Combined Face and Gait Recognition using Alpha Matte Preprocessing

Martin Hofmann[1], Stephan M. Schmidt[1], AN. Rajagopalan[1,2], Gerhard Rigoll[1]
[1] Institute for Human-Machine Communication, Technische Universität München, Germany
[2] Department of Electrical Engineering, Indian Institute of Technology Madras, India
martin.hofmann@tum.de, stephan.schmidt@mytum.de, raju@iitm.ac.in, rigoll@tum.de

Abstract

This paper presents advances on the Human ID Gait Challenge. Our method is based on combining an improved gait recognition method with an adapted low resolution face recognition method. For this, we experiment with a new automated segmentation technique based on alpha-matting. This allows better construction of feature images used for gait recognition. The same segmentation is also used as a basis for finding and recognizing low-resolution facial profile images in the same database. Both, gait and face recognition methods show results comparable to the state of the art. Next, the two approaches are fused (which to our knowledge, has not yet been done for the Human ID Gait Challenge). With this fusion gain, we show significant performance improvement. Moreover, we reach the highest recognition rates and the largest absolute number of correct detections to date.

1. Introduction

The focus of this paper is on recognizing people from larger distances. At a distance, many typical *physiologic* features, such as fingerprint, DNA, hand, ear, retina and face, are obscured or cannot be obtained at all. By contrast, *behavior based* features such as gait features can be extracted from walking people at a distance.

In our approach we make use of gait recognition combined with person identification based on low-resolution face profile images. As such we combine physiologic and behavior based features. We show that both modalities lead to good results on their own. When combining them, we observe a significant improvement in recognition performance, which demonstrates the strength of a multimodal approach.

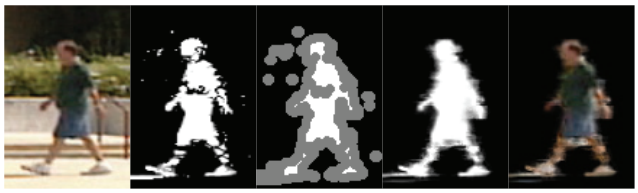
Primarily our approach is motivated by the success of gait recognition methods for recognition at a distance. In 1967, Murray [11] suggested that if all gait movements are considered, gait is unique. Early studies in 1977 by Cutting and Kozlowski [2] suggest that it is possible to recognize friends from just their way of walking. Later, Stevenage *et al*. [15] showed that people can be recognized without any information on the body-shape, only using gait features. A major advantage of these behavior based features over other physiologic features is the possibility to identify people from large distances and without the person's direct cooperation. Also no direct interaction with a sensing device is necessary, which allows for undisclosed identification. Thus gait recognition has great potential in video surveillance, tracking and monitoring.

For low resolution data, gait recognition has its clear advantages. However, in our approach, we also use low resolution face data. Even though face recognition has its performance peak at high resolution frontal face images, it can still be seen that facial profile recognition can contribute to the performance, when combined correctly.

A multitude of gait recognition algorithms (see Table 1) have so far been proposed, which leads to a rich set of results we can compare to. Most of these methods build solely on the binarized silhouette images. However we feel that a lot of identity information gets lost by this early binarization. Thus instead of binarizing, both our face and gait recognition methods build on a novel automated color foreground segmentation method based on alpha-matting. For gait recognition we use the continuous alpha-matte segmentation and show a small increase in performance. To our knowledge so far face recognition has not been applied to the Human ID Gait database [12], so we cannot compare these results directly. When fusing gait and face features we observe a significant performance gain, such that our combined method outperforms the state of the art.

2. Related Work

Generally speaking there are two kinds of gait recognition methods. On the one hand model-based methods, on the other hand model-free methods. Model based methods [1][21] define a (simplified) human model and match the gait sequences to this model. Gait recognition is then performed on the temporal change of the model parameters, such as leg angles [21]. Those methods are typically very



(a) (b) (c) (d) (e)

Figure 1: Left to right: input image; foreground segmentation; tri-state labeling with morphologic operations; alpha matte; final segmentation

demanding and good results are hard to achieve. Model-free methods [3][5][7][9][12][17][19][20] on the other hand have shown more success in the recent past. Here, the person identity is directly inferred from the features without an intermediate person model. Most methods build on a silhouette extraction for each frame in a gait cycle. Silhouettes are either averaged [3][9][19], or all silhouettes are used simultaneously [7][12][16]. Different classifiers ranging from nearest neighbor [3], SVM and HMM [7][16] have been applied with similarly good results.

