of the same action is marked with a white dot and the best overall match is marked with a white cross. The shown example should be interpreted as follows: the silhouette in the first input frame is closest to walking silhouette number 64, to jogging silhouette number 86, and to running silhouette number 70. These distances are used when calculating the action error. When all database silhouettes are considered together, the first input silhouette is closest to jogging silhouette number 86. This is used in the calculation of the two weights.

Table 1 shows the classification results in a confusion matrix. We have analyzed our classification errors in table 1 and found no significant correlation between the classification errors and the camera viewpoint (pan and tilt), the size and quality of the silhouette extracted, the frame rate, the image resolution, and the linearity of the path. Furthermore, we also evaluated the effect of the number of frames (number of gait cycles) in the sequences and found that our method classifies gait types correctly even when there are only a few cycles in the sequence. On the basis of these findings we note that our method is indeed invariant to the main factors relevant for gait classification.

The majority of the errors in table 1 occur simply because the action of jogging resembles that of running. These errors reflect the intuitive notion that jogging and running are similar actions that both differ more from walking than from each other. [12] shows results of a classification made by humans on a subset of the KTH data set (see figure 7). The classification rates are 100% for walking, 98% for jogging, and only 81% for running, which illustrates the inherent difficulty in distinguishing the two actions.

	Walk	Jog	Run
Walk	96.2	3.8	0.0
Jog	0.0	68.3	31.7
Run	0.0	2.3	97.7

Table 1. The percentage of correct matches listed in a confusion matrix. The rows represent the true classes and the columns show the classification.

The matching percentages in table 1 cannot directly be compared to the results of other methods of classification since we have included samples from different data sets to obtain more diversity. However, 75 of the sequences originate from the KTH data set. In table 2 we therefore list the percentage of correct matches for the methods we know of that report the best results on the KTH data set. We acknowledge that the KTH data set contains three additional actions (boxing, hand waving, and hand clapping). However, in the results reported in the literature the gait actions are not generally confused with the three hand actions, which means that our results are comparable.

Methods	Classification results in %			
	Total	Walk	Jog	Run
Our method	92.0	100.0	82.1	94.4
[12] 2007	84.3	98	79	76
[6] 2006	77.8	93.3	80.0	73.3
[4] 2005	77.3	90	57	85
[15] 2004	75.0	83.8	60.4	54.9

Table 2. Best classification results reported in the literature on the KTH data set. The listed results of our method are based on the 75 KTH sequences included in our test set.

8. Conclusion

In this paper we have presented a method for classifying three types of gait: running, jogging, and walking. The method is *not* based on statistical analysis of training data but rather on a general gait motion model synthesized using a computer graphics human model. This makes training (from different views) very easy and test results less biased. The method has been tested on different data sets containing all the important factors that such a method should be able to handle. The method performs well (both in its own right and in comparison to related methods) and we therefore conclude that it can be characterized as an *invariant* method of gait classification.

