Proceedings of the 14th International Conference
on Automation & Computing, Brunel University,
West London, UK, 6 September 2008

Low-resolution Human Detection and Gait Recognition in Natural Scenes

Andrzej Ruta, Yongmin Li, Xiaohui Liu
School of Information Systems, Computing and Mathematics
Brunel University
Uxbridge, Middlesex, UB8 3PH, United Kingdom
{Andrzej.Ruta, Yongmin.Li, Xiaohui.Liu}@brunel.ac.uk

***Abstract*—Fast and stable detection of humans in natural scenes is a challenging task due to the varying appearance of the target and diverse background. Under these circumstances a higher-level analysis, e.g. action classification, becomes even more difficult, as it is dependent on the quality of the preceding steps in the processing pipeline. In this paper we address this issue on both levels: low-level detection, and high-level classification. Detecting human figures is formulated as a problem of finding maxima of the distribution generated in the Haar cascade's response space. To find these maxima, we employ an augmented Mean Shift algorithm which assigns each hypothesis a confidence measure related to the distance of the classifier's response from the decision boundary. This confidence is incorporated in the Mean Shift formula to direct the search towards the more reliable hypotheses, which ensures a more accurate detection. On top of the detection framework we have developed a Hidden Markov Model action classifier. It is based on a discrete set of key poses obtained via clustering and utilises a simple motion feature representation. Our approach is exemplified by gait recognition. The model has been trained to recognise four different types, separately in the left and right direction, using the video sequences obtained with a stationary DV camcorder. Good performance on the unseen sequences has been observed which validates our approach and suggests it could be adopted for more general action recognition.**

Keywords-human detection; gait recognition; mean shift; motion feature representation

I. INTRODUCTION

Human detection and action recognition attracted much attention in the last decade, which can partly be explained by the redefined security requirements that emphasise the role of visual surveillance. Nevertheless, the gap between the state-of-the-art recognition systems and the capabilities of human eyes is still apparent. Observing humans in natural environment, we can capture the targets and recognise the activities instantly, regardless of the resolution and the appearance of the subjects and the background. The same ability is extremely hard to replicate on a computer machine and the risk of misinterpretation grows with the increasing complexity of the scene. The main obstacles include different clothing of humans, variability of the background, and a high computational cost of the processing.

In this work we attempt to address the above problems and present an integrated method for detecting humans and recognising their actions. Due to the characteristics of the data we possess, we put our attention to a specific subproblem: gait recognition. However, adaptation to other kinds of actions is straightforward. Our approach assumes a stationary camera and is primarily intended for the "medium field" targets, i.e. the humans seen at a considerable distance, where the motion information appears to be more reliable than the appearance features. Our contribution can be summarised in two points. In the detection stage we utilise Haar wavelet shape descriptors and a large, low-resolution image database to train a binary human classifier. Motivated by the observation that the classifier's responses form dense clusters around the true human figures, we propose to treat the response space as a probability distribution with the maxima to be found. To achieve this goal, we employ a modified Mean Shift algorithm which incorporates confidence of each individual human location hypothesis in weights that bias the direction of search towards the more reliably estimated locations. Secondly, we recognise the gait of detected humans using a Hidden Markov Model classifier trained on the real-life video sequences. The novelty of the approach lies in the underlying representation of the images which is constructed based on the temporal difference maps of the original video frames and the appropriate normalisation, thresholding and blurring. Performance of the system is evaluated on the unseen sequences depicting four types of gait, each distinguished in two versions, depending on the direction of motion.

The rest of this paper is organised as follows. In Section II the related work is reviewed. Section III discusses the basic human detection algorithm and the measures taken to improve its performance. In Section IV we develop a spatio-temporal representation of the image sequences and discuss how it is utilised to train a Hidden Markov Model action classifier. In Section V our experiments on gait recognition are summarised. Finally, conclusions are drawn in Section VI.

