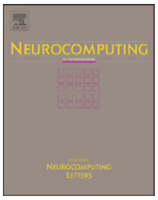
Neurocomputing 71 (2008) 1931–1938

Contents lists available at ScienceDirect

Neurocomputing

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



Fusing gait and face cues for human gender recognition

Caifeng Shan [a,*], Shaogang Gong [b], Peter W. McOwan [b]

[a] Philips Research, High Tech Campus 36, 5656 AE Eindhoven, The Netherlands
[b] Department of Computer Science, Queen Mary, University of London, Mile End Road, London E1 4NS, UK

ARTICLE INFO

Available online 18 March 2008

Keywords:
Gender classification
Canonical correlation analysis
Gait energy image
Support vector machine

ABSTRACT

Computer vision-based gender classification is an interesting and challenging problem, and has potential applications in visual surveillance and human–computer interaction systems. In this paper, we investigate gender classification from human gaits in image sequences, a relatively understudied problem. Moreover, we propose to fuse gait and face for improved gender discrimination. We exploit canonical correlation analysis (CCA), a powerful tool that is well suited for relating two sets of measurements, to fuse the two modalities at the feature level. Experiments demonstrate that our multimodal gender recognition system achieves the superior recognition performance of 97.2% in large data sets.

1. Introduction

Gender classification is one of the most important visual tasks for human beings, as many social interactions critically depend on the correct gender perception. As visual surveillance and human–computer interaction technologies evolves, computer vision systems for gender classification will play an increasing important role in our lives, e.g., collecting valuable demographic information in a social environment.

As human faces provide important visual information for gender perception, a very large number of psychophysical studies has investigated gender classification from face perception [3,32]. Recently this problem has been considered more technically using machine learning methods on large data sets [31,35]. However, in the real-world unconstrained situations, due to the arbitrary walking direction and continuously varying head pose, face information sometimes is unreliable or unavailable. More crucially, with people walking at a distance, face information cannot be measured reliably at low resolution. In these situations, human gait, or the style of walking, can provide important alternative cues for gender classification, as gaits can be detected and measured from arbitrary views and at a distance. Human walking manners contain subtle, yet informative variations. Psychophysical studies [24,1,29] have shown that, even for point-light displays (filmed by attaching small point-lights to the main joints of human body in a homogeneously dark background), people can recognize the gender of walkers. However, given the ability of humans to classify gender by the gaits, there have been few computer vision systems developed for gender recognition from gaits. Compared to facial gender classification, this problem is relatively understudied, although more recently some studies appeared [26,9]. However, there are some limitations in these existing tentative attempts, for example, point-light display from the aspect of biological motion, not visual features from images, was considered in [9], and relatively small data sets were used in [26].

In this paper, we investigate gender classification from human gaits in image sequences using machine learning methods on large data set. Considering each modality, face or gait, in isolation has its inherent weakness and limitations, we further propose to fuse gait and face for improved gender discrimination. We exploit canonical correlation analysis (CCA), a powerful tool that is well suited for relating two sets of signals, to fuse the two modalities at the feature level. Experiments demonstrate that our multimodal gender recognition system achieves the superior recognition performance of 97.2%. We plot in Fig. 1 the flow chart of our multimodal gender recognition system.

2. Previous work

Gender classification has been an active topic in psychological and cognitive literatures [24,1,3,32]. Machine learning-based computer vision methods have been proposed in recent years. In this section, we briefly review and summarize the previous work in visual gender classification.

[*] Corresponding author.
E-mail addresses: caifeng.shan@philips.com (C. Shan), sgg@dcs.qmul.ac.uk (S. Gong), pmco@dcs.qmul.ac.uk (P.W. McOwan).

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2007.09.023

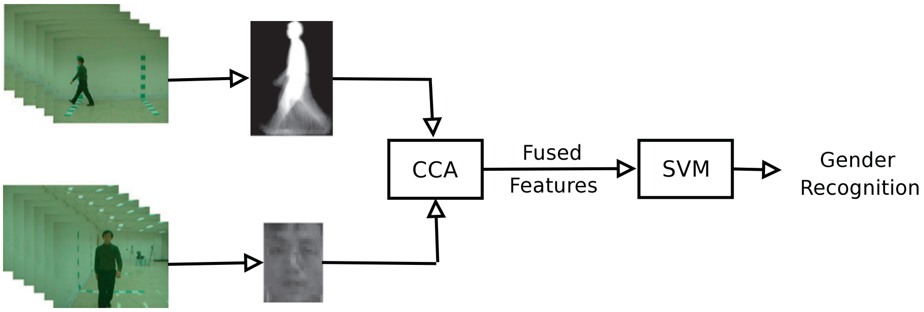


Fig. 1. The flow chart of our multimodal gender recognition system.

2.1. Learning gender from faces

Most of the existing work attempt to classify gender from human faces. In the early 1990s various neural network techniques were employed for gender classification from a frontal face [12,4,37,15]. Golomb et al. [12] trained a fully connected two-layer neural network, SEXNET, to identify gender from face images. Brunelli and Poggio [4] developed HyperBF networks for gender classification in which two competing networks, one for male and the other for female, are trained using 16 geometric features. Some of these techniques are appearance-based methods, that is, they learn the decision boundary between male and female classes from training images without extracting any geometrical features, while others are based on geometrical features.

Recently Moghaddam and Yang [31] investigated non-linear support vector machines (SVMs) for gender classification with low-resolution thumbnail face, and demonstrated the superior performance of SVMs to other classifiers. Graf and Wichmann [14] investigated the influence of two popular dimensionality reduction, principal component analysis (PCA) and LLE, on SVMs classification. Overall, PCA provides superior performance in classification and allowing linear separability. Walawalkar et al. [40] adopted SVMs for gender classification using audio and visual cues. Costen et al. [5] considered a class of sparse regularization functions to develop sparse classifiers for determining facial gender. The sparse classification method aims to both select optimal features and maximize the classification margin, in a manner similar to SVMs. Jain and Huang [21] adopted ICA to represent face images in low-dimensional subspace, and then used linear discriminant analysis (LDA) to perform gender recognition.

