Gender recognition from facial images: 2D or 3D?

WENHAO ZHANG, [*] **MELVYN L. SMITH, LYNDON N. SMITH, ABDUL FAROOQ**

Centre for Machine Vision, Bristol Robotics Laboratory, University of the West of England, T Block, Frenchay Campus, Coldharbour Lane, Bristol, BS16 1QY, UK
**Corresponding author: wenhao.zhang@uwe.ac.uk*

Received XX Month XXXX; revised XX Month, XXXX; accepted XX Month XXXX; posted XX Month XXXX (Doc. ID XXXXX); published XX Month XXXX

This paper seeks to compare encoded features from both 2D and 3D face images in order to achieve automatic gender recognition with high accuracy and robustness. The Fisher Vector encoding method is employed to produce 2D, 3D and fused features with escalated discriminative power. For 3D face analysis, a two-source Photometric Stereo (PS) method is introduced that enables 3D surface reconstructions with accurate details as well as desirable efficiency. Moreover, a 2D+3D imaging device, taking the two-source PS method as its core, has been developed that can simultaneously gather colour images for 2D evaluations and PS images for 3D analysis. This system inherits the superior reconstruction accuracy from the standard (3 or more light) PS method, but simplifies the reconstruction algorithm as well as the hardware design by only requiring two light sources. It also offers great potential for facilitating human computer interaction by being accurate, cheap, efficient and non-intrusive.

10 types of low-level 2D and 3D features have been experimented with and encoded for Fisher Vector gender recognition. Evaluations of the Fisher Vector encoding method have been performed on the FERET database, Colour FERET database, LFW database and FRGCv2 database, yielding 97.7%, 98.0%, 92.5% and 96.7% accuracy, respectively. In addition, the comparison of 2D and 3D features has been drawn from a self-collected dataset, which is constructed with the aid of the 2D+3D imaging device in a series of data capture experiments. With a variety of experiments and evaluations, it can be proved that the Fisher Vector encoding method outperforms most state-of-the-art gender recognition methods. It has also been observed that 3D features reconstructed by the two-source PS method are able to further boost the Fisher Vector gender recognition performance, i.e. up to a 6% increase on the self-collected database.

OCIS codes: *(100.5010) Pattern recognition; (110.2960) Image analysis; (100.3010) Image reconstruction techniques; (100.6890) Three-dimensional image processing; (110.6880) Three-dimensional image acquisition.*

1. INTRODUCTION

Gender has long been considered more than a matter of different biological or physical characteristics. It is also linked to certain social attributes and behavioural patterns in nature [1]. Being able to automatically recognise gender from facial images has become one of the main focuses in the field of Human Computer Interaction (HCI). Its significance is also justified by its contribution to other areas such as visual surveillance [2], data retrieval [3], directed advertising and marketing by providing demographic statistics.

Since a few decades ago, a variety of 2D facial features have been extracted, fused or encoded to achieve automatic gender recognition, but only recent years have witnessed the emergence of 3D imaging systems that are intended to obtain 3D facial features to better fulfil the recognition task [4]. As far as gender recognition is concerned, most techniques can be broadly classified into geometric [5] or appearance [6] based methods. The former aims to characterise anthropometric measures such as face width and length, distance between the eyes, or face boundaries extracted by the Active Shape Model (ASM) [7], while the latter describes the textures of facial skin where wrinkles, bulges, and furrows are present. While 2D features continue to draw extensive attention from researchers and lead to a wide range of applications, their inherent limitations are inevitable and are deemed volatile due to dynamic environmental and experimental factors such as illumination condition and head pose. These limitations should be resolved before gender recognition could become well suited for real-world implementations. A viable solution is facilitated by the increased availability and easier accessibility of 3D imaging devices. 3D facial features have reportedly contributed to higher accuracy and/or robustness in automatic face recognition, gesture recognition, facial expression recognition and gender recognition. These features can be extracted from 3D faces obtained via various means including structured-light [8] 3D scanning, laser scanning [9], stereoscopic systems, photometric systems, etc. While some 3D imaging systems are mechanically complex and expensive and other are computationally expensive, photometric vision systems manifest better feasibility for real-world applications in that 1) they require only inexpensive and simple settings, 2) they are capable of performing real-time 3D reconstruction [10] and that 3) they provide superior reconstruction results that reveal 3D textures. Photometric stereo methods are the core of photometric vision systems, which recover surface normals and albedos of a surface from three or more images where illumination directions vary but viewpoint remains fixed.

In this paper, we propose to encode low-level facial features as Fisher Vectors (FVs) for gender recognition and draw a comparison between utilising a number of 2D and 3D feature types. We first prove with two sets of experiments that by encoding 2D features, our method exhibits high accuracy and robustness in both controlled and uncontrolled environments. We then introduce a variation of PS that requires only two light sources to recover a 3D surface and present our 2D+3D imaging system based on this approach. Finally, we prove with the data gathered by our imaging system that 3D facial features can provide a further boost to the proposed method.

In summary, the main contribution of this paper is threefold.

1) A gender recognition method utilising the **FV encoding** approach – a generic method that can encode almost any type of features but still maintains low complexity in implementation. It also proves to have high accuracy and robustness against head poses. To the best of our knowledge, this is the first time that FVs have been employed for gender recognition.

2) A **two-source PS** method that can be applied to real-world scenarios for 3D surface reconstruction and a **2D+3D imaging system**. This method lowers the complexity of conventional PS methods by reducing the number of light sources (i.e. images per reconstruction). This will largely reduce the mechanical complexity in designing 3D imaging systems while improving real-time performance.

3) A comprehensive comparison of utilising 2D and 3D facial features for gender recognition by evaluating **10 types** of encoded low-level features. The superiority of our method and 3D features is manifested by our experiments on **4 publicly available databases** and **a self-collected database**.

2. RELATED WORK

We review in this section a number of 2D and 3D methods for gender recognition. In addition, 12 gender recognition approaches are further summarised in Table 1 with a detailed comparison to our method in terms of features/classifiers, recognition accuracy, database, validation method and limitation(s).

Gender recognition from facial images, i.e. gender classification, is a challenging task in that a face exhibits a wide range of intra-class variations due to diverse facial attributes or dynamic environmental factors. The former type of complication mainly includes age, ethnicity and makeup while the latter includes illumination condition, head pose, facial occlusion and camera quality.

Most high-performance gender recognition methods involve machine learning and follow four stages: face detection, facial image pre-processing, feature extraction and classification [11].

For the face detection stage, the Viola-Jones face detector [12] has been widely adopted due to its ease of implementation and relatively high accuracy [13]. For the image pre-processing stage, normalisation, i.e. contrast and brightness adjustment, image resizing and face alignment are commonly considered useful despite their varied implementation details. Among them, face alignment has been reported to be able to guarantee an increase in classification accuracy by a number of studies. For example, in a research [14] evaluating a number of gender classification methods, it was concluded that Support Vector Machines (SVMs) outperformed other classification methods with up to 86.54% accuracy, and that higher accuracy could be achieved by automatic face alignment methods. Another study [15] illustrated that face alignment brought an increase to classification accuracy for various methods including use of neural networks, SVMs, and AdaBoost.

For the feature extraction and selection stage, a wide range of features have been experimented with and evaluated in the literature. The need for higher robustness and discriminability has led to data capture that is 2D or 3D, and densely extracted or sparsely detected. Features include intensity values from greyscale images [16], Local Binary Patterns (LBP) [17,18], facial strips [19], Haar-like features [13], Scale Invariant Feature Transform (SIFT) features [20], etc. Depending on the type of features extracted, one or more face descriptors per image are obtained. These descriptors are commonly drawn to characterise facial texture, geometry or topology whose representations seek to obtain high robustness to intra-class variations (e.g. facial expression, head pose, illumination, etc.). For the classification stage, SVMs and neural networks have been the most popular classifiers. SVMs with different kernels were investigated in [16] and convolutional neural networks (CNNs) were adopted by [21] and [22] as gender classifiers.