II. RELATED WORK

A number of different approaches have been taken so far to detect humans in static images and video sequences. In majority of recent studies human detector is trained from a large database of example images by selecting a compact set of discriminative local features from an overcomplete feature space. Two most popular feature representations used are Haar wavelet features [3,7,10] and Histograms of Oriented Gradients (HOG) [16,18]. In the same time the best performing classifiers reported are combinations of linear classifiers obtained by AdaBoost, [10,21] or Support Vector Machines (SVM) [3,16]. A cascade organisation of the classifiers is frequently adopted to make the classifier both: more accurate and more efficient in quickly rejecting the most obviously irrelevant regions of the image. Several recently published algorithms differ significantly from this mainstream. In [7,21] separate specialised classifiers are trained for individual body parts and then combined to yield a final, more accurate decision. In [10] Viola, Jones and Snow extend the notion of Haar wavelets to the pairs of consecutive images, which helps capture the most discriminative features of the pedestrians as well as the short-term characteristics of their motion. Impressive detection results are reported by Tuzel, Porikli and Meer [22] who propose a novel algorithm to detect humans in static images utilizing covariance matrices as object descriptors. Leibe, Seemann and Schiele [17] have chosen a completely different strategy. They employ a generative model of a human based on a probabilistic assignment of the discriminative local appearances from a learned codebook to the observed image patches. Their results show a remarkable detection capability in crowded street scenes and for severe overlaps, but their approach seems not to be adequate for the low-resolution images and noisy video in particular. Many other studies on human detection introduce extra assumptions on the localisation, appearance or motion of the target, e.g. side-viewed humans, motion periodicity [5,12]. Finally, it is not rare to use dedicated hardware to facilitate human detection, e.g. stereo camera setup [4,6], or infrared cameras [13,23].

One of the standard methods for action recognition used in the past is via description of the human motion using Hidden Markov Models (HMM) [1]. For example, Meyer, Pösl and Niemann [2] generate motion features from the trajectories of body parts which are statistically modelled along with the static background. Another well-established technique utilises so-called temporal templates. In [8] the view-specific templates of distinctive human motions are built on the basis of Motion Energy Images (MEI) representing where the motion has occurred in an image sequence, and Motion History Images (MHI) containing the information about the recency of motion at the corresponding spatial locations. Promising results are reported for various actions performed indoors by the isolated humans. An extension of the Motion History Images to 3D, Motion History Volumes (MHV), is proposed in [19]. Based on the experimental results it is concluded that the discriminative 3D motion descriptors based on MHV can be used for efficient recognition of essentially all basic human actions, independent of the actual actor and the viewpoint. In the study of Efros, Berg, Mori and Malik [11] a robust motion descriptor for each frame of the input video is computed from the appropriately split, blurred, and normalised optical flow signal. The results obtained in the experiments involving various datasets reveal good classification performance which shows validity of the proposed algorithms. However, this success is achieved at the cost of making a very strong assumption: that the image sequence can be perfectly stabilised, i.e. that the moving figure is kept ideally in the centre of focus. In the recent study of Gorelick et al. [24] another strategy is taken. Action recognition is converted into a 3D shape classification problem, where the 3D motion shapes are constructed via tiling the 2D images obtained for human silhouettes using Poisson equation solver. This method is reported to offer an impressive recognition performance but seems to depend strongly on the quality of the input binary silhouettes.

III. Candidate Detection

Most of the automatic human detection algorithms aim at capturing humans that are sufficiently large to clearly distinguish the body parts and generate robust appearance features, e.g. SIFT descriptors [14], or gradient orientation histograms [16]. These features are further used to support detection. However, our eyes can rapidly distinguish between the humans and the other objects or the background even when they are seen at a large distance. In an attempt to replicate this impressive performance on a computer machine, we adopt the approach of Viola and Jones involving a trainable classifier cascade based on the Haar wavelet features [15]. It proves to be an adequate method for detecting patterns in a low resolution imagery and outperforms other techniques with regard to the computational complexity.