Recently gait recognition has been combined with face recognition [6][10][22]. Typical face recognition methods require a high resolution frontal face image. However for gait recognition, persons are only captured in low-resolution side view images. In [6], for face recognition, only the final segment of the gait video, where the person is visible in near frontal, is used. In [13], multiple cameras are used to ensure that both the side view, as well as the frontal view are available. To avoid these special cases, face recognition can be performed on the low-resolution side view images [22]. Our approach is similar to the latter ones, because we also do not depend on specialized data, but instead work directly on the low-resolution side view videos.

For performance evaluation, many databases have been recorded. However, the most popular and widely used database is probably the Human ID Gait database [12]. This database features video sequences of a total of 122 subjects, which walk perpendicular to the camera at a distance. While many methods have been applied to this dataset, so far no fusion method using gait and face was ever applied to this database.

3. Segmentation using Alpha Mattes

In this work, we investigate a new segmentation technique which we apply to both gait recognition as well as face recognition. Current gait recognition methods rely on good segmentation to extract the contour and the silhouettes of the foreground objects. Typically, a background is estimated by calculating the mean and variance of the scene over a certain period. Then the foreground is estimated by finding the pixels with significant deviation from the background model. This leads to a noisy, binary segmentation as depicted in Figure 1b). However, due to the nature of the image capturing, there is a band on the silhouette which belongs partially to foreground and partially to background. Thus at each pixel (x, y), the image I is modeled as a linear composition of the foreground F and the background B:

$$I(x,y) = \alpha(x,y)F(x,y) + (1 - \alpha(x,y))B(x,y) \quad (1)$$

Here, $\alpha(x, y)$ is the opacity of the pixel at (x, y). $F(x, y)$, $B(x, y)$ and $\alpha(x, y)$ are unknown. For a typical color image with three color channels we thus have 7 unknowns to solve for at each pixel. This kind of problem statement is typical for matting problems. To leverage the high number of unknowns, proximity and smoothness assumptions are made. Also the typical matting application has a human in the loop who has to provide some scribbles for foreground and background, leading to the so called *trimap*. This map contains regions which are definitely foreground ($\alpha(x, y) = 1$), some which are definitely background ($\alpha(x, y) = 0$) and some unknown regions for which the matting method determines the $\alpha(x, y)$.

However, for automated gait recognition it is infeasible to have a human in the loop. We therefore automatically generate the trimap from the noisy foreground segmentation. We get the definite-foreground regions ($\alpha(x, y) = 1$) by eroding the foreground segmentation with a circular structure element with radius $r = 4$. The definite-background regions are obtained by eroding the background region with the same circular structure element. The resulting trimap is shown in Figure 1c).

For background segmentation we use Gaussian mixture models [14], for alpha matting we used closed form matting [8].

The resulting foreground segmentation – the alpha-matte – is depicted in Figure 1d). It can be seen that this segmentation is superior to the initial background segmentation. Holes are closed, erroneous pixels are removed and most of all, the smooth transition of the foreground to the background is captured. Furthermore by $F(x, y) = I(x, y) \cdot \alpha(x, y)$ we can approximate a precise color segmentation of the foreground object (see Figure 1e)). This color segmentation is used for the face recognition part.

4. Gait recognition

4.1. Feature Extraction using α-GEI

For gait recognition we use a method based on the classical Gait Energy Image (GEI) [3]. However, instead of using binary silhouettes, we use the alpha channel from the alpha matting as described in the previous section. We call this the Alpha Gait Energy Image (α-GEI)

In essence, the Alpha Gait Energy Image is an arithmetic mean of the alpha channel. Denote α_t the alpha matte in frame t. Then, the α-GEI g is formally defined as the alpha matte average over one full gait cycle:

$$g(x,y) = \frac{1}{T}\sum_{t=1}^{T}\alpha_t(x,y) \tag{2}$$

4.2. Feature Space Reduction

The gait energy images $g(x,y)$ have a resolution of 88×128 pixels. Thus the feature vector is still large with 11264 dimensions. We apply principal component analysis (PCA) followed by multiple discriminant analysis (MDA) to reduce the size of the feature vector. A combination of PCA and MDA, as proposed in [4], results in the best recognition performance. While PCA seeks a projection that best represents the data, MDA seeks a projection that best separates the data.