Shakhnarovich et al. [35] developed a real-time face detection and demographic analysis (female/male and asian/non-asian) system using Adaboost [39], which delivers slight better performance than the non-linear SVMs [31] on unaligned faces from real-world unconstrained video sequences. Recently Wu et al. [42] presented a look up table (LUT) weak classifier-based Adaboost for gender classification. Sun et al. [36] employed Genetic Algorithms to select a subset of optimal features from the low-dimensional PCA subspace by disregarding certain eigenvectors that encode less gender information, reporting the best performance (95.3%) using the SVM classifier on 400 images. Wichmann et al. [41] investigated gender discrimination of human faces by combining psychological experiments with machine learning methods. They trained a set of linear classifiers, linear SVMs, relevance vector machines (RVMs), LDA and prototype (prot) classifiers, in the PCA subspace. The entire system acts as a linear classifier, allowing them to visualizing the *decision-image* corresponding to the normal vector of the separating hyperplanes of each classifier. A psychological discrimination experiment demonstrates that the female-to-maleness transition along the normal vector closely mimicking human classification is faster than the transition along any other direction. Their experiments also suggest that human subjects base their gender classification strongly on the eye and mouth regions of the face.

2.2. Learning gender from gaits

Gender recognition from point-light display of human walking has received much attention in psychological field during the past few decades. Kozlowski and Cutting [24] performed the first major experiment with six walkers (three females and three males) of approximately the same height and weight recorded at a sagittal view. They demonstrated that human observers could classify the gender of the walkers with average recognition rate of 63%, and alterations such as varying the arm swing, changing the walking speed, and occluding portions of the body do not significantly influence recognition performance. Barclay et al. [1] carried out further study by examining temporal and spatial factors. They reported that successful gender recognition required exposure to approximately two walking cycles, and the rendering speed has a strong influence over recognition. The effect of inversion on the point-lights was also investigated, and it is found that the gender assignments were significantly reversed. They proposed a view-based explanation based on the shoulder–hip ratio, in which men tend to have broader shoulders and smaller hips than women. Cutting et al. [6] supported the shoulder–hip concept and proposed a related center-of-moment feature of the torso. The shoulder–hip ratio and center-of-moment features [1,6] are mainly based on the structural differences between male and female walkers. However, there are certainly dynamic features of movement that contributes to recognition. By setting structural and dynamic features into confliction using a synthetic point-light walker, Mather and Murdoch [29] found that shoulder sway was an effective cue to gender at the frontal view. Although most of the study were conducted using a side-view presentation of the walkers to observers, the effect of view angle on gender recognition performance were examined in [18,29,38].

Much of the previous studies has focused on the manual identification of key features that enable the perceptual classification between female and male walking styles. Features related to speed, arm swing, shoulder–hip lengths, inversion, and body sway have been examined. However, to date there is no conclusive evidence as to which features actually drive the discrimination process. It seems that gender information is not a matter of a single feature, but rather involves multiple combined features. Troje [38] recently treated the analysis of biological motion as a linear pattern recognition problem, and presented a two-stage PCA framework for recognizing gender. The first PCA decomposed each walker's data into its eigenspace, and a second PCA was applied to all walker eigenspaces followed by a linear classifier. He reported 92.5% recognition rates. Davis and Gao [9,10] more recently presented an approach for gender recognition of point-light walkers using an expressive three-mode PCA model [7,8]. Their method first constructs a PCA representation of point-light trajectories for a prototype female and male walker. A large labeled set were are then used to automatically learn which trajectories in the prototype PCA representation best express the gender of the walkers. The non-expressive trajectories are removed and the remaining trajectories are weighted to bias the gender estimation method to produce the desired gender labels.

The aforementioned studies used point-light display from the aspect of biological motion. Lee and Grimson [25,26] adopted computer vision techniques to extract visual features of gaits from image sequences for gender classification. For each scale-normalized binary silhouette, they found the centroid and divided the silhouette into seven parts roughly corresponding to head/shoulder, arms/torso (front and back), thighs (front and back), and calve/feet (front and back), and then extracted moment-based features from each part to represent gait dynamics. Using SVMs as classifiers, their approach achieved performance of 84.5% on a small data set (10 women and 14 men). Recently Yoo et al. [43] studied gender discrimination by gaits using a much larger database (84 males and 16 females). They used a 2D stick figure (with eight sticks and six joint angles) to represent human body structure, which was extracted from body contour by determining body points. Gait features based on motion parameters were calculated from a sequence of stick figures, which were input into SVM classifiers for gender recognition. Their system produced average recognition performance of 96%. More recently, Li et al. [27,28] investigated gait-based gender recognition by segmenting human silhouettes into seven components, namely, head, arm, trunk, thigh, front-leg, back-leg, and feet. By adopting averaged gait images, the individual components and a number of combinations of components were studied for gender classification on a data set of 122 individuals (85 males and 37 females). Their extensive experiments demonstrate that the trunk and front-leg components are important for gender discrimination.

3. Gender recognition from gaits

3.1. Gait representation

Lee and Grimson [26] only considered dynamic features for gender representation. In our work, we investigate structural features and dynamic features of gaits for gender recognition, by adopting gait energy image (GEI) [16], a recently proposed spatio-temporal compact representation of gaits. GEI has been demonstrated to be effective for representing gaits in the human identification problem [16,45].

Using background substraction techniques, the walking subjects can be extracted from the original image sequences to derive binary silhouette image sequences. To make the gait representation insensitive to the distance between the camera and the subject, we perform silhouette preprocessing procedure including size normalization and horizontal alignment [16]. Some examples of normalized and aligned silhouette images are shown in Fig. 2. The entire human gait sequence can be divided into cycles as human walking repeats at a stable frequency. We decide the gait cycles by counting the number of foreground pixels in the bottom half of the silhouette [33], and the two consecutive strides in the variation of the number constitute a gait cycle.

Given the preprocessed binary silhouette image $B_t(x,y)$ at time t in a sequence, the GEI is defined as follows:

$$G(x,y) = \frac{1}{N}\sum_{t=1}^{N} B_t(x,y) \tag{1}$$

where N is the number of frames in the complete cycle(s) of a silhouette sequence, t is the frame number of the sequence, and x and y are values in the 2D image coordinate (see Fig. 2 for an example of GEI). GEI reflects shapes of silhouette and their changes over the gait cycle, and it is not sensitive to incidental silhouette errors in individual frames.