To train the classifier, we use the Daimler-Chrysler dataset [20] comprised of as small as 18x36 pixel images of pedestrians and non-pedestrians. For description of humans the rectangular filters shown in Fig. 1 are generated and the AdaBoost procedure is employed to select the most discriminative features. In order to maintain a manageable feature space, only the features satisfying w, h={4px, 8px} shifted by 1/4 the wavelet size in each direction are considered, giving 2622 features in total. To make the detector scale-invariant, we employ a pyramidal image representation. As a result, in runtime we capture humans at different distances from the camera in multiple subsampled copies of the original video frames, while the operational scale of the detector is fixed to 18x36 pixels.

Our adaptation of the Viola and Jones' detector [15] differs from the original approach in that an appropriate pre- and postprocessing are employed in order to improve the detection speed and accuracy. As we ultimately focus on the moving humans, in a preprocessing step we use a filter that retains only these detection windows where the average intensity of the contained pixels changes above threshold. The realisation of such a motion filter is based on the temporal difference maps that are computed from the pairs of consecutive frames. Then, for each such difference image an integral image [15] is calculated. With help of the latter we can compute motion density in any rectangular region R_j of size (w_j, h_j) and with the top-left corner at (x_j, y_j) in constant time:

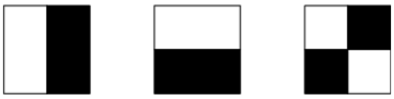

Figure 1. Haar wavelet filters used in our AdaBoost training of the human detector

$$\mu(R_j) = \frac{1}{w_j h_j}\left(\mathbf{I}_{\substack{x_j+w_j,\\ y_j+h_j}} - \mathbf{I}_{\substack{x_j,\\ y_j+h_j}} - \mathbf{I}_{\substack{x_j+w_j,\\ y_j}} + \mathbf{I}_{\substack{x_j,\\ y_j}}\right). \quad (1)$$

It should be noted that because we merely focus on thresholding, the division in (1) can be incorporated in the appropriately pre-multiplied threshold.

Motion density filter radically reduces the number of false positives but cannot cope with the clouds of redundant candidates detected around the true human figures, as shown in Fig. 2. Well-suited way of addressing this problem is by considering the detector's response space a probability distribution. In that sense the likely target objects correspond to the maxima of this distribution. In order to find these maxima, we employ a Mean Shift algorithm [9] which is modified in the following way. Each positive human location hypothesis, R_j, generated by the detector is characterised by the centroid $C_j = (x_j, y_j)$ and scale s_j. In addition, it is assigned a quality, $Q(R_j)$ which we relate to the confidence of the binary classifier's decision made for R_j. For a classifier trained in an AdaBoost manner, this confidence corresponds to the distance from the linear decision boundary:

$$Q(R_j) = \sum_{t=1}^{T} \alpha_t h_t(R_j), \quad (2)$$

where $h_t(R_j)$ denote the weak classifier responses, $\alpha_t = \log((1-e_t)/e_t)$, and e_t are the error rates of the weak classifiers. In the case of a more complex classifier cascade [15] the quality formula becomes less straightforward as the classifier's decision confidence is shared among the layers. However, we observed that simply a sum of $Q(R_j)$ terms over all layers gives similar results at an additional benefit of having a more efficient false candidate rejection mechanism.