Assume that the training set, consisting of N d-dimensional training vectors $\{g_1, g_2, \ldots, g_N\}$, is given. Then the projection to the $d' < d$ dimensional PCA space is given by

$$y_k = U_{pca}(g_k - \overline{g}), \qquad k = 1, \ldots, N \tag{3}$$

Here U_{pca} is the $d' \times d$ transformation matrix with the first d' orthonormal basis vectors obtained using PCA on the training set $\{g_1, g_2, \ldots, g_N\}$ and $\overline{g} = \sum_{k=1}^{N} g_k$ is the mean of the training set. After PCA, MDA is performed. It is assumed that the reduced vectors $\mathcal{Y} = \{y_1, y_2, \ldots, y_N\}$ belong to c classes. Thus the set of reduced training vectors $\mathcal{Y}$ is composed of its c disjunct subsets $\mathcal{Y} = \mathcal{Y}_1 \cap \mathcal{Y}_2 \cap \ldots \mathcal{Y}_c$. The MDA projection has by construction $(c-1)$ dimensions. These $(c-1)$ dimensional vectors z_k are obtained as follows

$$z_k = U_{mda} y_k, \qquad k = 1, \ldots, N \tag{4}$$

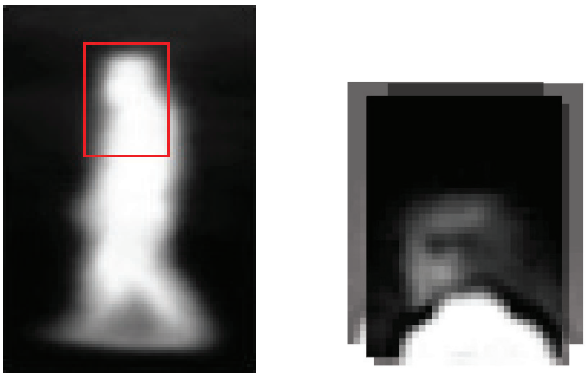
where U_{mda} is the transformation matrix obtained using MDA. This matrix results from optimizing the ratio of the between-class scatter matrix S_B and the within-class scatter matrix S_W:

$$J(U_{mda}) = \frac{|\widetilde{S}_B|}{|\widetilde{S}_W|} = \frac{|U_{mda}^T S_B U_{mda}|}{|U_{mda}^T S_W U_{mda}|}. \tag{5}$$

Here the within-class scatter matrix S_W is defined as $S_W = \sum_{i=1}^{c} S_i$, with $S_i = \sum_{y \in \mathcal{Y}_i}(y - m_i)(y - m_i)^T$ and $m_i = \frac{1}{N_i}\sum_{y \in \mathcal{Y}_i} y$. Where $N_i = |\mathcal{Y}_i|$ is the number of vectors in $\mathcal{Y}_i$. The between-class scatter S_B is defined as $S_B = \sum_{i=1}^{c} N_i(m_i - m)(m_i - m)^T$, with $m = \frac{1}{N}\sum_{i=1}^{c} N_i m_i$.

Finally, for each Gait Energy Image, the corresponding gait feature vector is computed as follows

$$z_k = U_{pca}U_{mda}(g_k - \overline{g}) = T(g_k - \overline{g}), \qquad k = 1, \ldots, N \tag{6}$$



(a) (b)

Figure 2: a) Rough definition of the pre-face around the face region. b) Registration of the pre-faces using sum of absolute differences.

4.3. Classification

Each class c is modeled with only one vector, which is the mean feature vector $\overline{z}_c$:

$$\overline{z}_c = \frac{1}{|\mathcal{Z}_c|}\sum_{z \in \mathcal{Z}_c} z. \tag{7}$$

For each α-GEI from the test set $\hat{g}_i$, we perform the transformation in Equation 6 to get the reduced feature vector $\hat{z}_i$. A distance $D_i^{gait}(c) = ||\hat{z}_i - \overline{z}_c||$ using Euclidean distance measure is defined. It defines for all sequences i, the distance to the c-th class. Final person identification using gait then becomes a nearest-neighbor classification. We assign a class label L_i to each test gait image according to

$$L_i = \operatorname*{argmin}_c D_i^{gait}(c) \tag{8}$$

5. Face recognition

5.1. Pre-faces

In the first part of the algorithm, the gallery set is processed. The goal is to find a 20×20 patch of the face profile of each person. To robustly achieve this and to avoid erroneous segmentations, first for each gallery sequence a pre-face is calculated. To this end, the mean of all frames in a sequence is calculated (similar to GEI), in order to find the person more precisely than using a bounding box. Over this mean image, a 30×40 patch is defined, which is used to cut the region for all frames (see Figure 2).