3.2. Gender classification: SVMs

A previous successful technique for gender classification is SVM [31,26,40], so we adopt SVM classifiers here. SVM is an optimal discriminant method based on the Bayesian learning theory. For the cases where it is difficult to estimate the density model in high-dimensional space, the discriminant approach is preferable to the generative approach. SVM performs an implicit mapping of data into a higher dimensional feature space, and then finds a linear separating hyperplane with the maximal margin to separate data in this higher dimensional space.

Given a training set of labeled examples $\{(x_i, y_i),\ i = 1, \ldots, l\}$ where $x_i \in R^n$ and $y_i \in \{1, -1\}$, a new test example x is classified by the following function:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b\right), \tag{2}$$

where α_i are Lagrange multipliers of a dual optimization problem that describe the separating hyperplane, $K(\cdot,\cdot)$ is a kernel function, and b is the threshold parameter of the hyperplane. The training sample x_i with $\alpha_i > 0$ is called the *support vector*, and SVM finds the hyperplane that maximizes the distance between the support vectors and the hyperplane. Given a non-linear mapping Φ that embeds the input data into the high-dimensional space, kernels have the form of $K(x_i, x_j) = \langle \Phi(x_i) \cdot \Phi(x_j) \rangle$. SVM allows domain-specific selection of the kernel function. Though new kernels are being proposed, the most commonly used kernel functions are the linear, polynomial, and radial basis function (RBF) kernels.

4. Fusing gaits and faces for gender recognition

Each modality, gait or face, has its inherent weakness and limitations. Fusing gait and face cues in image sequences is a potential way to accomplish effective gender discrimination. In this study, we further present a multimodal gender recognition system by fusing gaits and faces.

Recently several attempts [34,22,45] have been made to integrate face and gait cues for the human identification problem. Shakhnarovich and Darrell [34] computed an image-based visual hull from a set of monocular views which is used to render virtual canonical views for frontal face recognition and side-view gait

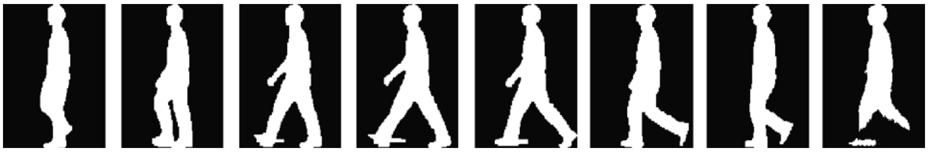


Fig. 2. Examples of normalized and aligned silhouette images. The rightmost image is the corresponding GEI.

recognition. Zhou and Bhanu [45] more recently combined side face and gait cues for human identification. All these existing studies have focused on the decision-level fusion of face and gait, while the feature-level fusion is understudied. This is mainly because the two modalities may have incompatible feature sets and the relationship between the different feature spaces is unknown. Here we propose to fuse face and gait cues at the feature level using CCA. Our motivation is that, as face and gait are two sets of measurements for human gender, conceptually the two modalities are correlated, and their relationship can be established using CCA. CCA derives a semantic "gender" space, in which the gait features and face features are compatible and can be effectively fused.

4.1. *Canonical correlation analysis*

CCA is a statistical technique developed by Hotelling [19] for measuring linear relationships between two multidimensional variables. It finds pairs of base vectors (i.e., canonical factors) for two variables such that the correlations between the projections of the variables onto these canonical factors are mutually maximized. Recently CCA has been applied to computer vision and pattern recognition problems [2,30,17,23,11]. Borga [2] adopted CCA to find corresponding points in stereo images. Melzer et al. [30] applied CCA to model the relation between an object's poses with raw brightness images for appearance-based 3D pose estimation. Harsoon et al. [17] presented a method using CCA to learn a semantic representation to web images and their associated text.

Given two zero-mean random variables $\mathbf{x} \in R^m$ and $\mathbf{y} \in R^n$, CCA finds pairs of directions $\mathbf{w}_x$ and $\mathbf{w}_y$ that maximize the correlation between the projections $x = \mathbf{w}_x^{\mathrm{T}}\mathbf{x}$ and $y = \mathbf{w}_y^{\mathrm{T}}\mathbf{y}$. The projections x and y are called *canonical variates*. More formally, CCA maximizes the function:

$$\rho = \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} = \frac{E[\mathbf{w}_x^{\mathrm{T}}\mathbf{x}\mathbf{y}^{\mathrm{T}}\mathbf{w}_y]}{\sqrt{E[\mathbf{w}_x^{\mathrm{T}}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{w}_x]E[\mathbf{w}_y^{\mathrm{T}}\mathbf{y}\mathbf{y}^{\mathrm{T}}\mathbf{w}_y]}} = \frac{\mathbf{w}_x^{\mathrm{T}}\mathbf{C}_{xy}\mathbf{w}_y}{\sqrt{\mathbf{w}_x^{\mathrm{T}}\mathbf{C}_{xx}\mathbf{w}_x\mathbf{w}_y^{\mathrm{T}}\mathbf{C}_{yy}\mathbf{w}_y}}, \tag{3}$$

where $\mathbf{C}_{xx} \in R^{m\times m}$ and $\mathbf{C}_{yy} \in R^{n\times n}$ are the *within-set covariance matrices* of $\mathbf{x}$ and $\mathbf{y}$, respectively, while $\mathbf{C}_{xy} \in R^{m\times n}$ denotes their *between-sets covariance matrix*. A number of at most $k = \min(m,n)$ canonical factor pairs $\langle \mathbf{w}_x^i, \mathbf{w}_y^i \rangle$, $i = 1,\ldots,k$ can be obtained by successively solving $\arg\max_{\mathbf{w}_x^i,\mathbf{w}_y^i}\{\rho\}$ subject to $\rho(\mathbf{w}_x^j, \mathbf{w}_x^i) = \rho(\mathbf{w}_y^j, \mathbf{w}_y^i) = 0$ for $j = 1,\ldots,i-1$, i.e., the next pair of $\langle \mathbf{w}_x, \mathbf{w}_y \rangle$ are orthogonal to the previous ones. Apparently the canonical variates x_i and y_i (corresponding to $\mathbf{w}_x^i$ and $\mathbf{w}_y^i$) are uncorrelated with the previous pairs x_j and y_j, $j = 1,\ldots,i-1$.