Let $x_j = [x_j, y_j, s_j]$, $j = 1, \dots, n$ be the d=3-dimensional feature vectors characterising positive hypotheses in the response space of the human detector, and $f(x)$ be the underlying distribution of x. Our goal is to find the modes of this distribution given by the kernel density estimator:

$$\hat{f}_{h,K}(x) = \frac{c_{k,d}}{nh^d} \sum_{j=1}^{n} k\left(\left\| \frac{x - x_j}{h} \right\|^2\right), \quad (3)$$

where $k(x)$ is the profile of the symmetric kernel $K(x)$, h is the bandwidth parameter, and $c_{k,d}$ is a normalisation constant which makes $K(x)$ integrate to one. Modes $\hat{f}_{h,K}(x)$ are located among the zeros of its gradient given by the gradient density estimator:

$$\nabla \hat{f}_{h,K}(x) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{j=1}^{n} (x - x_j) k'\left(\left\| \frac{x - x_j}{h} \right\|^2\right). \quad (4)$$

Substituting $g_{x,j} = -k'\left(\left\|(x - x_j)/h\right\|^2\right)$ and introducing $g(x)$ into (4) yields:

$$\nabla \hat{f}_{h,K}(x) = \frac{2c_{k,d}}{nh^{d+2}} \left[\sum_{j=1}^{d} g_{x,j}\right] \left[\frac{\sum_{j=1}^{n} x_j g_{x,j}}{\sum_{j=1}^{n} g_{x,j}} - x\right]. \quad (5)$$

The latter term in square brackets, known as *mean shift*, is responsible for pushing the centre of the kernel window in the direction of maximum increase in the density and hence ensures convergence of the algorithm. However, it cannot take account of the possibly varying confidence of the measurements encoded in the feature vectors. By incorporating confidence terms, $q_t = Q(R_j)$ in (5), the mean shift vector becomes:

$$m_{h,G} = \left(\sum_{j=1}^{n} x_j q_j g_{x,j}\right) \Big/ \left(\sum_{j=1}^{n} q_j g_{x,j}\right) - x, \quad (6)$$

which is equivalent to amplifying the density gradients pointing towards the more reliably detected human locations. Such a modification of the original Mean Shift is useful as it increases the detector's accuracy and prevents it from getting locked in the regions containing dense positive yet weak hypotheses. In Fig. 2 several examples of mode estimation have been illustrated to show that our weighted Mean Shift algorithm converges in more accurate human locations.

Having found the modes of the detector's response distribution, the final human candidates are only considered in the locations the detector can confirm. It means that for each retrieved triple of parameters $x_k = [x_k, y_k, s_k]$ the detector is run at point (x_k, y_k) of the input image stored at the pyramid level corresponding to the discrete scale nearest to s_k. The hypothesis is retained only if the corresponding location is positively re-classified.

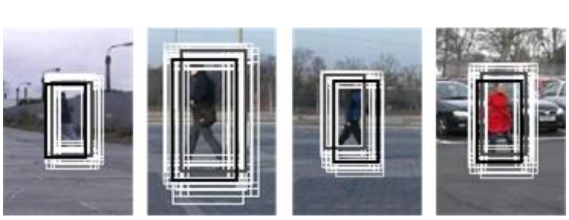

Figure 2. Typical output of the attentive Haar cascade. Clouds of hypotheses marked white are visible around the human figures. Gray boxes correspond to the modes found by the standard Mean Shift, while black boxes denote the modes found by our weighted Mean Shift algorithm.

IV. Gait Recognition

The motivation for our gait recognition approach described below is based on how easily people can recognise common human movements from the low-resolution imagery. We claim that it is possible based solely on the analysis of the local motion of body parts if the image sequence is stabilised, i.e. the detected subject is always roughly in the centre of focus. In the previous section it has been outlined how to achieve such a stabilisation on a reasonable level under a natural, possibly difficult background. Below, we will show that the sequence stabilisation facilitates construction of a robust spatio-temporal motion descriptor based on image differencing. Further, we will incorporate this descriptor into a Hidden Markov Model framework to obtain a gait classifier that can prove useful in real-life scenarios.