Following the 4 major stages, a number of approaches have reported relatively high classification rates on publicly available datasets. A decision-fusion based method was presented by [23] that utilised multiple SVMs to classify greyscale values, LBP and histograms of edge directions as features. Then all classification results were integrated to make the final decision by means of majority voting, leading to 99.07% accuracy. However this result was obtained from a small subset of the Grey FERET database and their validation method was not sophisticated enough to reflect the performance of their approach objectively. Similarly, [19] employed 10 regression functions to conduct region-based classifications and then fed the vector of classification results into an SVM to generate the final decision. Despite the 98.8% accuracy on the Grey FERET database they reported, they did not illustrate their evaluation method and the split of training and testing data. The reappearance of the same subject(s) in both training and testing sets may account for the high accuracy they obtained. A face alignment scheme is compulsory to their approach, the absence of which leads to a 6% drop in the classification rate, bringing 98.8% down to 92.8%. [24] proposed a fusion-based method for gender recognition by integrating different facial regions using the 'matcher weighing fusion' method. The facial landmarks for segmenting the face into its sub-regions were detected by a profile-based method and a curvature based method. Interestingly they proved experimentally that the fusion of multiple facial sub-regions was superior to the complete face region alone and that the upper face contained more discrimination power regarding gender classification. Another type of facial feature, i.e. two-directional Principal Component Analysis on real Gabor space, was explored by Rai and Khanna [25,26] and achieved up to 98.4% gender classification rate on the Colour FERET database.

While features obtained from colour or greyscale images are still attracting huge research focus, the trend has now been directed toward the study of 3D face features, mainly including 3D appearance features and 3D geometric features. In [27], two types of appearance features, i.e. the LBP features and the shape index features, were fused to characterise facial textures and shapes. The resulting classification rate on the FRGCv2 dataset was up to 93.7%. As for the utilisation of geometric features, [28] performed the Random Forest algorithm on 3D morphological features and used the votes to represent the magnitude of sexual dimorphism. 97.18% accuracy was achieved on the FRGCv2 dataset. Similarly, [29] investigated 'gender strength', which was aimed at replacing the conventional binary gender labels by introducing a continuous gender class variable. Another study [24] achieved high accuracy by integrating multiple facial regions segmented by a selection of facial landmarks. 3D features extracted from these regions were then classified by a SVM. This study evaluated the contributions of individual facial regions and concluded that the upper facial region contained the highest gender discriminability. Since 3D features have shown promising results, they have also been combined with 2D features such that the merits from both types of features can be inherited. [30] introduced a type of LBP based feature descriptor to encode 3D facial features for gender classification. A combination scheme making use of both depth images and greyscale images was proposed, which showed enhanced classification accuracy on both high and low resolution data. [31] also proposed a fusion method for gender recognition that combined the shape and texture features extracted from 3D meshes and greyscale images. The fused features proved to outperform individual types.

With individual works reviewed, we draw conclusions from the literature regarding the preferences and trends in gender classification. 1) SVMs and neural networks are the most popular classifiers. 2) LBP and its variations are the most popular features. 3) Most studies use the FERET database as the standard evaluation database. 4) Most works are carried out under a well-controlled environment while real-world implementation and evaluation lack exploitation. 5) Most works incorporate facial landmark detection and face alignment in the pre-processing stage. 6) Most works employ the 5-fold cross validation for accuracy estimation. 7) Although 3D features have received an increasing amount of study, the corresponding 3D imaging systems and implementations are not keeping the pace with the algorithms. Therefore, the mainstream for gender recognition continues to be 2D methods.

3. FISHER VECTOR FOR GENDER RECOGNITION

As seen from the literature, most methods regarding gender recognition lack robustness and suffer from various limitations in real-world scenarios. To bridge these gaps, we explore novel methods that can boost the discriminative power of facial features while overcoming challenging environmental variations. In this section, we introduce a gender recognition method that utilises FVs as encoded features.

A. Fisher Vector principle

A Fisher Vector (FV) is an encoded vector that applies Fisher kernels on visual vocabularies where the visual words are represented by means of a Gaussian Mixture Model (GMM). The Fisher kernel function is derived from a generative probability model, and provides a generic mechanism that combines the advantages of generative and discriminative approaches. As a core component of a FV, a GMM is a parametric probability density function represented as a weighted sum of Gaussian component densities as given by Eq. (1) [32]

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{N} \omega_i \, g(\boldsymbol{x}|\boldsymbol{\mu_i}, \boldsymbol{\sigma_i}) \tag{1}$$

where $\boldsymbol{x}$ is a D-dimensional data vector, $\lambda = \{\omega_i, \boldsymbol{\mu_i}, \boldsymbol{\sigma_i}, i = 1,2, \dots, N\}$ is the collective representation of the GMM parameters – ω_i the mixture weights, $\boldsymbol{\mu_i}$ the mean vector and $\boldsymbol{\sigma_i}$ the covariance matrix. N is the number of Gaussians. The component $g(\boldsymbol{x}|\boldsymbol{\mu_i}, \boldsymbol{\sigma_i})$ is further described in Eq. (2).

$$g(\boldsymbol{x}|\boldsymbol{\mu_i}, \boldsymbol{\sigma_i}) = \frac{e^{\{-\frac{1}{2}(x-\mu_i)'\boldsymbol{\sigma_i}^{-1}(x-\mu_i)\}}}{(2\pi)^{D/2}|\boldsymbol{\sigma_i}|^{1/2}} \tag{2}$$

The mixture weights are subject to the constraint in Eq. (3).

$$\sum_{1}^{N} \omega_i = 1 \tag{3}$$

The covariance matrices are assumed to be diagonal since any distribution can be decomposed into a number of weighted Gaussians with diagonal covariances. Let $\boldsymbol{X} = \{\boldsymbol{x_t}, t = 1,2, \dots, T\}$ be the set of descriptors of low-level features extracted from an image, and it is assumed that all the descriptors are independent. Eq. (4) can be found:

$$\log p(\boldsymbol{X}|\lambda) = \sum_{1}^{T} \log p(\boldsymbol{x_t}|\lambda) \tag{4}$$

The descriptors $\boldsymbol{X}$ can be described by the gradient vector:

$$\psi_\lambda^{\boldsymbol{X}} = \frac{\nabla_\lambda \, logp(\boldsymbol{X}|\lambda)}{T} \tag{5}$$

A natural kernel on these gradients is:

$$\varkappa(\boldsymbol{X}, \boldsymbol{Y}) = \psi_\lambda^{\boldsymbol{X}'} \mathcal{F}_\lambda^{-1} \psi_\lambda^{\boldsymbol{Y}} \tag{6}$$

where $\mathcal{F}_\lambda = \mathcal{L}_\lambda{}'\mathcal{L}_\lambda$ is the Fisher information matrix and $\Psi_\lambda^{\boldsymbol{X}} = \mathcal{L}_\lambda \, \psi_\lambda^{\boldsymbol{X}}$ is referred to as the Fisher Vector of $\boldsymbol{X}$. Let $\gamma_t(i)$ denote the soft assignment of descriptor $\boldsymbol{x_t}$ to the Gaussian component i:

$$\gamma_t(i) = p(i|\boldsymbol{x_t}, \lambda) = \frac{\omega_i g(\boldsymbol{x_t}|\boldsymbol{\mu_i}, \boldsymbol{\sigma_i})}{\sum_{j=1}^{N} \omega_j \, g(\boldsymbol{x_t}|\boldsymbol{\mu_j}, \boldsymbol{\sigma_j})} \tag{7}$$

The gradients of Gaussian component i with respect to the mean $\boldsymbol{\mu_i}$ and the covariance $\boldsymbol{\sigma_i}$ respectively are:

$$\Psi_{\mu,i}^{\boldsymbol{X}} = \frac{1}{T\sqrt{\omega_i}} \sum_{1}^{T} \gamma_t(i) \left(\frac{\boldsymbol{x_t} - \boldsymbol{\mu_i}}{\boldsymbol{\sigma_i}}\right) \tag{8}$$

$$\Psi_{\sigma,i}^{\boldsymbol{X}} = \frac{1}{T\sqrt{2\omega_i}} \sum_{1}^{T} \gamma_t(i) \left[\frac{(\boldsymbol{x_t} - \boldsymbol{\mu_i})^2}{\boldsymbol{\sigma_i}^2} - 1\right] \tag{9}$$

Finally a FV is represented as:

$$\Phi = \{\Psi_{\mu,1}^{\boldsymbol{X}}, \Psi_{\sigma,1}^{\boldsymbol{X}}, \dots, \Psi_{\mu,N}^{\boldsymbol{X}}, \Psi_{\sigma,N}^{\boldsymbol{X}}\} \tag{10}$$

A FV is derived from GMM parameters by comparing the Gaussian distribution of one image with that of the entire training data, and then capturing the variation. As a result, a FV as a feature vector is provided with contextual definition and enhanced saliency for classification.