The maximization problem can be solved by setting the derivatives of Eq. (3), with respect to $\mathbf{w}_x$ and $\mathbf{w}_y$, equal to zero, resulting in the eigenvalue equations as

$$\begin{cases} \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{w}_x = \rho^2\mathbf{w}_x \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{w}_y = \rho^2\mathbf{w}_y \end{cases}. \tag{4}$$

Matrix inversions need to be performed in Eq. (4), leading to numerical instability if $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are rank deficient. Alternatively, $\mathbf{w}_x$ and $\mathbf{w}_y$ can be obtained by computing principal angles, as CCA is the statistical interpretation of principal angles between two linear subspace [13] (see [23] for details).

Like PCA and LDA, CCA also reduces the dimensionality of the original variables, since only a few factor pairs are normally needed to represent the relevant information. However, they serve different purposes: whilst PCA aims to minimize the reconstruction error, and LDA derives a discriminant function that maximizes between-class scatter and minimize within-class scatter, CCA seeks directions for two sets of variables to maximize their correlations.

4.2. *Feature fusion of gait and face*

Given $G = \{\mathbf{x}|\mathbf{x} \in R^m\}$ and $F = \{\mathbf{y}|\mathbf{y} \in R^n\}$, where $\mathbf{x}$ and $\mathbf{y}$ are the feature vectors extracted from gaits and faces, respectively, we apply CCA to establish the relationship between $\mathbf{x}$ and $\mathbf{y}$. Suppose $\langle \mathbf{w}_x^i, \mathbf{w}_y^i \rangle$, $i = 1,\ldots,k$ are the canonical factor pairs obtained, we can use d $(1 \leqslant d \leqslant k)$ factor pairs to represent the correlation information. With $\mathbf{W}_x = [\mathbf{w}_x^1,\ldots,\mathbf{w}_x^d]$ and $\mathbf{W}_y = [\mathbf{w}_y^1,\ldots,\mathbf{w}_y^d]$, we project the original feature vectors as $\mathbf{x}' = \mathbf{W}_x^{\mathrm{T}}\mathbf{x} = [x_1,\ldots,x_d]^{\mathrm{T}}$ and $\mathbf{y}' = \mathbf{W}_y^{\mathrm{T}}\mathbf{y} = [y_1,\ldots,y_d]^{\mathrm{T}}$ in the lower dimensional correlation space, where x_i and y_i are uncorrelated with the previous pairs x_j and y_j, $j = 1,\ldots,i-1$. We then combine the projected feature vector $\mathbf{x}'$ and $\mathbf{y}'$ to form the new feature vector as

$$\mathbf{z} = \begin{pmatrix} \mathbf{x}' \\ \mathbf{y}' \end{pmatrix} = \begin{pmatrix} \mathbf{W}_x^{\mathrm{T}}\mathbf{x} \\ \mathbf{W}_y^{\mathrm{T}}\mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{W}_x & 0 \\ 0 & \mathbf{W}_y \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}. \tag{5}$$

This fused feature vector effectively represents the multimodal information in a joint feature space for gender discrimination.

5. Experiments

5.1. *Data*

We carried out experiments on the CASIA Gait Database (Dataset B) [44], currently one of the largest gait databases in the gait-research community. The database consists of 124 subjects aged between 20 and 30 years, of which 93 were male and 31 were female, and 123 were Asian and one was European. Each subject first walked naturally along a straight line six times, then put on his/her coat and walked twice, and finally walked twice carrying a bag (knapsack, satchel, or handbag). Each subject walked a total of 10 times in the scene (six normal + two with a coat + two with a bag). Eleven cameras were uniformly set on the left-hand side, with view angle interval of 18°, so 11 video sequences from different views were captured simultaneously for every walking scenario (see Fig. 3). There are a total of 13,640 $(124 \times 10 \times 11)$ video sequences in the database, with 2–3 gait cycles in each sequence. The frame size is 320-by-240 pixel, and the frame rate is 25 fps.

In our experiments we used video sequences from two views for gender recognition: frontal view for face cues and side view for gait cues. We selected video sequences of 119 subjects (88 male and 31 female) that are suitable for gait and face analysis. In total 2380 $(119 \times 10 \times 2)$ video sequences were used in our experiments. Compared to the small data set (24 subjects) used in the previous work [26], our study was performed on a much larger data set.

As the database was collected for human gait analysis, there was no specific consideration of face data collection. Human faces were captured in an unconstrained environment like real-world surveillance scenario. The sequences contain facial expression changes, head pose variations, hair and glasses presented in the low-resolution faces. We first adopted a AdaBoost-based face detector to detect face regions in each video sequence. Then, for simplicity, we manually labeled the three points (two eyes and the mouth) of the detected face with the best resolution in a sequence, and normalized the face as a 30-by-22 pixel thumbnail to represent the video sequence. That is, we extracted a face image

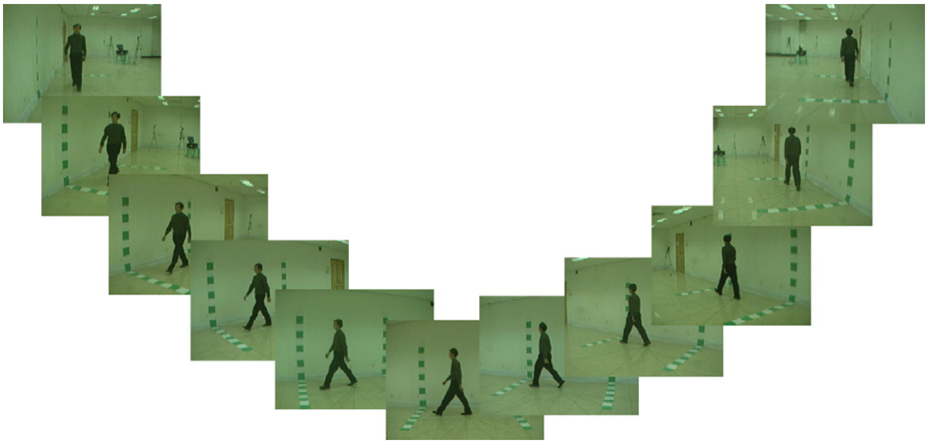


Fig. 3. The walking sequences captured from 11 different views.