Because viewing direction and body positions slightly changes when the person walks across the scene, instead of only extracting one face per sequence, multiple such faces, which are evenly spread over the sequence, are extracted. This ensures that as much information about the person is

(a)

(b)

Figure 3: a) The alpha matte based segmentation; the roughly cropped pre-face and the final face segmentation. b) Several sub faces of a specific sequence. It can be clearly seen that the appearance of a face changes within the sequence.

captured as possible. Thus, always five consecutive pre-faces are combined. Those five pre-faces are registered using sum of absolute differences. After registration, the mean is taken to find the averaged pre-face.

Finally, to find the precise head location within the 30×40 pixel pre-face, a simple threshold method is used to find the highest point (top of head) and the left-most point (nose). Using these two points, a 20×20 pixel patch is extracted, which captures the final segmentation of the face. Results of segmentation can be seen in Figure 3. Note that due to the alpha matte preprocessing the segmentations contain only color foreground regions. Disturbing background pixels are eliminated.

The same segmentation is carried out on the test sequences. The splitting of the test sequences has the advantage, that for each sequence, multiple sub faces of each person can be used for classification. This way, multiple aspects of the person are captured and in addition, the influence of erroneous segmentations is reduced.

5.2. Eigenface Calculation

We apply the classical eigenface method [18] for face recognition. This means that the average face is calculated by taking the mean. This average face is subtracted from the gallery faces and a covariance matrix is estimated from the gallery data. Thus a PCA is performed. In order to capture color information like skin and hair color, all three color channels are appended and used for the calculation of the covariance matrix.

Let $\{f_1, f_2, \ldots, f_M\}$ be the set of M $20 \times 20 \times 3$ color face patches in the gallery set. Here M is number of all sub faces, so it is roughly 40 times larger than the number of people in the gallery set. Then the resulting transformation is

$$v_k = U_{face}(f_k - \overline{f}) \tag{9}$$

where $\overline{f} = \sum_{k=1}^{N} f_k$ is the mean face and U_{face} is the transformation matrix learned by PCA.

5.3. Classification

Face recognition is done similarly to gait recognition. However, instead of having one average gait template, we have several sub faces for each sequence as described above. Typically one would use k-nearest neighbor in such a case. For the later fusion step, however, we need a continuous score for each potential class. Thus for each of the sub faces of a test sequence we calculate the distance to all sub faces of all trainings sequences (see Figure 4). Out of these matches, we only keep the k nearest matches. Within these k matches, the average distance to all comprised classes is averaged, thus resulting in a distance $D_i^{face}(j)$. If a class c is not comprised in the k best matches at all, then the distance is set to $D_i^{face}(j) = \infty$. In our experiments we set $k = 100$, however the method is not sensitive to this value as long as it is big enough (> 10).

For pure face classification the class c with the minimal distance $\underset{c}{\operatorname{argmin}}\, D_i^{face}(c)$ is taken as the recognition result.

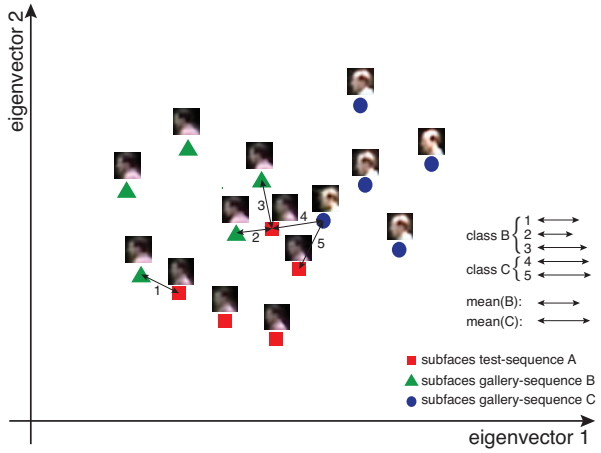


Figure 4: Illustration of the face classification (shown for the first two eigenvalues). For a given test sequence A, the k closest matches are found (here $k = 5$). Within those top k matches, the class averages (here, to class B and C, respectively) are a measure for the similarity to these classes.