B. Fisher Vector encoding for gender recognition

FVs have been used for face recognition and have proved to be an excellent encoding method [33]. The FV encoding approach consists of 5 main stages:

1. Face pre-processing

The techniques experimented at this stage include face detection, image resizing, histogram equalisation and face alignment. In the proposed method, the Viola-Jones face detector [12] is used to obtain a face region in the first place. We then reshape the face region obtained by the face detector so that it incorporates the hair region and the chin. All the face regions obtained are further resized to a uniform size in the next step. We have also experimented with face alignment (using eye centres) and histogram equalisation so as to evaluate their impacts.

2. Low-level feature and face descriptor computation

We extract dense descriptors at every pixel location. Firstly, we segment a face image into a number of overlapping patches of the same size. Specifically, these patches are obtained by sliding a $ps \times ps$ window across an image horizontally and vertically by a predefined sampling step ss ($ss \in \mathbb{Z}$). One descriptor per patch rather than one descriptor per image is calculated.

3. Dimension reduction and feature selection.

As one image produces multiple descriptors, linearly increasing the total number of observations as opposed to conventional methods, dimension reduction is essential in that it cuts down memory consumption and that it potentially compresses raw features into more discriminative representations. We employ Principal Component Analysis (PCA) to reduce the dimensionality to 64 features per descriptor (see Fig. 7 for the impact of PCA). The centre position of a patch, after being scaled to [0.5, 0.5], is then appended to the end of the corresponding descriptor in the form of a 2D vector [33], increasing the dimensionality from 64 to 66.

4. FV encoding

The aggregate of all descriptors from training images trains a GMM and yields model parameters which, according to Eq. (8) – (10), lead to FVs as encoded features. In the experiments, we utilised a publicly available toolbox [34] for GMM training and SIFT feature extraction. In this stage, the decomposed patches are reunited to characterise a complete image, in the form of derivatives of all the Gaussian components. For computation of FVs, ℓ_2 normalisation and power normalisation are applied to the vectors since they have been reported to bring improved classification performance [35].

5. Classifier training

The proposed algorithm learns a SVM classifier from all training FVs. In our experiments, a linear SVM and a SVM with the Radial Basis Function (RBF) kernel were trained individually for predicting image labels. The reason for this is to demonstrate that the proposed algorithm only requires a linear SVM to achieve high accuracy.

4. TWO-SOURCE PHOTOMETRIC STEREO FOR 3D FACE RECONSTRUCTION

This section introduces a 3D face reconstruction method such that 3D features can be utilised by the proposed FV encoding method for higher accuracy and robustness. 3D facial features reveal facial topology by providing geodesic distances and surface curvatures. They have thus shown promise for bringing higher accuracy to face recognition; and improved robustness to practical applications where scenes are complex and dynamic. However, the exploitation of 3D vision is not currently sufficient to enable a wide array of 3D vision based applications. This is mainly due to 3D reconstruction techniques being isolated from 3D imaging systems which are commonly seen in bulky and expensive setups. Consequently, many algorithms struggle to find their way into real-world scenarios. In this section, we introduce a two-source variation of the photometric stereo (PS) method and, for the first time, apply it to various types of realistic data. Our aim is to provide a 3D reconstruction algorithm, together with a stereo imaging system, suitable but not restricted to the gender recognition task.

A. Photometric Stereo Principles

PS allows estimation of surface normals from reflectance maps obtained from images of the same object captured under different illumination directions. It was first introduced by [36] which illustrates that three views are sufficient to uniquely determine surface normal as well as albedo at each image point, provided that the directions of incident illumination are not collinear in azimuth. Other works employ four views for improved reconstruction performance. PS techniques are superior in capturing detailed high-frequency 3D textures and are less affected by image noise compared to triangulation based techniques [37]. In addition, PS methods normally require only one camera for image capture, simplifying the calibration process and allowing for high efficiency. In contrast, binocular stereo, for example, recovers depth of surface rather than surface orientations, which would likely introduce noise and artefacts.

Let $I_1(x,y), I_2(x,y)$ and $I_3(x,y)$ be three images captured under varied illumination directions. By varying the illumination direction, the reflectance map is changed accordingly, giving Eq. (11):

$$\begin{cases} I_1(x,y) = R_1(p,q) \\ I_2(x,y) = R_2(p,q) \\ I_3(x,y) = R_3(p,q) \end{cases} \tag{11}$$

A general reflectance map in gradient representation of the surface orientation and illumination direction is expressed in Eq. (12).

$$R(p,q) = \frac{\varrho(1+pp_s+qq_s)}{\sqrt{1+p^2+q^2}\sqrt{1+{p_s}^2+{q_s}^2}} \tag{12}$$

where ϱ is the albedo, $\vec{\boldsymbol{N}} = [-p,-q,1]$ defines the surface normal vector, and $\vec{\boldsymbol{L}} = [-p_s,-q_s,1]$ defines the illumination direction. Let the surface be $z = f(x,y)$, the gradients in the x and y directions become:

$$\begin{cases} p = -\dfrac{\partial f(x,y)}{\partial x} \\ q = -\dfrac{\partial f(x,y)}{\partial y} \end{cases} \tag{13}$$

These equations are derived under the assumptions that 1) the object size is small relative to the viewing distance; 2) the surface is Lambertian; and 3) the surface is exempt from cast-shadows or self-shadows. To simplify the expression, the light vector is further normalised to a unit vector $\overrightarrow{\boldsymbol{L_n}} = [a_x, a_y, a_z]$. The relationship between a greyscale image and a reflectance map can also be written as:

$$I(x,y) = \frac{\varrho \cdot <\vec{\boldsymbol{N}}, \overrightarrow{\boldsymbol{L_n}}>}{|\vec{\boldsymbol{N}}|} = \varrho \cdot \frac{-pa_x - qa_y + a_z}{\sqrt{1+p^2+q^2}} \tag{14}$$

From Eq. (11) to Eq. (14), it is known that with three greyscale images $I_1(x,y), I_2(x,y)$ and $I_3(x,y)$, along with three known light vectors $\overrightarrow{\boldsymbol{L_{n1}}}, \overrightarrow{\boldsymbol{L_{n2}}}$ and $\overrightarrow{\boldsymbol{L_{n3}}}$ pointing in the directions of their respective light source, the surface normal and albedo at each image point can be uniquely determined.

B. A two-source PS method

Although the standard 4-source PS method is highly accurate and relatively efficient in recovering 3D surface normals, it is however not ideal for facilitating real-world applications. Generally, its implementation is prohibited by the need for capturing at least 3 (or more commonly 4) images at high frame rate for every reconstruction. Another limitation is posed by the complex structure of data capture system where a large set of light sources need to be deployed, which are likely to lead to an inconvenient and hazardous system in some cases. We propose to simplify both data capture and hardware design by employing a two-source PS variation where only two light sources are required and therefore only two images need to be captured per reconstruction. In general, the equations in (11) are nonlinear, and therefore the two-source PS problem is not well posed. This ambiguity problem is illustrated in Fig. 1.