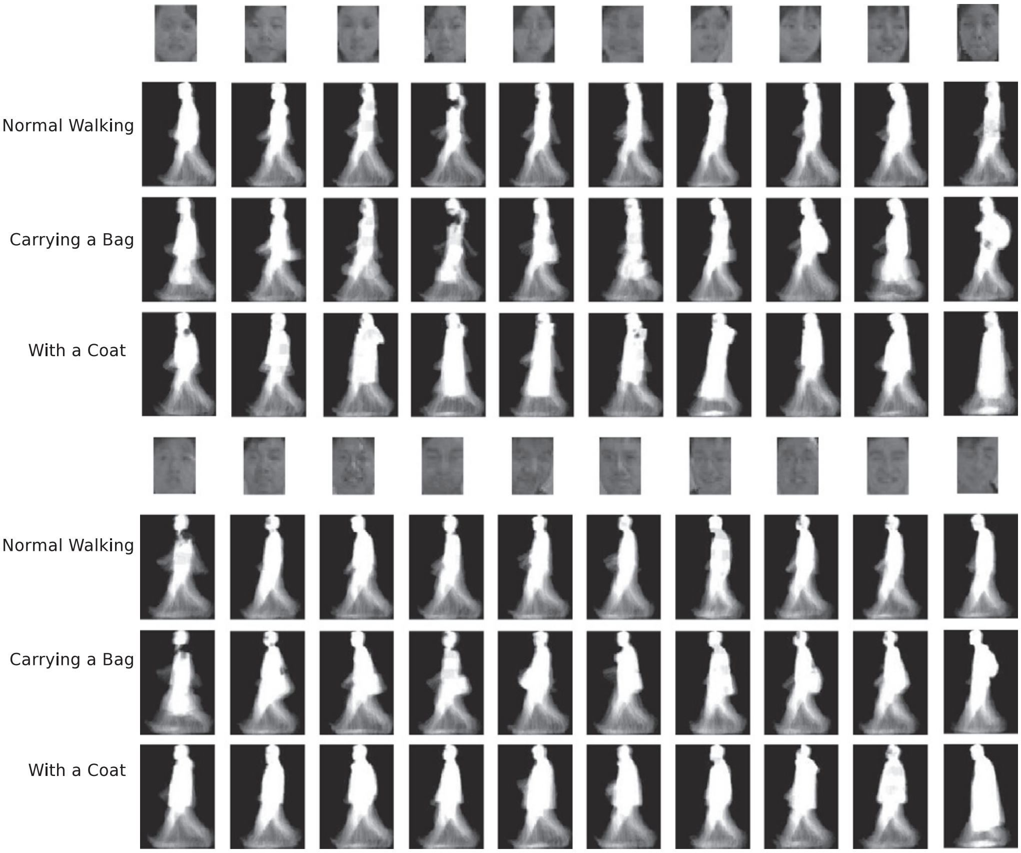


Fig. 4. The extracted face images and GEIs of 20 subjects. (Top) Female; (Bottom) male.

for each video sequence. Video-based facial gender classification is a subject of our future research. To derive gait data, we computed the GEI for each video sequence. We show the processed face images and GEIs of 20 subjects (10 female + 10 male) in Fig. 4, where the first row of GEIs are normal walking, and the second row is carrying a bag, while the bottom row is with wearing his/her coat.

5.2. *Gender recognition from gaits*

To evaluate the algorithms' generalization ability, we adopted a 5-fold cross-validation test scheme in all recognition experiments. That is, we divided the data set randomly into five groups with roughly equal (female and male) subjects, and then used the data from four groups for training and the left group for testing; the process was repeated five times for each group in turn to be tested. We report the average recognition rates (with the standard deviation) here. In all experiments, we set the soft margin C value of SVMs to infinity so that no training error was allowed. Meanwhile, each training and testing vector was scaled to be between -1 and 1. With regard to the hyperparameter selection of polynomial and RBF kernels, as suggested in [20], we carried out grid-search on the kernel parameters in the 5-fold cross-validation. The parameter setting producing the best cross-validation accuracy was picked. We used the SVM implementation

Table 1
Experimental results of gait-based gender recognition

Classifier	Recognition rates		
	Overall (%)	Male (%)	Female (%)
SVM (linear/polynomial)	94.2 ± 2.1	97.5 ± 3.2	84.7 ± 10.4
SVM (RBF)	93.6 ± 2.3	96.8 ± 3.9	84.4 ± 10.7
PCA + LDA	94.5 ± 1.9	98.0 ± 2.4	84.6 ± 9.6

Table 2
Experimental results of face-based gender recognition

Classifier	Recognition rates		
	Overall (%)	Male (%)	Female (%)
SVM (linear/polynomial)	87.5 ± 1.8	92.3 ± 2.1	74.3 ± 10.3
SVM (RBF)	90.4 ± 1.8	96.0 ± 2.1	74.6 ± 9.7
PCA + LDA	76.2 ± 1.8	79.6 ± 3.5	66.2 ± 7.7

in the publicly available machine learning library SPIDER[1] in our experiments.

We report the results of gait-based gender recognition in Table 1. It is observed that GEIs-based SVMs produce high overall recognition rates (93–94%), and the linear kernel and the (1st degree) polynomial kernel provide the same performance, slightly better than the RBF kernel. The number of support vectors of SVMs with different kernels were 13–16% of the total number of training samples. It is indicated that, for the GEI-based gait representation, the linear decision surface is able to effectively classify gender, although there are many variations in GEIs due to wearing a coat or carrying a bag (as shown in Fig. 4). To verify this, we further performed experiments with the linear subspace method PCA + LDA, which has frequently been used for the appearance-based object recognition. PCA reduces the dimension of feature space, and LDA identifies the most discriminant features. A nearest-neighbor classifier was used in our experiments. The experimental results summarized in Table 1 show that PCA + LDA achieves similar performance to the linear/polynomial kernels. Therefore, GEI is an effective gait representation for gender recognition, based on which the linear decision surface can discriminate gender with high confidence. The performance of GEI is also much better than that of dynamic features (84.5%) used in [26].

