Computer Vision and Image Understanding 149 (2016) 32–50

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Gender and gaze gesture recognition for human-computer interaction

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

ARTICLE INFO

Article history:
Received 17 April 2015
Revised 16 March 2016
Accepted 18 March 2016
Available online 29 March 2016

Keywords:
Assistive HCI
Gender recognition
Eye centre localisation
Gaze analysis
Directed advertising

ABSTRACT

The identification of visual cues in facial images has been widely explored in the broad area of computer vision. However theoretical analyses are often not transformed into widespread assistive Human-Computer Interaction (HCI) systems, due to factors such as inconsistent robustness, low efficiency, large computational expense or strong dependence on complex hardware. We present a novel gender recognition algorithm, a modular eye centre localisation approach and a gaze gesture recognition method, aiming to escalate the intelligence, adaptability and interactivity of HCI systems by combining demographic data (gender) and behavioural data (gaze) to enable development of a range of real-world assistive-technology applications.

The gender recognition algorithm utilises Fisher Vectors as facial features which are encoded from low-level local features in facial images. We experimented with four types of low-level features: greyscale values, Local Binary Patterns (LBP), LBP histograms and Scale Invariant Feature Transform (SIFT). The corresponding Fisher Vectors were classified using a linear Support Vector Machine. The algorithm has been tested on the FERET database, the LFW database and the FRGCv2 database, yielding 97.7%, 92.5% and 96.7% accuracy respectively.

The eye centre localisation algorithm has a modular approach, following a coarse-to-fine, global-to-regional scheme and utilising isophote and gradient features. A Selective Oriented Gradient filter has been specifically designed to detect and remove strong gradients from eyebrows, eye corners and self-shadows (which sabotage most eye centre localisation methods). The trajectories of the eye centres are then defined as gaze gestures for active HCI. The eye centre localisation algorithm has been compared with 10 other state-of-the-art algorithms with similar functionality and has outperformed them in terms of accuracy while maintaining excellent real-time performance.

The above methods have been employed for development of a data recovery system that can be employed for implementation of advanced assistive technology tools. The high accuracy, reliability and real-time performance achieved for attention monitoring, gaze gesture control and recovery of demographic data, can enable the advanced human-robot interaction that is needed for developing systems that can provide assistance with everyday actions, thereby improving the quality of life for the elderly and/or disabled.

1. Introduction

With the emergence of personal computing in the late 1970s, the concept of Human-Computer Interaction (HCI) was pushed rapidly and steadily into the environment where the individual role of science or engineering was not sufficient for addressing the urgent need for increasing the usability of computer software and operating systems (Grudin, 2011). The potential of increased accessibility to personal computers and demand for higher usability (Karray et al., 2008) of computer platforms called for the practical need for HCI and a synthesis of science and engineering. As personal computers present more digital information to people, the way people perceive, process and respond to such data has largely altered. HCI has therefore not only become central to information science theoretically and professionally, but also has stepped into peoples' lives, offering multiple types of communication channels, i.e. modalities. These modalities gain their input via various types of sensors, mimicking human sensors including visual, audio and haptic sensors (Karray et al., 2008). They individually or combined give rise to a wide array of HCI systems. Among the various types of sensors, the visual sensor and the resulting visual based interaction modality are the most widespread, taking visual signals as

* Corresponding author: Tel.: +44 11732 86806.
E-mail address: wenhao.zhang@uwe.ac.uk (W. Zhang).

http://dx.doi.org/10.1016/j.cviu.2016.03.014
1077-3142/

inputs, which include but are not limited to images/videos of faces, bodies and hands (Jaimes and Sebe, 2007). The corresponding research spans the areas of gaze tracking, gait analysis, and recognition of face, gender, age, gesture and facial expression. While face recognition reveals an individual's identity, gender/age recognition aims to gather human demographics for the system to understand human characteristics. Facial expression recognition further estimates human affection states and gathers emotional cues while gaze tracking serves to extract users' attentive information and to predict their intentions.

These modalities, independent or combined, have played different assistive roles in their corresponding HCI applications. As an example of gaze tracking, smart solutions are available that monitor the gaze direction of a driver in order to identify driver distraction/drowsiness and provide timely alerts (Tawari et al., 2014). These driver assistance systems, capable of detecting and acting on driver inattentiveness, are of great value to road safety. In the area of gesture recognition, a sterile browsing tool, 'Gestix', has been designed for doctor-computer interaction. This assistive technology provides doctors the sterility needed in an operation room where radiology images can be browsed in a contactless manner. A doctor's hand is tracked by a segmentation algorithm using colour model back-projection and motion cues from image frames (Wachs et al., 2008). Moreover, when expression recognition is concerned, an intelligent tutoring system, namely a Learning Companion, is proposed to predict when a learner might be frustrated, initiate interaction depending on the user's affective state and provide support accordingly (Kapoor et al., 2007).

With regard to the visual modality, we categorise these tasks as 'demographic recognition' (e.g. age and gender recognition) or 'behavioural recognition' (e.g. gaze analysis), according to the source and the utilisation of visual signals. In this paper, we propose a HCI strategy by combining demographic and behavioural recognition such that more natural, interactive and user-centred HCI environments can be created. More specifically, although demographic data reveal user characteristics and help define the 'initial state' and the 'general theme' of a HCI session, they cannot address the dynamic nature of a HCI session where user behaviours are constantly changing. Therefore, on the one hand, behavioural recognition reflects user attentions and intentions; on the other hand, it allows a user to issue commands and to actively interact with a HCI system through various means. As a result, the combination of demographic and behavioural recognition provides fused knowledge throughout a HCI session and better mimics a natural face-to-face interaction.