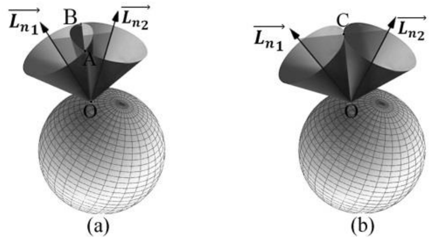
$\overrightarrow{L_{n1}}$
B
$\overrightarrow{L_{n2}}$
A
O
(a)
$\overrightarrow{L_{n1}}$
C
$\overrightarrow{L_{n2}}$
O
(b)

Fig. 1. An illustration of PS ambiguity

When only one light source is concerned, the surface normal vectors that produce a specific intensity value at point O form a cone with apex at this point and axis in the direction of illumination $\overrightarrow{\boldsymbol{L_{n1}}}$. In the case of two illuminators, the surface normals should belong to two such cones and therefore exist at the intersections of the two cones. Two cones with the same apex either have two intersections or one intersection (the case of no intersection does not occur for PS images). These two scenarios corresponding to ambiguous solutions and a unique solution respectively are shown in Fig. 1(a) and Fig. 1(b) where $\overrightarrow{OA}$ and $\overrightarrow{OB}$ represent the ambiguous solutions and $\overrightarrow{OC}$ represents the unique solution. This can be mathematically explained by deriving a pair of equations as in the general form of Eq. (14). If constant albedo is assumed for simplicity, and let $\overrightarrow{\boldsymbol{L_{n1}}} = [a_x, a_y, a_z]$, $\overrightarrow{\boldsymbol{L_{n2}}} = [b_x, b_y, b_z]$, two PS images yield:

$$\begin{cases} I_1 = \dfrac{-pa_x - qa_y + a_z}{\sqrt{1+p^2+q^2}} \\ I_2 = \dfrac{-pb_x - qb_y + b_z}{\sqrt{1+p^2+q^2}} \end{cases} \tag{15}$$

The solutions for p and q are produced in a similar way to [38]. Let

$$T \triangleq \sqrt{1+p^2+q^2} \tag{16}$$

Rearranging Eq. (16) yields

$$\begin{cases} pa_x + qa_y = a_z - I_1T \\ pb_x + qb_y = b_z - I_2T \end{cases} \tag{17}$$

Solving Eq. (17) for p and q in terms of T produces equations of the form

$$\begin{cases} p = \varepsilon_1 T + \varepsilon_2 \\ q = \varepsilon_3 T + \varepsilon_4 \end{cases} \tag{18}$$

where $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$ are functions of known values $I_1, I_2, a_x, a_y, a_z, b_x$, b_y and b_z. Combining Eq. (16) and (18) provides a quadratic equation for T of the form

$$\lambda_2 T^2 + \lambda_1 T + \lambda_0 = 0 \tag{19}$$

where λ_0, λ_1 and λ_2 can be calculated as functions of known values I_1, $I_2, a_x, a_y, a_z, b_x, b_y$ and b_z. Solving this quadratic equation ($\lambda_2 \neq 0$ in this case) gives:

$$T_{1,2} = \frac{-\lambda_1 \pm \sqrt{{\lambda_1}^2 - 4\lambda_2\lambda_0}}{2\lambda_2} \tag{20}$$

The two pairs of derivatives then become:

$$\begin{cases} p_{1,2} = \dfrac{a_z b_y - b_z a_y + I_2 T_{1,2} a_y - I_1 T_{1,2} b_y}{a_x b_y - b_x a_y} \\ q_{1,2} = \dfrac{a_x b_z - a_z b_x + I_1 T_{1,2} b_x - I_2 T_{1,2} a_x}{a_x b_y - b_x a_y} \end{cases} \tag{21}$$

It should be noted that $a_x b_y - b_x a_y \neq 0$, in order to produce meaningful solutions. To remove the ambiguity, we follow [38] and enforce integrability and continuity properties of a surface, assuming that the surface normals are continuous and that the surface height is twice differentiable. According to the continuity property of a surface, this study suggests that an arbitrary surface can be divided into connected regions R_c ($c \in \mathbb{Z}$), where there exists either a unique solution for T or a pair of solutions. On the other hand, integrability provides Eq. (22):

$$\int_{(x,y)\in R_c} \left(\frac{\partial p_{1,2}}{\partial y} - \frac{\partial q_{1,2}}{\partial x}\right)^2 = 0 \tag{22}$$

The pairs of p and q that agree with this equation are the true gradients and therefore correspond to the true surface normals. It is worth noting that when a unique solution can be found, the surface normal lies on the plane defined by the two lighting vectors (illumination directions). In the case of two ambiguous solutions, the surface normals lie on the different sides of this plane. We combine this with Eq. (22) so that the false solution can be discarded and the ambiguity can be removed.

5. Experiments and Results

A. 3D reconstruction results

We demonstrate the accuracy of the two-source PS method by applying this algorithm to a series of image sets from the Photoface database [39]. The accuracy is then compared to that of the standard (4-source) PS method by evaluating the ℓ_2-norm error for surface normals and the root-mean-square (RMS) error for surface depth.

The Photoface database is one of the few databases containing images captured under PS settings. It is deemed a suitable representation for realistic data as the data capture device was placed at the entrance to a workplace to ensure casual usage. We firstly show an example of 3D face reconstruction by applying the two-source PS method to the image set of the first subject in this database (Fig. 2).

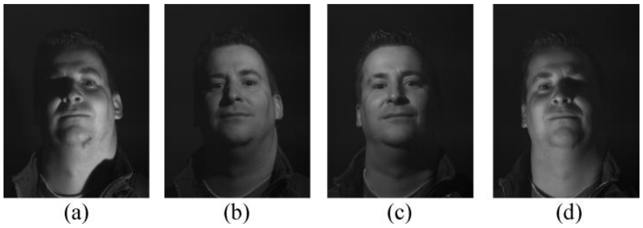

Fig. 2. A sample PS image set from the Photoface database

Note that the proposed two-source PS method only employs the two images illuminated from the top (Fig. 2(b) and (c)) while the standard PS method uses all the four images for reconstruction. It has been experimentally observed that illuminating from the top-left and the top-right directions creates relatively fewer self-cast shadows. The recovered x gradient image and the depth image are compared with those from the standard PS in Fig. 3. The recovery of depth images in this paper is based on the algorithm introduced in [40].

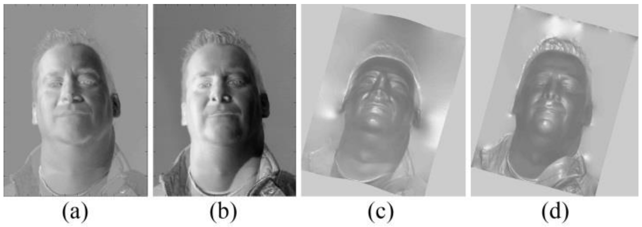

Fig. 3. A comparison of a 3D face reconstructions between the two-source PS and the 4-source PS. (a) and (b) are the x gradient images from 2-source PS and 4-source PS, respectively. (c) and (d) are the corresponding depth images.

The difference between a pair of reconstructions by the two-source and four-source PS method is more likely to be seen in facial regions that are less continuous (e.g. eye regions). Possible causes include sharp changes in face depth, non-Lambertian reflectance and shadows. Overall, comparable reconstruction results can be reflected by this example in this particular visualisation form. We further provide a comparison of the reconstructed depth images (Fig. 4) for 4 other subjects (subject 1002 to 1005) from the Photoface database.

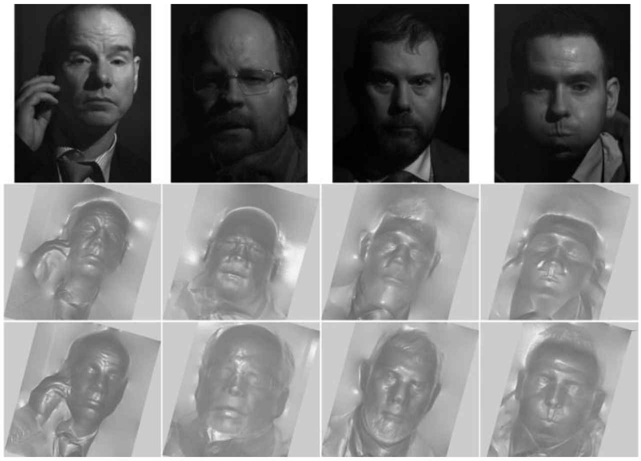

Fig. 4. 3D face reconstructions for 4 subjects from the Photoface database using the two-source PS and the standard 4-source PS methods. Top to bottom: one of the PS images, reconstructions by the two-source PS and reconstructions by the standard PS.