A. Spatio-temporal Motion Descriptor

Our motion descriptor is computed within the bounding boxes of the detected human candidates at time t, cropped from the temporal difference image of the entire scene $\Delta I = I_t - I_{t-1}$. No additional computational cost is introduced as ΔI is already used in the detection stage to capture the presence of motion (see section III for details). Implications of this choice are significant. First, this method is much faster than the flow field computation [11]. Second, the image differencing proceeds between physically the same pixels, not based on the correspondences inferred from the observation of the target at different time points. This eliminates the danger of minor, even single-pixel misalignments affecting the obtained motion image. On the minus side, along with the relative motion of the body parts, also the irrelevant global motion of the subject as a whole is captured. However, this negative effect can be minimised as described below.

Having the region corresponding to the hypothesis at time t cropped from the appropriate difference image, the obtained motion profile map is normalised. Further, this map is thresholded at a fixed level which for a moving pedestrian typically produces a thick and sometimes broken contour. Finally, the resulting binary map is smoothed using a Gaussian kernel and the Principal Component Analysis (PCA) is applied to the output image. The latter step effectively reduces the dimensionality of the data (to 20 in our experiments). The final motion descriptor is a feature vector constructed from the obtained PCA coefficients. Motion descriptor construction scheme is shown in Fig 3.

It is important to note that PCA applied to the smoothed feature image ensures that only the maximum variance patches are incorporated into the target representation. As a result, the contribution of global motion, common to different actions, is minimised and the discriminative power of the motion descriptor increased. Secondly, the role of thresholding is important. It changes the continuous-valued image into a binary map and hence helps achieve a better representation's invariance of the diverse background and subject's clothing. The differencing itself is insufficient as, say, the high-contrasting top of a human may produce large intensities in the upper part of the motion image, while the low-contrasting trousers will yield much smaller values in the lower part. This increases a chance of different spatial distributions of the figure-background contrast to be learned as independent modes of the motion profile.

B. Classification with Hidden Markov Model

Different gaits are characterised using Hidden Markov Models (HMM). The following steps are taken to determine the number of states in all models. First, our detection module is used to produce stabilised training sequences representing gaits to be recognised. Afterwards, the images from all sequences altogether are transformed as described in section IVA. Obtained feature vectors are then subject to divisive clustering. The number of resulting clusters, interpreted as modes of the human motion profile, determines the number of states in HMMs.

For each gait we train a separate Hidden Markov Model using several sample sequences. Each HMM is parametrised by $\lambda_k = (A_k, B_k, \pi_k)$. Full connectivity between the states is assumed, and the initial state probabilities, π_i, as well as transition probabilities, a_{ij}, are uniform. The distribution of the observation vector $x_t = [x_{t1}, \ldots, x_{td}]$ in each state q_i is modelled with a Gaussian Mixture:

$$b_i(x_t) = \sum_{j=1}^{M_i} c_{ij} N(x_t, \mu_{ij}, \Sigma_{ij}), \qquad (7)$$

where M_i is the number of the i-th mixture's components. The means and covariances of each mixture's components are initialised randomly, but in the latter case the non-diagonal elements are set to zero for simplicity. The Baum-Welch algorithm is used to determine the maximum likelihood estimates of the model parameters. To avoid problems related to the numerical underflow, the scaling technique is used and each operation involving multiplication of probabilities is checked against the numerical floor.

Once the individual gait models have been trained, a new, unseen sequence up to time t, $X_t = [x_1, \ldots, x_t]$, can be classified by picking the model yielding the maximum probability $p(X_t \mid \lambda)$:

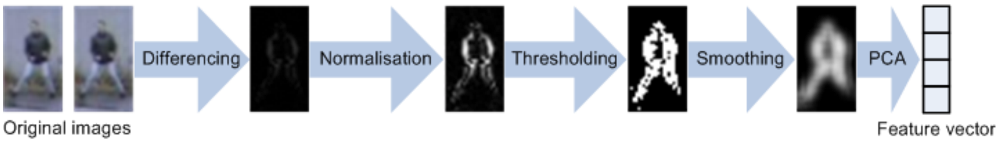

Figure 3. Motion descriptor construction scheme.