5.3. Gender recognition from faces

Before fusing gait and face modalities, we first performed gender recognition with faces, and report the results in Table 2. By comparing Tables 1 and 2, we can see that recognition results based on faces alone were consistently inferior to that based on gaits, which indicates that it is hard to learn human gender from low-resolution faces captured in unconstrained environments. For face-based gender recognition, SVMs have a clear margin of superiority over the linear subspace method PCA + LDA; the polynomial kernel also achieved the same performance with the linear kernel, but RBF kernel was found to perform best. The results we obtained reinforce the findings reported in [31]. This indicates that the face data can be better gender classified by the non-linear decision surfaces. The number of support vectors of the linear/polynomial kernels were 23–24% of the total number of training samples, while the RBF kernel employed 25–39%. The SVMs' performance of 87–90% we obtained is inferior to that reported in [31]. This is because our face data were captured in an unconstrained real-world scenario, with the presence of facial expression changes, head pose variations, various hair styles and glasses, so it is more complex than the face images of FERET database used in [31].

5.4. Gender recognition from gaits and faces

We then fused gait and face cues at the feature level using CCA. Different numbers of CCA factor pairs can be used to project the original gait and face feature vectors to a lower dimensional CCA feature space, and the recognition performance varies with the dimensionality of the projected CCA features. We first tested SVM (linear) with the CCA features of different dimensions. We plot in Fig. 5 the average recognition rates of SVM (linear) versus CCA dimensionality reduction. It is observed that the projected CCA features of gaits and faces with 90-dimension provide the best performance. Hence we carried out subsequent experiments with CCA features of 90-dimension.

To verify its effectiveness, we compared the presented CCA feature fusion with another three feature fusion methods: (1) direct feature fusion, that is, concatenating the original gait and face feature vectors to derive a single feature vector; (2) PCA feature fusion: the original gait and face feature vectors are first projected to the PCA space, respectively, and then the PCA features are concatenated to form the single feature vector. In our experiments, all principle components were kept; (3) PCA + LDA feature fusion: for each modality, the derived PCA features are further projected to the discriminant LDA space; the LDA features are then combined to derive the single feature vector. We report the experimental results of different feature fusion schemes in Table 3, where it shows the linear kernel also achieves the same performance as the polynomial kernel. We also plot bar graphs of the recognition performance in Fig. 6. We can see that the direct feature fusion and PCA + LDA feature fusion outperform slightly the single modality, while the PCA feature fusion provides the performance that is better than that of face cues but inferior to that of gait cues. In contrast, our proposed CCA feature fusion consistently achieves the best recognition results, producing considerable performance improvement over the single modality. This is because CCA captures the relationship between the feature

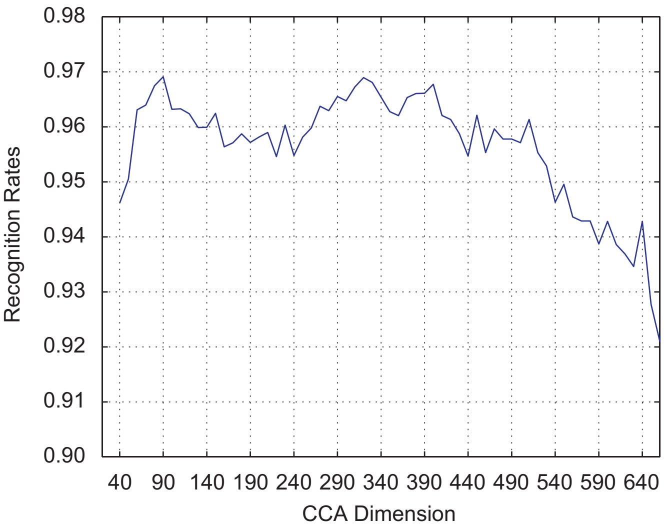


Fig. 5. Recognition rates of SVM (linear) versus dimensionality reduction of CCA.

[1] http://www.kyb.tuebingen.mpg.de/bs/people/spider/index.html.

Table 3
Experimental results of gender recognition by fusing gaits and faces

		Recognition rates			Feature dimension
		Overall (%)	Male (%)	Female (%)	
Direct fusion	SVM (linear/polynomial)	95.6 ± 1.7	98.3 ± 2.4	88.0 ± 8.6	4160
	SVM (RBF)	94.5 ± 1.8	97.4 ± 3.1	86.3 ± 8.5	
CCA fusion	SVM (linear/polynomial)	96.9 ± 1.1	99.0 ± 1.1	91.0 ± 5.2	180
	SVM (RBF)	97.2 ± 0.8	99.0 ± 1.3	92.0 ± 4.6	
PCA fusion	SVM (linear/polynomial)	92.3 ± 0.9	94.6 ± 1.9	85.6 ± 6.9	1600
	SVM (RBF)	92.5 ± 1.3	95.8 ± 1.1	83.1 ± 7.1	
PCA + LDA fusion	SVM (linear/polynomial)	95.6 ± 1.9	98.3 ± 1.7	87.9 ± 8.0	2
	SVM (RBF)	95.6 ± 1.9	98.3 ± 1.7	87.9 ± 8.0	

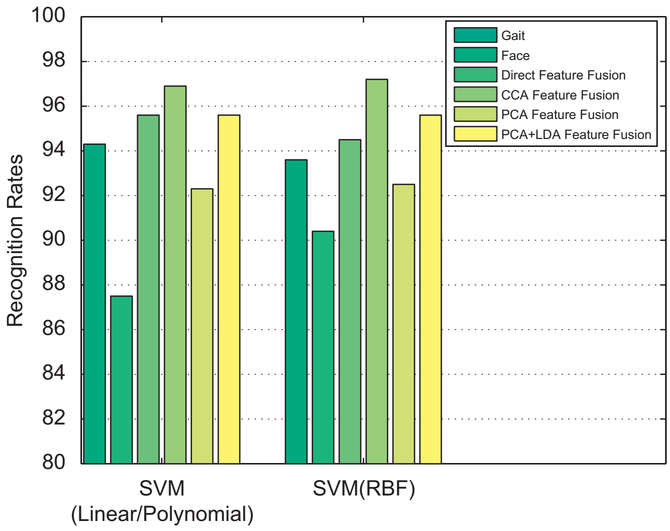


Fig.~6. Gender recognition using different features.

sets in different modalities, and the fused CCA features effectively represent information in each modality, removing noisy and redundant data. More crucially, the CCA feature fusion bring significant time and space benefit, for example, compared to the high dimensionality (4160) in the direct feature fusion. The compact 180-dimension CCA features reduce the memory space by order 23. Another strength of the CCA feature fusion is that it always produces the smallest standard deviation of cross-validation, which demonstrate it is more robust than each single modality and other feature fusion schemes. The performance 97.2% that the CCA feature fusion-based SVM (RBF) obtained is better than 96.6% reported in [31], and, to our best knowledge, is the best gender recognition performance reported so far in the published literature. A supplementary *video demonstration* is available at http://www.dcs.qmul.ac.uk/~cfshan/research/gender.html.