While the x gradient images and the depth images offer a visual comparison, a statistical analysis follows, which evaluates the surface normals and the depth images by measuring the ℓ_2-norm and the RMS errors for the first 100 subjects. As [41] has already calculated the errors between the standard PS and the ground truth obtained by a 3dMD projected pattern range finder [42], we are able to set the standard PS method as a reference for our evaluation. We followed [41] and cropped all reconstructions to 160×200 regions centred on the nose tip in order that the evaluations are consistent. While [41] measured the ℓ_2-norm and the RMS errors for 8 subjects from the Photoface database, we calculated the errors for the first 100 subjects for a more objective evaluation. These results can be seen in Fig. 5.

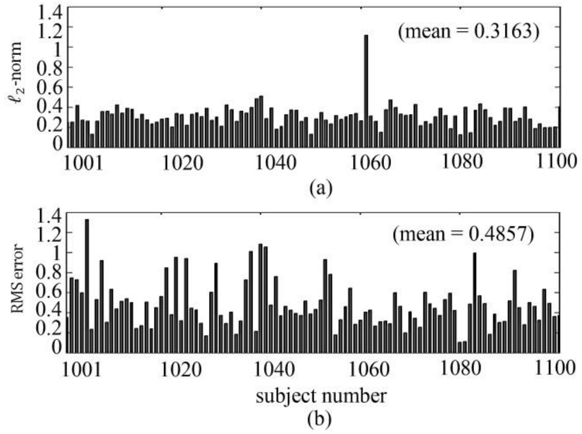

Fig. 5. (a) The ℓ_2-norm errors for surface normals and (b) the RMS errors for surface depth, for the first 100 subjects in the Photoface database, when the two-source PS method and the standard PS method are used for reconstructions.

Note that the ℓ_2-norm error for subject No. 1061 is relatively large due to the extreme head pose and the large background area. The average ℓ_2-norm and the RMS errors for the 100 subjects are 0.3163 and 4.8757, respectively. Consider a Cartesian coordinate system with x, y and z axes, a ℓ_2-norm error is caused by a unit vector deviating in the x and/or y directions. This is similarly to angular deviations in a spherical coordinate system represented by radial distance, azimuthal angle, and polar angle. ℓ_2-norm error of 0.3163 corresponds to an error of only 5.74 degrees when a unit surface normal vector [x, y, z] deviates in either the x direction or the y direction. On the other hand, The RMS error of 4.8757 pixels corresponds to only 1.12% of the average length of the 100 face images (i.e. 420 pixels). Therefore, both the 3D face visualisation and the statistical study can validate that, when realistic data are concerned, the two-source PS method has achieved comparable results to those from the standard PS method.

B. Gender recognition evaluations on public databases

In order to evaluate the performance of the proposed gender recognition algorithm under controlled environments and real-world conditions, the Grey FERET database [43], the Colour FERET database [44], the FRGCv2 database [45] and the Labelled Face in the Wild (LFW) database [46] are employed. The Grey FERET database (referred to as the FERET database in the rest of the paper) and the Colour FERET database consist of images of 1762 subjects in greyscale and 1199 subjects in colour, respectively. Although captured under controlled environment, they still pose great challenge as they accommodate different ethnicities, facial expressions, facial accessories, facial makeup and illumination conditions. The FRGCv2 database includes 4007 depth images belonging to 466 subjects. These data also include different ethnicities and age groups. The LFW database is considered one of the most challenging databases and has become the evaluation benchmark for face recognition under unconstrained environments. It contains 13233 colour facial images of 5749 subjects collected from the web in inconsistent image quality.

Different evaluation methods were employed in our experiments, such as 5-fold cross validation and 2-fold cross validation, detailed in Table 1. We also evaluated different types of features including the greyscales, LBP (extracted using the circularly symmetric neighbour sets with 8 neighbouring pixels and radius of 1 [47]), LBP histogram with uniform pattern and SIFT. Their respective performances on the FERET database are summarised in Fig. 6.

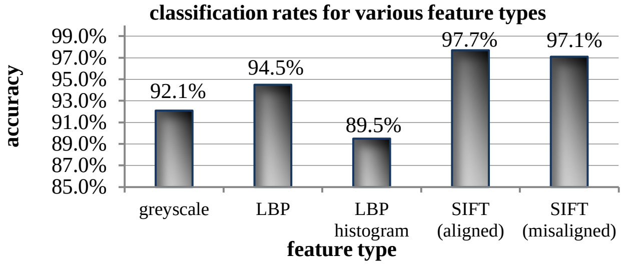

Fig. 6. Gender classification rates for various feature types, tested on the FERET fa partition. Results were obtained by a linear SVM.

In the evaluation experiments, we investigated various parametric settings and identified the parameters that gave the highest accuracy. Four groups of experiments were conducted to inspect image size, sampling window size, sampling step and Gaussian component number. Note that only one parameter varied in each group of experiments, where the parameter that yielded the highest classification rate was selected to form the optimal parametric setting. This was found to be 1) image size: 160×120 pixels (the largest we experimented with for the FERET database), 2) sampling window size: 24×24 pixels, 3) sampling steps: 4 pixels, and 4) Gaussian component number: 512.

The impact of PCA was also explored by reducing the original facial descriptors to dimensionality of 32, 64 and 96, respectively. It can be shown in Fig. 7 that when the first 64 principal components are employed, the highest classification rate can be achieved.

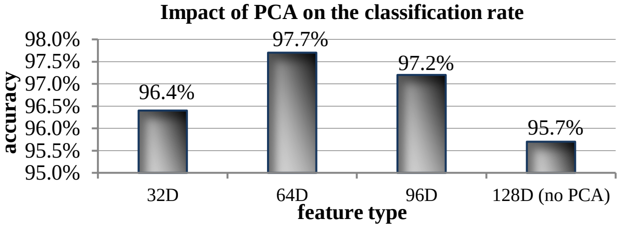

Fig. 7. Impact of PCA on gender classification, tested on the FERET fa partition. Results were obtained by a linear SVM.

As SIFT features yielded the highest classification accuracy, it was further tested on the other databases with the optimal parametric setting. Note that our SIFT features were only extracted at one scale since we did not observe any improvement when multiple scales are used. In addition, we only employ a linear SVM since the RBF kernel in the experiments did not contribute to higher classification rate. The classification rates for aligned and misaligned images in the LFW database are 92.5% and 92.3%, respectively. The classification rates for misaligned images in the Colour FERET database and the FRGCv2 database are 98.0% and 96.7%, respectively. To further validate the proposed method, we compare these results to 12 other gender recognition studies in the literature, detailed in Table 1.

Table 1. A comparison of the proposed method with other gender recognition methods on public databases [a]

Method	Description	Database	Validation method	Accuracy	Limitation
the proposed method	*FV encoding*	*FERET fa 1762*	*5-CV*	*97.7%*	*slow at training stage*
		FERET fa 1762	*50%/50%*	*96.9%*	
		FERET fa 1762	*50%/50%**	*97.9%*	
		FERET fa+fb 900 (u)	*5-CV*	*96.1%*	
		FERET fa+fb 2400	*50%/50%*	*98.3%*	
		FERET fa+fb 2400	*50%/50%**	*99.5%*	
		Colour FERET 700	*2-CV**	*98.0%*	
		LFW all (u)	*5-CV*	*92.5%*	
		FRGCv2 depth all 466 subjects (u)	*5-CV*	*96.7%*	
[14]	classifier fusion	FERET fa+fb 900 *(u)*	5-CV	92.9%	6 classifiers needed
[16]	RBF-SVM	FERET thumbnail1855	5-CV*	96.6%	*
[17]	refined LBP	LFW 7443 selected	5-CV	94.8%	manual data selection
[18]	LBP, wavelet transform	FERET fa+fb 2400	50%/50%*	99.3%	manual data selection
[19]	facial strips	FERET fa 1763	not specified	98.8%	slow & need alignment
[21]	CNN	FERET fa 1762	5-CV*	97.2%	*
[22]	CNN	FERET fa 1762	5-CV*	96.4%	*
[48]	geometric facial features	Indian face	trained with 40 subjects	95.6%	database of small size
[25]	2DPCA Gabor space	Colour FERET 700	2-CV*	98.4%	*
		LFW all	2-CV*	89.1%	
[26]	2DPCA Gabor space	Colour FERET	2-CV*	98.2%	*
		LFW all	2-CV*	88.3%	
[27]	LBP, shape index	FRGCv2 depth all 466 subjects	5-CV*	93.7%	*
[28]	Random Forest votes	FRGCv2 depth all 466 subjects	leave-one-out CV	97.2%	/

[a] The '*' notation indicates that training data and testing data may contain different images of the same subject(s); '(u)' represents unique subjects; '5-CV' and '2-CV' represent five-fold and two-fold cross validation, respectively; '50%/50%' represents half the data for training and the other half for testing.