$$L(X_t) = \arg\max_k (p(X_t \mid \lambda_k)) . \qquad (8)$$

V. Experimental Results

To train and evaluate our gait recognition system, we have collected a number of outdoor sequences depicting 5 actors and several accidental pedestrians in different clothing, at a varying distance from the camera, and in different places. The actors represent four relatively similar types of gait: walking, running, side jumping, and marching[1]. All gaits are distinguished in two directions: left and right, giving a total of 8 actions. The gait differs from person to person and sometimes the actors were deliberately instructed to change their own style. The detector was set to process as many as 10 distinct scales (differing by a factor of 0.9) in a pyramid fashion. 30 sequences, between 3 and 5 per action, were used for training. The confusion matrices in Tab. 1 show the performance of the classifier obtained for 99 unseen sequences depicting 103 pedestrians. This result is compared with the performance of the same classifier trained on only a single sequence per action.

The overall recognition rate of the classifier reached 91.3%. Note that this result highly depends on the number of sequences used for model training. The average processing speed of approximately 25fps was achieved on a modern PC. However, most of the time required to process a single frame is consumed by the pyramidal processing of the scene. Taking the difficulty of the sequences and the operating resolution of both the detector and the classifier into account, the obtained results are promising. More importantly, the errors are mainly caused by the insufficient training data rather than an inherent weakness of the algorithm. For example, the two confusions between walking left and walking right occurred for a woman wearing a long coat, mostly obscuring the important leg motion information. In the same time, the persons wearing a coat were not presented to the classifier in the training stage. Several other confusions were caused by the inaccuracy of the detector introduced when approximating the scale of a candidate, as described in section III. When a cluster of candidates is merged into a final hypothesis, the optimal scale may differ from the actual discrete approximation. This, in turn, can cause minor misalignments in a stabilised sequence. A special remark is required for explanation of the confusions between the same gaits but observed for the opposite directions of movement. Distinguishing between these actions is trivial based on the analysis of the humans' positions in the consecutive frames. However, we have deliberately chosen not to do it for two reasons. First, from the point of view of a stabilised sequence these actions are not only perfectly legitimate, but also the likelihood of confusion between them is equally high as the likelihood of confusion between any pair of other actions. Second, the introduction of the direction-based variants of the same gait helped us increase the complexity of the classification problem for demonstration purposes at a very low cost.

Evolution of the classifier's decision over time is an interesting problem itself. In Fig. 4. we compare the log likelihoods of 3 sample observation sequences in each HMM to illustrate the confidence of the actual decision. As seen, a reliable decision is typically reached after some period of time. It is caused by the fact that in the first frames it is prevalently dependent on the initial state probabilities. The latter may be significantly biased in the case of few training sequences and periodical actions. An additional irregularity in the expected linearly growing distance between the likelihood scores of the correct and the incorrect classes may be introduced by the inaccuracy of the detector. This is especially the case when the discrete scale of the human bounding box estimate changes at some point to fit the smoothly changing size of the figure being tracked. The Gaussian smoothing used in the construction of the motion descriptor reduces this effect to some extent. In general however, if the model is trained on a sufficiently diverse set of sequences, the correct decision should be reached in the end.

VI. Conclusions

In this paper we have presented a comprehensive approach to detection of moving humans and recognition of their gaits in the real-life scenarios. In the detection stage we utilise a Haar rejection cascade within a pyramid of images to capture the human candidates in the scene at multiple scales simultaneously. For performance optimisation, each frame is pre-filtered to retain only these fragments of the scene where the sufficiently high motion density is recorded. In order to improve the detection accuracy, the detector's discrete response space is treated as a continuous probability distribution with the maxima to be found. Our main contribution in this part is a method of capturing these maxima via a confidence-weighted Mean Shift algorithm, where the confidence of each positive hypothesis of the detector is understood as its distance from the decision boundary. For gait recognition, we have trained a separate Hidden Markov Model per gait using the stabilised training sequences obtained by our detector. The robustness of the classifier is achieved thanks to the proposed representation of the images which relies solely on the motion information extracted via an appropriate combination of simple image processing steps: temporal differencing, normalisation, thresholding, and smoothing. Overall, this method provides a computationally inexpensive yet powerful alternative to a general action recognition in low-resolution video.