We note that, in Tables 1–3, all the female recognition rates are poorer than the male (with larger variance). In previous studies [31,35], different classifiers also had higher error rates in classifying females. This phenomenon is possibly because the female gaits and faces have less prominent and distinct features, for example, the female has much variation in their hair styles and clothing. Another possible reason is the unbalance data set (88 male and 31 female) in our experiments. An encouraging observation is the female recognition performance based on each single modality is improved much by the CCA feature fusion (from 74–84% to 91–92%) which is significant.

In the above experiments, the face images were manually aligned. To investigate the effect of the misalignment of faces on final fusion results, we further carried out experiments by taking the face images directly from the face detector. Due to the unaligned face, the recognition performance of different fusion methods degrades to 85–91%, although the CCA feature fusion still provides the best performance.

6. Conclusions

In this paper, we investigate an important but understudied problem, gender classification from human gaits, which has important applications in intelligent visual surveillance and human–computer interaction. Our extensive experiments demonstrate that visual gender recognition from human gaits is very effective. Considering each modality in isolation has its limitations, we also propose a method to effectively fuse gait and face at the feature level for improved gender discrimination. Experiments demonstrate that our multimodal gender recognition system achieves the superior recognition performance of 97.2% in large data sets.

References

[1] C.D. Barclay, J.E. Cutting, L.T. Kozlowski, Temporal and spatial actors in gait perception that influence gender recognition, Percept. Psychophys. 23 (2) (1978) 145–152.

[2] M. Borga, Learning multidimensional signal processing, Ph.D. Thesis, Dissertation No. 531, Linkoping University, Linkoping, Sweden, 1998.

[3] V. Bruce, A.M. Burton, N. Dench, E. Hanna, P. Healey, O. Mason, A. Coombes, R. Fright, A. Linney, Sex discrimination: How do we tell the difference between male and female faces, Perception 22 (1993) 131–152.

[4] R. Brunelli, T. Poggio, Hyperbf networks for gender classification, in: DRAPA Image Understanding Workshop, 1992, pp. 311–314.

[5] N.P. Costen, M. Brown, S. Akamastu, Sparse models for gender classification, in: IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2004.

[6] J.E. Cutting, D.R. Proffitt, L.T. Kozlowski, A biomechanical invariant for gait perception, J. Exp. Psychol. Human Percept. Perform. 4 (3) (1978) 357–372.

[7] J.W. Davis, H. Gao, Recognizing human action efforts: an adaptive three-mode PCA framework, in: IEEE International Conference on Computer Vision (ICCV), 2003, pp. 1463–1469.

[8] J.W. Davis, H. Gao, An expressive three-mode principal components model of human action style, Image Vision Comput. 21 (11) (2003) 1001–1016.

[9] J.W. Davis, H. Gao, Gender recognition from walking movements using adaptive three-mode PCA, in: IEEE CVPR Workshop on Articulated and Nonrigid Motion, 2004, p. 9.

[10] J.W. Davis, H. Gao, An expressive three-mode principal components model for gender recognition, J. Vision 4 (5) (2004) 362–377.

[11] R. Donner, M. Reiter, G. Langs, P. Peloscheck, H. Bischof, Fast active appearance model search using canonical correlation analysis, IEEE Trans. Pattern Anal. Mach. Intell. 28 (10) (2006) 1690–1694.

[12] B.A. Golomb, D.T. Lawrence, T.J. Sejnowski, Sexnet: a neural network identifies sex from human faces, in: Advances in Neural Information Processing Systems (NIPS), 1991, pp. 572–577.

[13] G.H. Golub, H. Zha, The canonical correlations of matrix pairs and their numerical computation, Technical Report, Stanford, CA, USA, 1992.

[14] A.B.A. Graf, F.A. Wichmann, Gender classification of human faces, in: International Workshop on Biologically Motivated Computer Vision, 2002, pp. 491–500.

[15] S. Gutta, H. Wechsler, P.J. Phillips, Gender and ethnic classification, in: IEEE International Conference on Automatic Face & Gesture Recognition (FG), 1998, pp. 194–199.

[16] J. Han, B. Bhanu, Individual recognition using gait energy image, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2) (2006) 316–322.

[17] D. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis; an overview with application to learning methods, Neural Comput. 16 (12) (2004) 2639–2664.

[18] S. Hirashima, Recognition on the gender of point-light walkers moving in different directions, Jpn. J. Psychol. 70 (2) (1999) 149–153.

[19] H. Hotelling, Relations between two sets of variates, Biometrika 8 (1936) 321–377.

[20] C.-W. Hsu, C.-C. Chang, C.-J. Lin, A practical guide to support vector classification, Technical Report, Taipei, 2003.

[21] A. Jain, J. Huang, Integrating independent component analysis and linear discriminant analysis for gender classification, in: IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2004.

[22] A. Kale, A.K.R. Chowdhury, R. Chellappa, Fusion of gait and face for human identification, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2004, pp. 901–904.

[23] T.-K. Kim, J. Kittler, R. Cipolla, Learning discriminative canonical correlations for object recognition with image sets, in: European Conference on Computer Vision (ECCV), 2006, pp. 251–262.

[24] L.T. Kozlowski, J.E. Cutting, Recognizing the sex of a walker from dynamic point-light display, Percept. Psychophys. 21 (6) (1977) 575–580.

[25] L. Lee, Gait analysis for classification, Technical Report 2003-014, MIT AI Lab, June 2003.

[26] L. Lee, W.E.L. Grimson, Gait analysis for recognition and classification, in: IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2002, pp. 155–162.