With a number of evaluation methods implemented on different databases and database partitions, it can be seen from Table 1 that the proposed gender recognition algorithm outperforms most studies in comparison under the same experimental settings. Although its classification rate evaluated on the Colour FERET database is 0.4% lower than the 2DPCA based methods [25,26], the results on unconstrained data (i.e. the LFW database) are 3.4% higher under cross-validation evaluations. In the following subsection, we illustrate how the proposed gender recognition method can be further boosted by the utilisation of reconstructed 3D facial data.

C. Gender recognition evaluations on 3D reconstructions from PS

In the preceding subsection, we demonstrated that even with greyscale images, the proposed FV encoding method can achieve excellent gender classification rates comparable to state-of-the-art studies. As much as this gives promising reliability, the inherent limitation of utilising greyscale or colour images is mainly posed by dynamic lighting conditions. This cannot be resolved by designing classifiers with higher discriminative power, but can be tackled by seeking light-independent features – the 3D facial features.

In this subsection, we draw a comparison between utilising 2D facial features and 3D features for FV gender recognition. For 2D facial features, we evaluated the SIFT features as a reference, which had gained the highest classification rate in our previous experiments. For 3D facial features, we firstly performed 3D reconstructions by employing the two-source PS method and then extracted 1) the x gradient features, 2) depth features, 3) SIFT features extracted from the x gradient images (gradient-SIFT), 4) SIFT features extracted from the depth images (depth-SIFT), 5) LBP features extracted from the x gradient images (gradient-LBP) and 6) LBP features extracted from the depth images (depth-LBP).

Currently, we have not discovered any public databases that contain both PS image sets and the corresponding greyscale/colour images which, at the same time, have a suitable male-female ratio. Therefore we developed a 2D+3D data capture system that could gather both PS images and colour images at high frame rate. These data were then employed for our gender recognition evaluations.

1. Development of a 2D+3D data capture system and data capture experiments

We developed a 2D+3D data capture system, shown in Fig. 8, intended to gather PS facial data as well as colour images in real-world environments. The design of this system consists of 1) a high-definition (HD) 47-inch display, 2) a webcam (referred to as camera 1 in the rest of the paper) operating at 640×480 resolution, 3) two near-infrared (NIR) LEDs (SFH4232 with 850 nm wavelength) for PS illumination, 4) a Point Grey GS3-U3-41C6NIR-C camera (referred to as camera 2 in the rest of the paper) operating at 2048×800 resolution, with a 850nm +/-5nm NIR band pass filter 5) a PC in the cabinet for data storage and processing, and 6) a control unit that synchronises NIR LEDs with the cameras. The cameras are 2.1 metres from the floor and the NIR illuminators are both 0.75 metres from the cameras.

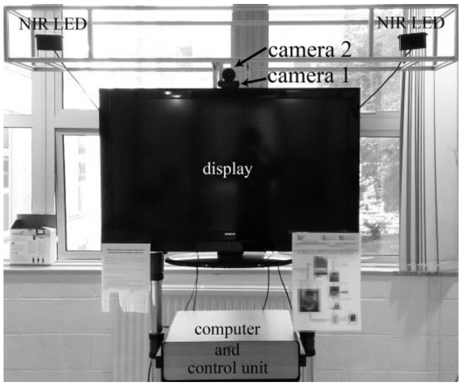

Fig. 8. System structure for the 2D+3D imaging system

The data capture process is described as follows:

1) The data capture system was firstly placed at a university public kitchen area with the presence of only artificial light sources. A total number of 45 volunteers participated in this experiment in 2 different recording sessions.

2) The system was then placed at a university library foyer where lighting conditions were affected by both lamps installed in the foyer and sunshine through the window. A total number of 127 volunteers participated in this experiment in 4 different recording sessions.

3) In the overall 6 experiment sessions, every volunteer was asked to stand at 1 metre away from the display and look at the display (each face was therefore near frontal as we explored the impact of dynamic illumination). The two NIR LEDs then lighted up alternatingly when camera 1 captured colour images of the volunteer and camera 2 captured PS image sets (two NIR images in each set) of the volunteer. It should be noted that since camera 2 was covered by a NIR filter, the ambient light posed negligible impact on camera 2 while the NIR LEDs had minimal impact on camera 1 due to their invisible spectrum.

4) The overall 6 recording sessions gathered image data of 90 male subjects and 82 female subjects. The image data contain Caucasian, Asian and African faces, with an age range from 18 to 58. A few image

sets were stripped where image frames only contained partial faces due to the subjects standing at improper locations while being recorded. In the end, we employed image data of 75 female subjects and an equal number of male subjects (150 faces overall) for the gender recognition evaluations.

A sample image captured by camera 1 and a PS image set captured by camera 2 are displayed in Fig. 9.

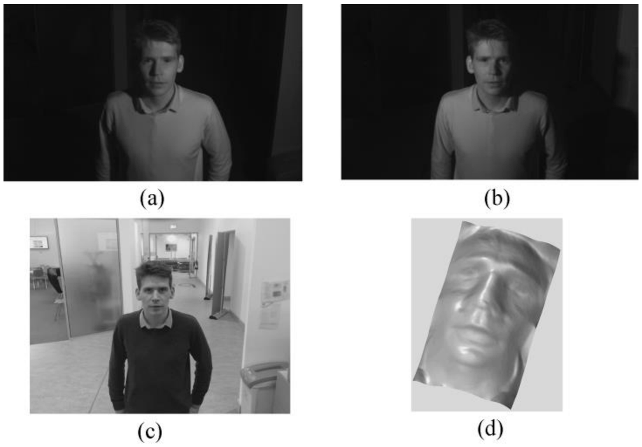

Fig. 9. (a) and (b) are the PS image set used for the 3D reconstruction in (d); (c) is the 2D colour image of the same subject captured under the same environmental setting, displayed in greyscale.

Apart from these raw images, the depth image for the same subject is displayed in Fig. 9(d) where the reconstruction is achieved by the two-source PS method. Other than being able to gather data, this system offers significant potential in the advancement of HCI. This is enabled by the employment of the two-source PS method that facilitates the design of a system with hardware simplicity as well as desirable real-time performance.

2. Gender recognition results on 2D and 3D face images

For all the 150 subjects, we employed both the colour images (which are converted to greyscale images before FV encoding) and the PS images, and performed evaluations respectively. Various types of features were extracted from them and were encoded into FVs. The optimal parameters summarised in the preceding subsection were employed while the images were resized to 192×144 for a consistent aspect ratio.

The classification rates resulted from 6 types of 3D features were compared with the 2D SIFT features as well as the fused features, illustrated by Fig. 10.

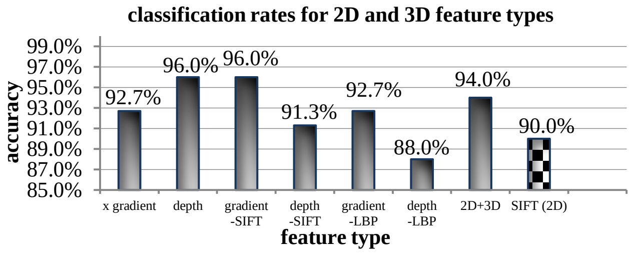

Fig. 10. Evaluations of 2D and 3D features for FV encoding. Results were obtained by a linear SVM.