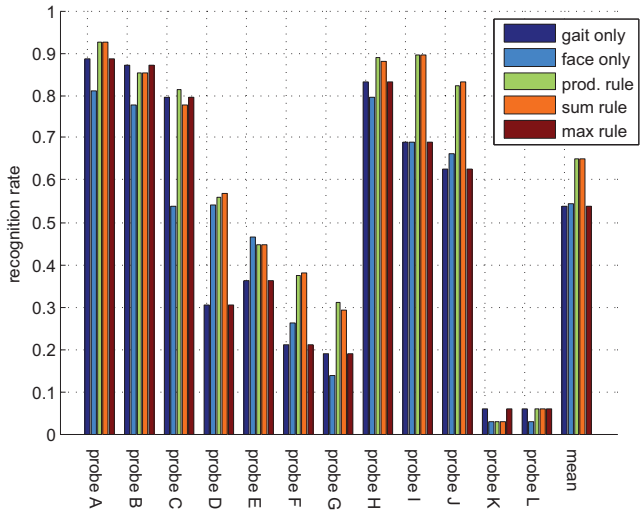


Figure 5: Quantitative results on the Human ID Gait database [12]. (1) using only gait information, (2) using only face, (3) fusion using product rule, (4) fusion using sum rule, (5) fusion using max rule

6. Fusion of Face and Gait

In this work we use score level fusion. This means that the distance scores $D^{gait}(c)$ and $D^{face}(c)$ are combined before decision making. There are multiple ways of fusing the results. We use max, product and sum rules:

$$D_i(c) = D_i^{gait}(c) \cdot D_i^{face}(c) \quad (10)$$

$$D_i(c) = D_i^{gait}(c) + D_i^{face}(c) \quad (11)$$

$$D_i(c) = \max(D_i^{gait}(c), D_i^{face}(c)) \quad (12)$$

$$(13)$$

The distances result from different modalities, thus the values are not directly comparable. Therefore normalization of the vectors is of central importance. Before fusion, the vectors are normalized to have unit length, i.e. $D(c) \leftarrow D(c)/\sum_{\hat{c}} D(\hat{c})$.

7. Results and Comparison

Figure 5 shows the quantitative results on the Human ID Gait database. It can be seen that fusion using either the product rule or the sum rule greatly improves the recognition rates, except Probe B, where fusion slightly reduces recognition rates of gait, but greatly increases results of face recognition. The max rule shows inferior performance.

For performance evaluation, we compare our method to several state-of-the-art results. Summarizing results are shown in Table 1 (largely taken from [5]). Here, recognition rates for all 12 experiments, as well as the weighed recognition average are shown.

It can be seen that our α-GEI $(53.6.0\%)$ - which does not use synthetic images as in [3] - outperforms the standard GEI (48.2%). This demonstrates the effectiveness of the alpha matte preprocessing and it can be foreseen that when implementing synthetic images, recognition rates can be even improved further. We cannot compare our α-eigenface method, since currently no other face recognition method was applied to the Human ID Gait database.

Both our face $(54, 6\%)$ and our gait recognition method $(53, 6\%)$ alone cannot compete with the current state of the art. However, when combining these multimodal methods, recogniton rates exceed all previous approaches. This shows the importance of simultaneously using multiple modalities and fusing them. It can be seen that simple product and sum rules lead to good fusion results and to adramatic increase in performance.

8. Conclusion and Outlook

In this work, a new preprocessing method using closed form alpha matting was introduced. It was applied to both face and gait recognition. In order to use this method, which typically requires a "human in the loop", an automated generation of the trimap was presented. Using this preprocessing it was possible to increase the performance of the standard Gait Energy Image.

Combining both the modified face and gait recognition method, it was possible to achieve unprecedented performance results on the Human ID Gait challenge. Similar fusion techniques have currently only been carried out on other (smaller) datasets.

For future work, stronger and better face and gait methods should be combined. It can be foreseen that recognition rates could improve even further.

References

[1] C. BenAbdelkader, R. Cutler, and L. Davis. Stride and cadence as a biometric in automatic person identification and verification. In *Proceedings Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 372–377. IEEE, 2002.

[2] J. Cutting and L. Kozlowski. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, 9(5):353–356, 1977.

[3] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 316–322, 2006.

[4] P. Huang, C. Harris, and M. Nixon. Recognising humans by gait via parametric canonical space. *Journal of Artificial Intelligence in Engineering*, 13(4):359–366, November 1999.

[5] Y. Huang, D. Xu, and T.-J. Cham. Face and human gait recognition using image-to-class distance. *IEEE Trans. Circuits Syst. Video Techn.*, 20(3):431–438, 2010.