[27] X. Li, S.J. Maybank, D. Tao, Gender recognition based on local body motions, in: IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2007.

[28] X. Li, S.J. Maybank, S. Yan, D. Tao, D. Xu, Gait components and their applications to gender recognition, IEEE Trans. Syst. Man Cybern. Part C 38 (2) (2008) 145–155.

[29] G. Mather, L. Murdoch, Gender discrimination in biological motion displays based on dynamic cues, Proc. R. Soc.: Biol. Sci. 258 (1353) (1994) 273–279.

[30] T. Melzer, M. Reiter, H. Bischof, Appearance models based on kernel canonical correlation analysis, Pattern Recognition 39 (9) (2003) 1961–1973.

[31] B. Moghaddam, M. Yang, Learning gender with support faces, IEEE Trans. Pattern Anal. Mach. Intell. 24 (5) (2002) 707–711.

[32] A.J. O'Toole, T. Vetter, N.F. Troje, H.H. Bulthoff, Sex classification is better with three-dimensional structure than with image intensity information, Perception 26 (1997) 75–84.

[33] S. Sarkar, P.J. Phillips, Z. Liu, I.R. Vega, P. Grother, K.W. Bowyer, The HumanID gait challenge problem: data sets, performance, and analysis, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2) (2005) 162–177.

[34] G. Shakhnarovich, T. Darrell, On probabilistic combination of face and gait cues for identification, in: IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2002, pp. 169–174.

[35] G. Shakhnarovich, P.A. Viola, B. Moghaddam, A unified learning framework for real time face detection and classification, in: IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2002, pp. 14–21.

[36] Z. Sun, G. Bebis, X. Yuan, S.J. Louis, Genetic feature subset selection for gender classification: a comparison study, in: IEEE Workshop on Application of Computer Vision, 2002.

[37] S. Tamura, H. Kawai, H. Mitsumoto, Male/female identification from 8 × 6 very low resolution face images by neural network, Pattern Recognition 29 (2) (1996) 331–335.

[38] N.F. Troje, Decomposing biological motion: a framework for analysis and synthesis of human gait patterns, J. Vision 2 (5) (2002) 371–387.

[39] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2001, pp. 511–518.

[40] L. Walawalkar, M. Yeasin, A. Narasimhamurthy, R. Sharma, Support vector learning for gender classification using audio and visual cues, Int. J. Pattern Recognition Artif. Intell. 17 (3) (2003) 417–439.

[41] F.A. Wichmann, A.B.A. Graf, E.P. Simoncelli, H.H. Bulthoff, B. Scholkopf, Machine learning applied to perception: decision-images for gender classification, in: Advances in Neural Information Processing Systems (NIPS), 2004.

[42] B. Wu, H. Ai, C. Huang, S. Lao, LUT-based Adaboost for gender classification, in: International Conference on Audio and Video-Based Person Authentication (AVBPA), 2003.

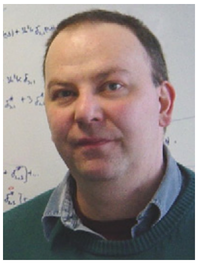
[43] J.-H. Yoo, D. Hwang, M.S. Nixon, Gender classification in human gait using support vector machine, in: Proceedings of Advanced Concepts for Intelligent Vision Systems, 2005, pp. 138–145.

[44] S. Yu, D. Tan, T. Tan, A framework for evaluating the effect of view angle clothing and carrying condition on gait recognition, in: International Conference on Pattern Recognition (ICPR), 2006, pp. 441–444.

[45] X. Zhou, B. Bhanu, Integrating face and gait for human recognition, in: IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2006, p. 55.

Caifeng Shan received the B.Eng. degree in computer science from the University of Science and Technology of China (USTC), Hefei, China in 2001, the M.Eng. degree in Pattern Recognition and Intelligent System from National Laboratory of Pattern Recognition, Chinese Academy of Sciences (CAS), Beijing, China in 2004, and the Ph.D. degree in computer science from Queen Mary, University of London, London, UK in 2007. Currently he is a Research Scientist at Philips Research. He was awarded the Queen Mary Research Studentship (2004–2007). His research interests include computer vision, pattern recognition, and image/video processing.

Shaogang Gong is Professor of Visual Computation at Queen Mary, University of London, elected a Fellow of the Institution of Electrical Engineers and a Member of the UK Computing Research Committee. He heads both Queen Mary Computer Vision Research Group and Queen Mary Mathematics, Electronic Engineering and Computer Science Interdisciplinary Consortium on Digital Media and Complexity Sciences. He received his D.Phil. in computer vision from Oxford University in 1989 with a thesis on the computation of optic flow using second-order geometric analysis. He was a recipient of a Queen's Research Scientist Award in 1987, a Royal Society Research Fellow in 1987 and 1988, and a GEC-Oxford fellow in 1989. He twice won the Best Science Prize of the British Machine Vision Conferences (1999, 2001) and once won the Best Paper Award (2001) of the IEEE International Workshops on Recognition, Analysis and Tracking of Faces and Gestures. He has published over 160 papers in computer vision and machine learning, and a book on Dynamic Vision: From Images to Face Recognition. His work focuses on the detection, tracking and recognition of motion objects; video based face and expression recognition; gesture recognition for visually mediated interaction, video behaviour profiling, recognition and abnormality detection.



Peter McOwan is Professor of Computer Science at Queen Mary, University of London. He was born in the UK in 1962. He received the B.Sc. degree in Physics from Edinburgh University, Edinburgh, UK in 1984, the M.Sc. degree in Medical Physics from Aberdeen University, Aberdeen, UK in 1985, the Ph.D. degree in computer generated holography from King's College London, London, UK in 1990, and the M.Sc. degree in experimental methods in psychology from University College London in 1995. After holding a Wellcome Trust Mathematical Biology fellowship at University College London, he was a Lecturer in the Department of Cybernetics, University of Reading, Reading, UK, between 1996 and 1997, a Lecturer in the Department of Mathematics and Computer Science at Goldsmiths College, University of London, between 1998 and 1999, and has since been with the Department of Computer Science, Queen Mary College, London, where he currently is Director of Teaching. His research interests are in the area of biologically inspired computing and cognitive science. He is a member of the IEE, the Institute of Physics, and the Optical Society of America.