We have evaluated our algorithms on a set of video sequences showing eight different types of gait. When recording these sequences, we have explicitly included many types of variations such as different actors, clothings, distances from camera, and the background. A promising recognition rate of over 90% has been reached for the unseen sequences. In the same time the average processing speed of 25fps of both detection and recognition has been reported. This figure is however a rather conservative estimate, as the current prototype system has been set to process as many as 10 distinct scales simultaneously. This number can be reduced to not more than 2-3 scales for more specialised applications.

[1] Sample sequences are available at http://dimas.webd.pl/ICAC/video/

TABLE I. RECOGNITION PERFORMANCE FOR DIFFERENT GAITS: WALKING LEFT/RIGHT (WL/WR), RUNNING LEFT/RIGHT (RL/RR), JUMPING LEFT/RIGHT (JL/JR), AND MARCHING LEFT/RIGHT (ML/MR), OBTAINED FOR 99 UNSEEN SEQUENCES. A 6-STATE HMM CLASSIFIER WAS TRAINED USING A SINGLE SEQUENCE PER TYPE (LEFT), AND BETWEEN 3 AND 5 SEQUENCES PER TYPE (RIGHT).

	WL	WR	RL	RR	JL	JR	ML	MR
WL	11	0	1	2	0	0	8	0
WR	4	10	0	1	1	0	3	0
RL	0	0	7	0	1	0	1	0
RR	0	0	0	7	0	0	2	0
JL	0	0	0	0	6	5	1	0
JR	0	0	0	0	2	8	1	1
ML	0	0	0	0	0	0	10	0
MR	2	0	0	0	0	0	8	0

	WL	WR	RL	RR	JL	JR	ML	MR
WL	21	0	0	0	0	0	0	1
WR	2	16	0	0	0	0	0	0
RL	0	0	8	0	1	0	0	0
RR	0	1	0	9	0	0	0	0
JL	0	0	0	0	10	1	1	0
JR	0	0	0	0	0	12	0	0
ML	0	0	0	0	0	0	10	0
MR	1	0	0	0	0	0	1	8

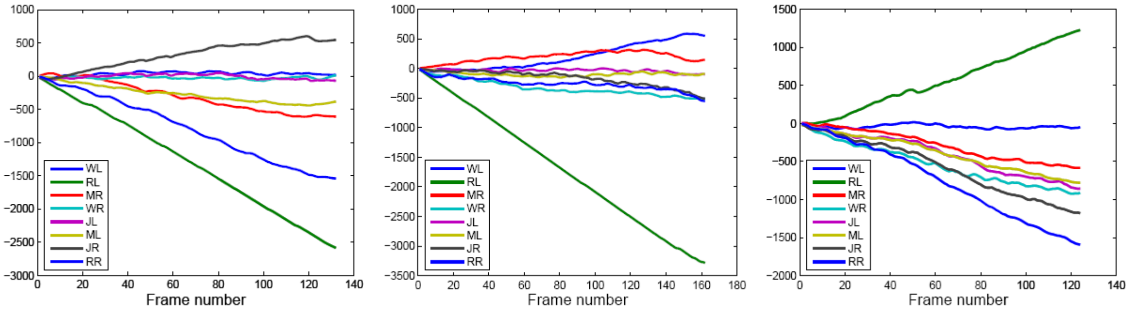

Figure 4. Evolution of the classifier's decision in the sample image sequences depicting different gaits: jumping right (left), running right (centre), running left (right). This figure is best viewed in colour.