It can be seen from Fig. 10 that 3D features have generally resulted in superior classification accuracies over 2D features and fused features, except for the depth-LBP features which are most likely to have be interfered by image noise. The highest accuracy obtained in this experiment surpassed the 2D SIFT features by 6%. In this experiment, this increase in accuracy is reflected when the illumination conditions are acceptable. Therefore, it can be inferred that when the illumination condition worsens, 2D features will suffer more while 3D features are much less influenced, hence at such times greater superiority of 3D features is observed. Moreover, being illumination independent, 3D features can tolerate variations in dynamic scenes which commonly render 2D features unstable and incapable of discriminating gender groups. However when we fused the greyscale-SIFT features with gradient-SIFT features (the top 256 Gaussian components were used), we observed a decrease in accuracy. This is likely due to the 2D features being considered redundant and interfering in the new feature space and therefore lowered the overall discriminative power of the fused features.

6. Discussion

A. The merits and limitations of the Fisher Vector encoding method

FVs bring various benefits to object classification tasks. As well as the high discriminability they add to facial features, which results in superior accuracy, they offer the following advantages:

1. Fisher vectors are of uniform length (i.e. dimensionality). This allows different types of low level dense features to be encoded into feature vectors with the same dimensionality. As a result, various feature types, regardless of the source they are extracted from, can be easily manipulated, e.g. feature fusion.

2. This method is versatile in that it can encode almost any type of features. In our experiments, we have encoded and evaluated as many as 10 different feature types.

3. At the classification stage, only a linear SVM is needed for the best accuracy. This makes the approach more efficient. The reason that SVMs with non-linear kernels do not offer additional benefits in our experiments is that "learning a kernel classifier using the kernel is equivalent to learning a linear classifier on the Fisher vectors" [49].

4. FVs are robust to head pose and therefore face alignment can be eliminated. This approach samples and encodes dense facial patches. Although features within a patch are bounded by local geometry, they are globally independent from facial geometry as individual patches are treated equally.

However, the FVs are of high dimensionality. Therefore, at the training stage, it requires a large memory space for data storage. This also causes a prolonged offline training time. However, computers nowadays can be easily equipped with large memories, and as soon as the classifier offline training is completed, the classification can be achieved online in real time.

B. Databases and evaluation method

Four public databases have been employed to evaluate our method in laboratory environments as well as in real-world scenes. For 3D face reconstruction and 3D feature analysis, we used self-collected data obtained in a series of data capture experiments since public PS based databases (e.g. the Photoface database) either contain too few subjects or contain biased male-female subject ratio.

We recommend that, in the evaluation process, the 5-fold cross validation technique should be adopted for strict and efficient evaluations. In addition, evaluations with different images of the same subjects in both training set and testing set should be avoided. The increase of accuracy (normally 1% or more) caused by this is most likely due to classification of gender-specific features that are already learnt by a classification model. Other validation methods have also been adopted in our evaluations, in order to provide an objective comparison to other gender recognition studies (see Table 1).

The evaluation of the proposed method is two-fold. Firstly, we prove with 2D features that FV encoding is a superior method for automatic gender recognition. Subsequently, we prove that the employment of 3D features can further boost the performance of the proposed method. Therefore, it can be concluded that the FV encoding method when employing 2D or 3D features exhibits superior excellence in automatic gender recognition.

C. The efficiency of PS 3D reconstruction and gender recognition

We evaluated the times spent for the first 100 reconstructions in the 3D face reconstruction experiments. The computational times for the 2-source PS and the 4-source PS are plotted in Fig. 11 for a comparison.

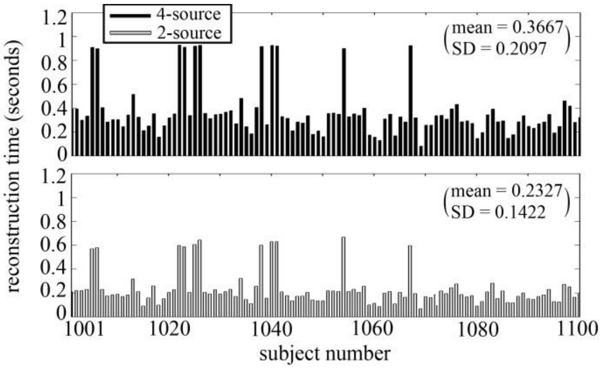

Fig. 11. An evaluation of 3D face reconstruction efficiency. Note that the times include both surface normal recovery and depth estimation. The mean and the standard deviation (SD) of reconstruction times are calculated for both the 2-source and the 4-source PS method.

This test was performed with Matlab R2014a and on a computer with an Inter(R) Core(TM) i5-4570 CPU and 12G memory. It can be seen from Fig. 11 that, on average, the two-source PS method only consumes less than two thirds of computational time required by the standard PS in the 3D reconstruction stage. Note that the reconstruction times for the 100 subjects vary mainly due to their different image sizes (average size: 733 × 624 pixels).

Moreover, in the image capture stage, the two-source PS method halves the time consumption by capturing only 2 images per reconstruction instead of the commonly required 4 images. The standard PS has been able to spawn real-time 3D imaging systems [10], the two-source PS variation should promise to boost the efficiency of 3D imaging systems to a higher level.

Overall, the reduced image capture time and the shortened 3D reconstruction time are the two main factors that can result in high efficiency of the two-source PS method and therefore bring huge application potential.

We also evaluated the efficiency of the proposed gender recognition method. Under the same hardware and software setting, FV encoding for 1200 images took 37 seconds (0.03 seconds per image), while SVM gender classification took 1.19 seconds (less than 0.001 seconds per image). This indicates that even with a Matlab implementation, the proposed algorithm can gain excellent real-time performance (over 30 images per second). This is mainly due to the fact that the FV encoding method only requires a linear SVM to gain high accuracy. It should be noted that the dimensionality of a FV is only determined by the number of Gaussian components in a GMM (which is normally a fixed value as we have identified the optional parametric setting, i.e. 512 in the experiments). Therefore when different types of features are concerned, computational time would be similar.

D. 2D facial feature vs. 3D facial features

3D face reconstructions from our self-collected data and the Photoface database show that the reconstruction accuracy deteriorates as the camera-subject distance increases. This is likely to be caused by illumination attenuation from the NIR LEDs, a reduction in image resolution and the change in light (NIR LEDs) positions relative to the subject.

Therefore, from an applied point of view, we recommend a combination of 2D and 3D features for gender recognition, and design of an adaptive mechanism for real-world HCI system implementations. For example, when the camera-subject distance is within 2 metres, 3D features should be used for classification where possible; otherwise 2D features should be employed.

7. Conclusion

This paper employs the FV encoding method for automatic gender recognition and draws a comparison between 2D and 3D facial features by encoding up to 10 different feature types for a comprehensive study. A two-source PS variation is also proposed to achieve 3D face reconstruction with high accuracy and efficiency. Tested on 4 publicly available databases and a self-collected database, the proposed gender recognition method has proved to be highly accurate while the employment of 3D features has further boosted the classification accuracy and robustness. The 2-source PS 3D reconstruction method has been tested on 100 subjects from the Photoface database and has been compared to the standard PS method in terms of accuracy and efficiency. A 2D+3D imaging system has been developed that is capable of collecting colour images as well as PS image sets. This system has the potential for facilitating HCI with an inexpensive, simple, accurate and efficient hardware device. We conclude that although 3D features promise to be more discriminant and invariant, they should be combined with 2D features in real-world implementations to adapt to dynamic environments.