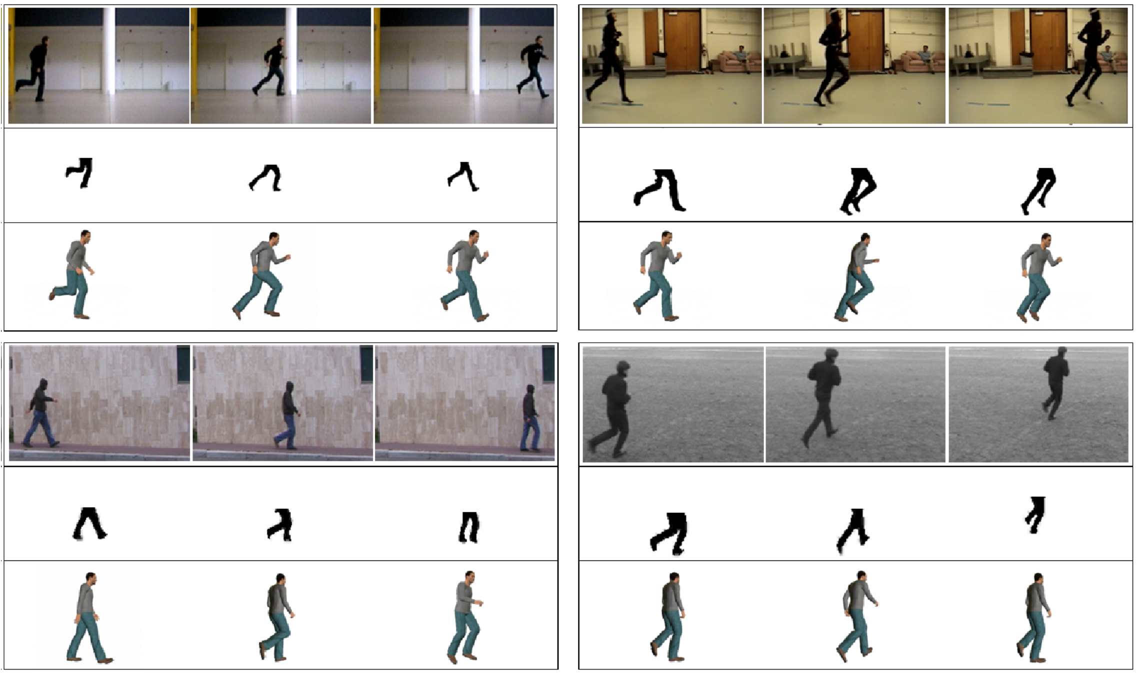

Figure 7. Samples from the 4 different data sets used in the test together with the extracted silhouettes of the legs used in the database comparison, and the best matching silhouette from the database. Top left: data from our own data set. Bottom left: data from the Weizmann data set [2]. Top right: data from the CMU data set [16]. Bottom right: data from the KTH data set [15].

Acknowledgements

This work is supported by the EU project HERMES (FP6 IST-027110).

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape Matching and Object Recognition Using Shape Contexts. *PAMI*, 24(4), 2002.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes. In *ICCV*, 2005.

[3] R. Collins, R. Gross, and J. Shi. Silhouette-Based Human Identification from Body Shape and Gait. In *FGR*, 2002.

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior Recognition via Sparse Spatio-Temporal Features. In *VS-PETS*, 2005.

[5] P. Fihl, R. Corlin, S. Park, T. Moeslund, and M. Trivedi. Tracking of Individuals in Very Long Video Sequences. In *Int. Symposium on Visual Computing*, Lake Tahoe, Nevada, USA, November 6-8 2006.

[6] H. Jiang, M. Drew, and Z. Li. Successive Convex Matching for Action Detection. In *CVPR*, 2006.

[7] K. Kim, T. Chalidabhongse, D. Harwood, and L. Davis. Real-time Foreground-Background Segmentation using Codebook Model. *Real-time Imaging*, 11(3), 2005.

[8] Z. Liu, L. Malave, A. Osuntugun, P. Sudhakar, and S. Sarkar. Towards Understanding the Limits of Gait Recognition. In *Int. Symposium on Defense and Security*, Orlando, Florida, USA, April 12-16 2004.

[9] Z. Liu and S. Sarkar. Improved Gait Recognition by Gait Dynamics Normalization. *PAMI*, 28(6), 2006.

[10] O. Masoud and N. Papanikolopoulos. A Method for Human Action Recognition. *Image and Vision Computing*, 21(8), 2003.

[11] C. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Courier Dover Publications, Mineola, NY, USA, 1998.

[12] A. Patron and I. Reid. A Probabilistic Framework for Recognizing Similar Actions using Spatio-Temporal Features. In *British Machine Vision Conference*, 2007.

[13] Y. Ran, I. Weiss, Q. Zheng, and L. Davis. Pedestrian Detection via Periodic Motion Analysis. *IJCV*, 71(2), 2007.

[14] N. Robertson and I. Reid. Behaviour Understanding in Video: A Combined Method. In *ICCV*, 2005.

[15] C. Schüldt, I. Laptev, and B. Caputo. Recognizing Human Actions: a Local SVM Approach. In *ICPR*, 2004.

[16] *http://mocap.cs.cmu.edu/*. CMU Graphics Lab Motion Capture Database, March 2007.

[17] *http://www.e-frontier.com/go/poser/*. Poser (ver. 6.0.3.140), February 2007.

[18] A. Veeraraghavan, A. Roy-Chowdhury, and R. Chellappa. Matching Shape Sequences in Video with Applications in Human Movement Analysis. *PAMI*, 27(12), 2005.

[19] P. Viola, M. Jones, and D. Snow. Detecting Pedestrians Using Patterns of Motion and Appearance. *IJCV*, 63(2), 2005.

[20] C. Yam, M. Nixon, and J. Carter. On the Relationship of Human Walking and Running: Automatic Person Identification by Gait. In *ICPR*, 2002.