REFERENCES

[1] L. R. Rabiner and B. H. Juang, "An introduction to Hidden Markov Models", IEEE ASSP Magazine, p. 4–15, 1986.

[2] D. Meyer, J. Pösl and H. Niemann, "Gait classification with HMMs for trajectories of body parts extracted by mixture densities", Proc. of the British Machine Vision Conference, p. 459–468, 1998.

[3] C. Papageorgiou and T. Poggio, "A trainable system for object detection", International Journal of Computer Vision, 38(1), p. 15–33, 2000.

[4] A. Broggi, M. Bertozzi, A. Fascioli and M. Sechi, "Shape-based pedestrian detection", Proc. of the IEEE Intelligent Vehicles Symposium, p. 215–220, 2000.

[5] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications", IEEE Trans. on PAMI, 22(8), p. 781–796, 2000.

[6] L. Zhao and C. E. Thorpe, "Stereo- and neural network-based pedestrian detection", IEEE Trans. on Intelligent Transportation Systems, 1(3), p. 148–154, 2000.

[7] A. Mohan, C. Papageorgiou and T. Poggio, "Example-based object detection in images by components", IEEE Trans. on PAMI, 23(4), p. 349–360, 2001.

[8] A. F. Bobick and J. W. Davis, "The representation and recognition of action using temporal templates, IEEE Trans. on PAMI, 23(3), p. 257–267, 2001.

[9] D. Comaniciu and P. Meer, "Mean shift: a robust approach towards feature space analysis", IEEE Trans. on PAMI, 24(5), p. 603–619, 2002.

[10] P. Viola, M. Jones and D. Snow, "Detecting pedestrians using patterns of motion and appearance", Proc. of the 9th Int. Conf. on Computer Vision, vol. 2, p. 734–742, 2003.

[11] A. A. Efros, A. C. Berg, G. Mori and J. Malik, "Recognizing action at a distance", Proc. of the 9th Int. Conf. on Computer Vision, vol. 2, p. 726–734, 2003.

[12] G–J. Tian, F–Y. Hu and R–C. Zhao, "Gait recognition based on Fourier descriptors", Proc. of Int. Symposium on Intelligent Multimedia, Video and Speech Processing, p. 29–32, 2004.

[13] M. Bertozzi, A. Broggi, A. Fascioli, T. Graf and M–M. Meinecke, "Pedestrian detection for driver assistance using multiresolution infrared vision", IEEE Trans. on Vehicular Technology, 53(6), p. 1666–1678, 2004.

[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, 60(2), p. 91–110, 2004.

[15] P. Viola and M. Jones, "Robust real-time object detection", International Journal of Computer Vision, 57(2), p. 137–154, 2004.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, p. 886–893, 2005.

[17] B. Leibe, E. Seemann and B. Schiele, "Pedestrian detection in crowded scenes", Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, p. 878–885, 2005.

[18] Q. Zhu, S. Avidan, M. C. Yeh and K. T. Cheng, "Fast human detection using a cascade of histograms of oriented gradients", Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, p. 1491–1498, 2006.

[19] D. Weinland, R. Ronfard and E. Boyer, "Free viewpoint action recognition using motion history volumes", Computer Vision and Image Understanding, 104(2), p. 249–257, 2006.

[20] Daimler-Chrysler pedestrian classification benchmark, http://www.science.uva.nl/research/isla/downloads/pedestrians/, 2006.

[21] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features", Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, p. 1–8, 2007.

[22] O. Tuzel, F. Porikli and P. Meer, "Human detection via classification on riemannian manifolds", Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, p. 1–8, 2007.

[23] Z. Li, W. Bo and R. Nevatia, "Pedestrian detection in infrared images based on local shape features", Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, p. 1–8, 2007.

[24] L. Gorelick, M. Blank, E. Shechtman, M. Irani and R. Basri, "Actions as space-time shapes", IEEE Trans. on PAMI, 29(12), p. 2247–2253, 2007.