Probe Set	A	B	C	D	E	F	G	H	I	J	K	L	avg.
Probe Size	122	54	54	121	60	121	60	120	60	120	33	33	-
Baseline [12]	73	78	48	32	22	17	17	61	57	36	3	3	41,0
HMM [7]	89	88	68	35	28	15	21	85	80	58	17	15	53,5
IMDE [19]	75	83	65	25	28	19	16	58	60	42	2	9	42,9
IMDE + LDA [19]	88	86	72	29	33	23	32	54	62	52	8	13	48,6
GEI [3]	87	85	76	31	30	18	21	63	59	54	3	6	48,2
GEI + Synth [3]	91	94	81	51	57	25	29	62	60	57	9	12	55,8
GDN [9]	85	89	72	57	66	46	41	83	79	52	15	24	62,8
MMFA [20]	89	94	80	44	47	25	33	85	83	60	27	21	59,9
GTDA [17]	91	93	86	32	47	21	32	95	90	68	16	19	60,6
I-to-C [5]	93	89	81	54	52	32	34	81	78	62	12	9	61,2
our α-eigenface	81	78	54	54	47	26	14	80	69	66	6	3	54,6
our α-GEI	89	87	79	30	36	21	19	83	69	63	6	6	53,6
our fusion	93	85	81	56	45	38	31	89	90	82	3	6	**65,2**

Table 1: Comparison of our method to other methods (all rank 1): Baseline [12], Hidden Markov Models (HMM) [7], IMage Euclidean Distance (IMED) [19], Gait Energy Image (GEI) [3], Gait Dynamics Normalization (GDN) [9], General Tensor Discriminant Analysis (GTDA) [17], Image-to-Class Distance (I-to-C) [5]

[6] A. Kale, A. Roychowdhury, and R. Chellappa. Fusion of gait and face for human identification. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, volume 5, pages V – 901–4 vol.5, may 2004.

[7] A. Kale, A. Sundaresan, A. Rajagopalan, N. Cuntoor, A. RoyChowdhury, and V. Krueger. Identification of humans using gait. *IEEE Transactions on Image Processing*, 13(9):1163–1173, 2004.

[8] A. Levin, D. Lischinski, and Y. Weiss. A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:228–242, 2008.

[9] Z. Liu and S. Sarkar. Improved gait recognition by gait dynamics normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 863–876, 2006.

[10] Z. Liu and S. Sarkar. Outdoor recognition at a distance by fusing gait and face. *Image Vision Comput.*, 25:817–832, June 2007.

[11] M. Murray. Gait as a total pattern of movement. *American Journal of Physical Medicine*, 46(1):290, 1967.

[12] S. Sarkar, P. Phillips, Z. Liu, I. Vega, P. Grother, and K. Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE transactions on pattern analysis and machine intelligence*, pages 162–177, 2005.

[13] G. Shakhnarovich and T. Darrell. On probabilistic combination of face and gait cues for identification. In *Automatic Face and Gesture Recognition, 2002. Proceedings*, pages 169 –174, may 2002.

[14] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:2246, 1999.

[15] S. Stevenage, M. Nixon, and K. Vince. Visual analysis of gait as a cue to identity. *Applied Cognitive Psychology*, 13(6):513–526, 1999.

[16] A. Sundaresan, A. Chowdhury, and R. Chellappa. A hidden markov model based framework for recognition of humans from gait sequences. *Proceedings IEEE International Conference on Image Processing*, 2:II – 93–6 vol.3, 2003.

[17] D. Tao, X. Li, X. Wu, and S. Maybank. General tensor discriminant analysis and gabor features for gait recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(10):1700 –1715, oct. 2007.

[18] P. Viola and M. Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001.

[19] L. Wang, Y. Zhang, and J. Feng. On the euclidean distance of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1334 –1339, aug. 2005.

[20] D. Xu, S. Yan, D. Tao, S. Lin, and H.-J. Zhang. Marginal fisher analysis and its variants for human gait recognition and content- based image retrieval. *Image Processing, IEEE Transactions on*, 16(11):2811 –2821, nov. 2007.

[21] C. Yam, M. Nixon, and J. Carter. Automated person recognition by walking and running via model-based approaches. *Pattern Recognition*, 37(5):1057–1072, 2004.

[22] X. Zhou and B. Bhanu. Feature fusion of face and gait for human recognition at a distance in video. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 4, pages 529 –532, 0-0 2006.