References:

1. J. Marchbank, and G. Letherby, *Introduction to gender: Social science perspectives*, (Routledge, 2014).
2. F. Ahmad, Z. Ahmed, and A. Najam, “Soft biometric gender classification using face for real time surveillance in cross dataset environment,” in *Proceedings of the16th International Multi Topic Conference* (IEEE, 2013), pp. 131-135.
3. G. Toderici, S. M. O’malley, G. Passalis, T. Theoharis, and I. A. Kakadiaris, “Ethnicity-and gender-based subject retrieval using 3-D face-recognition techniques,” International Journal of Computer Vision **89**, 382-391 (2010).
4. A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino, “2D and 3D face recognition: A survey,” Pattern Recogn. Lett., **28**, 1885-1906 (2007).
5. L. Ballihi, B. Ben Amor, M. Daoudi, A. Srivastava, and D. Aboutajdine, “Boosting 3-D-geometric features for efficient face recognition and gender classification,” IEEE Trans. Information Forensics and Security **7**, 1766-1779 (2012).
6. A. Hadid and M. Pietikäinen, “Combining appearance and motion for face and gender recognition from videos,” Pattern Recognition **42**, 2818-2827 (2009).
7. A. M. Baumberg, and D. C. Hogg. “An efficient method for contour tracking using active shape models,” in Proc. *IEEE Workshop on Motion of Non-Rigid and Articulated Objects* (IEEE, 1994), pp. 194-199.
8. D. Scharstein and R. Szeliski, “High-accuracy stereo depth maps using structured light,” in *Proceedings of 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE, 2003) pp. I-195.
9. V. Blanz and T. Vetter, “Face recognition based on fitting a 3D morphable model,” IEEE Trans. Pattern Anal. Mach. Intell. **25**, 1063-1074 (2003).
10. T. Malzbender, B. Wilburn, D. Gelb and Ambrisco, B, “Surface Enhancement Using Real-time Photometric Stereo and Reflectance Transformation,” Rendering Techniques, (2006).
11. C. B. Ng, Y. H. Tay, B. M. Goi, “Vision-based human gender recognition: A survey,” arXiv preprint arXiv:1204.1611, (2012).
12. P. Viola and M. J. Jones, “Robust real-time face detection,” International journal of computer vision **57,** 137–154 (2004).
13. P. Viola and M. J. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings of IEEE Computer Society Conference on CVPR* (IEEE, 1988), pp. I-511–I-518 (2001).
14. E. Mäkinen and R. Raisamo, “An experimental comparison of gender classification methods,” Pattern Recogn. Lett. **29**, 1544–1556 (2008).

15. E. Mäkinen and R. Raisamo, "Evaluation of gender classification methods with automatically detected and aligned faces," IEEE Trans. Pattern Anal. Mach. Intell. **30,** 541–547 (2008).

16. B. Moghaddam, M. H. Yang, Gender classification with support vector machines, in *Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition* (IEEE, 2000), pp. 306–311 (2000).

17. C. Shan, "Learning local binary patterns for gender classification on real-world face images," Pattern Recogn. Lett. **33**, 431–437 (2012).

18. I. Ullah, M. Hussain, H. Aboalsamh, G. Muhammad, A.M. Mirza, and G. Bebis, "Gender recognition from face images with dyadic wavelet transform and local binary pattern," in *Advances in Visual Computing* (Springer 2012), pp. 409–419.

19. P. H. Lee, J. Y. Huang, and Y. P. Huang, "Automatic gender recognition using fusion of facial strips," in *Proceedings of 20th IEEE International Conference on Pattern Recognition* (IEEE, 2010), pp. 1140–1143.

20. J. G. Wang, J. Li, W. Y. Yau, and E. Sung, "Boosting dense SIFT descriptors and shape contexts of face images for gender recognition," in *Proceedings of IEEE Computer Society Conference on CVPRW* (IEEE, 2010), pp. 96–102.

21. F. H. C. Tivive and A. Bouzerdoum, "A gender recognition system using shunting inhibitory convolutional neural networks," in *Proceedings of IEEE International Joint Conference on Neural Networks* (IEEE, 2006), pp. 5336–5341.

22. S. L. Phung, A and Bouzerdoum, "A pyramidal neural network for visual pattern recognition," IEEE Trans. Neural Networks **18**, 329–343 (2007).

23. L. A. Alexandre, "Gender recognition: A multiscale decision fusion approach," Pattern Recogn. Lett. **31**, 1422–1427 (2010).

24. Y. Hu, J. Yan, and P. Shi, "A fusion-based method for 3D facial gender classification," in *Proceedings of the 2nd International Conference on Computer and Automation Engineering* (IEEE, 2010), pp. 369–372.

25. P. Rai and P. Khanna, "A gender classification system robust to occlusion using Gabor features based $(2D)^2$ PCA," Journal of Visual Communication and Image Representation **25**, 1118-1129 (2014).

26. P. Rai and P. Khanna, "An illumination, expression, and noise invariant gender classifier using two-directional 2DPCA on real Gabor space," Journal of Visual Languages & Computing **26**, 15-28 (2015).

27. X. Wang and C. Kambhamettu, "Gender classification of depth images based on shape and texture analysis," in Proceedings of *Global Conference on Signal and Information Processing* (IEEE, 2013), pp. 1077-1080.

28. B. Xia, B. B. Amor and M. Daoudi, "Exploring the Magnitude of Human Sexual Dimorphism in 3D Face Gender Classification," in *Proc. Computer Vision–ECCV 2014 Workshops* (Springer, 2014), pp. 697-710.

29. J. Fagertun, T. Andersen and R. R. Paulsen, "Gender recognition using cognitive modelling," in *Proc. Computer Vision–ECCV 2012* (Springer, 2012), pp. 300-308.

30. T. Huynh, R. Min, and J. L. Dugelay, "An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data," in *Proc. Computer Vision–ECCV 2012* (Springer, 2012), pp. 133-145.

31. B. Xia, B. Ben Amor, D. Huang, M. Daoudi, Y. Wang,, and H. Drira, "Enhancing gender classification by combining 3d and 2d face modalities," in *Proceedings of the 21st IEEE European Signal Processing Conference (EUSIPCO)* (IEEE, 2013), pp. 1-5

32. D. Reynolds, Gaussian Mixture Models, in *Encyclopaedia of Biometrics* (Springer, 2009), pp. 659–663.

33. K. Simonyan, O. Parkhi, A. Vedaldi and A. Zisserman, "Fisher Vector Faces in the Wild," in *Proceedings of British Machine Vision Conference* (BMVA Press, 2013), pp. 8.1–8.12.

34. A. Vedaldi, B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proceedings of the international conference on Multimedia* (ACM, 2010), pp. 1469–1472.

35. F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. Computer Vision–ECCV 2012* (Springer, 2010), pp. 143–156.

36. R. J. Woodham, "Photometric method for determining surface orientation from multiple images," *Optical engineering*, ***19***, 139-144 (1980).

37. S. Herbort, and C. Wöhler, "An introduction to image-based 3D surface reconstruction and a survey of photometric stereo methods," 3D Research ***2***, 1-17 (2011).

38. R. Onn and A. Bruckstein, "Integrability disambiguates surface recovery in two-image photometric stereo," International Journal of Computer Vision **5**, 105-113 (1990).

39. S. Zafeiriou, M. Hansen, G. Atkinson, V. Argyriou, M. Petrou, Smith, M. and L. Smith, "The photoface database," in *Proceedings of IEEE Computer Society Conference on CVPRW* (IEEE, 2011), pp. 132-139.

40. R. T. Frankot and R. Chellappa, "A method for enforcing integrability in shape from shading algorithms," IEEE Trans. on Pattern Analysis and Machine Intelligence **10**, 439-451 (1988).

41. M. F. Hansen, G. A. Atkinson, L. N. Smith, and M. L. Smith, "3D face reconstructions from photometric stereo using near infrared and visible light," Comput. Vis. Image Und. **114**, 942-951 (2010).

42. "3dMDface System," http://www.3dmd.com/3dmd-systems/3d-systems/3dmdface/.

43. P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," Image and vision computing **16**, 295–306 (1988).

44. P. J. Phillips, H. Moon, S. A. Rizvi and P. J. Rauss, "The FERET Evaluation Methodology for Face Recognition Algorithms," IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1090-1104 (2000).

45. P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min and W. Worek, "Overview of the face recognition grand challenge," in *Proceedings of IEEE Computer Society Conference on CVPR* (IEEE, 2005), pp. 947-954 (2005).

46. G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in: University of Massachusetts Technical Report 07-49, 1–11 (2007).

47. T. Ojala, M. Pietikäinen, and T. Mäenpää, "Gray scale and rotation invariant texture classification with local binary patterns," in *Proc. Computer Vision–ECCV 2000* (Springer, 2000), pp. 404-420.

48. S. Kalam and G. Guttikonda, "Gender Classification using Geometric Facial Features," International Journal of Computer Applications **85**, 32-37 (2014).

49. F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. Computer Vision–ECCV 2010* (Springer, 2010), pp. 143-156.