Although research in HCI has become increasingly active and sophisticated, a few issues remain unresolved that restrict most works from being transformed into assistive HCI systems that can benefit the daily life of human beings. We summarise the three major general issues that undermine the practicability of these works as follows:

(1) Lack of accuracy in real-world scenarios. Many research works are tested on controlled databases where ideal illuminations, high-resolution images and desirable viewpoint are available. When tested under various types of scenes with dynamic environmental factors, their performance will drop severely.
(2) Undesirable real-time performance. As powerful as they might be, sophisticated algorithms often incur large computational cost, rendering them unsuitable for real-time implementation.
(3) High dependence on expensive or inconvenient hardware configuration. The cost and the complexity of algorithm implementation will limit the usability and applicability of any method. Cheap yet effective methods are in high demand in order to boost assistive technologies.

In order to fill these gaps, in Section 3, we present an accurate gender recognition method that exhibits high accuracy and robustness under controlled and uncontrolled environments. This method puts Fisher Vectors at its core and encodes low-level local features (e.g. greyscale values, Local Binary Pattern (LBP), LBP histograms and the Scale Invariant Feature Transform (SIFT)) into more discriminative features for gender classification. State-of-the-art accuracy is achieved by only a linear Support Vector Machine (SVM), which further confirms the superiority of Fisher Vectors.

In Section 4, we propose a modular eye centre localisation method that makes use of isophote and gradient features extracted by its two modules. The first module performs an initial estimation of eye centre locations using isophote features from face images and filters eye centre candidates for the second module. The second module then updates the eye centre locations using only local gradient features. This coarse-to-fine and global-to-regional scheme ensures that this method is fast and accurate.

To further explore the localised eye centres, in Section 5, we introduce gaze gesture recognition, i.e. classification of the trajectory patterns of eye centre locations in consecutive frames. These gaze gestures provide contactless substitutes for mouse/keyboard input. Gaze gesture recognition therefore offers great assistance to the elderly and the disabled in accessing a variety of digital systems.

As our algorithms only require a standard webcam to capture image data, we implemented all the above mentioned algorithms and integrated them in a directed advertising system – which is one of the assistive technology tools our algorithms can bring to real-world applications. The directed advertising system combines the interpretation of demographic data (gender) and behavioural data (gaze) in order to deliver customised advertisements to its users and allow them to remotely browse advertising messages.

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

(1) A novel gender recognition method utilising the Fisher Vector encoding method – a generic method that can encode almost all types 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 Fisher Vectors have been employed for gender recognition.
(2) A modular eye centre localisation method consisting of two modules and a Selective Oriented Gradient (SOG) filter. The eye centre localisation method has proved to be accurate, fast and, more importantly, robust to in-plane and out-of-plane head rotations. The SOG filter is specifically designed to detect and remove strong gradients from eyebrows, eye corners and self-shadows that many other methods suffer. The SOG filter also resolves general tasks regarding the detection of curved shapes.
(3) Design of gaze gestures and a gaze gesture recognition algorithm. The algorithm captures the relative attention of the users and enables them to control HCI systems by issuing gaze gestures. This will largely enhance the interactivity of current HCI systems.
(4) Development of an intelligent and assistive case study system. By combining demographic data and behavioural data, this system makes digital out-of-home advertising more adaptive and interactive so that it is of great value to both advertising agencies and consumers. This system also provides enabling technology for assisting diverse groups of people (including the elderly and the disabled), with accessing HCI systems with ease and convenience, in a wide range of applications.

2. Related work

A number of assistive technologies that utilise vision modalities have been introduced in the preceding section. In this section, we explore the combination of gender recognition and eye/gaze analysis and review a number of related state-of-the-art methods. We provide a summary of the current research state for gender recognition, eye centre localisation and gaze/gaze gesture analysis; and we discuss their potential benefits for assistive HCI applications, and identify the limitations that will critically hinder advancement in this area.

2.1. Gender recognition

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 facial attributes or environmental factors. The former complications mainly include age, ethnicity and makeup while the latter include 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 (Ng et al., 2012).

For the face detection stage, the Viola-Jones face detector (Viola and Jones, 2004) has been widely adopted due to its ease of implementation and relatively high accuracy. It is essentially a face detector that employs Haar-like features, a classifier learning with AdaBoost and a cascade structure (Viola and Jones, 2001). It can operate in real time and has allowed many practical applications to boom in the last decade.

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 the classification accuracy by a number of studies. For example, in a research (Mäkinen and R., 2008) evaluating a number of gender classification methods, it is concluded that Support Vector Machine (SVM) outperformed other classification methods with 86.54% accuracy on 36×36 aligned images, and that higher accuracy could be achieved by improving the implementation of the automatic alignment methods. Another research (Mäkinen and Raisamo, 2008) with regard to gender classification illustrated that face alignment brought an increase to the classification accuracy for various methods including use of neural networks, SVM, and Adaboost. Different pre-processing methods are experimented with in our method with the results reported in Section 3.

For the feature extraction and selection stage, a wide range of features are experimented with and evaluated in the literature. They include intensity values from greyscale images (Moghaddam and Yang, 2000), LBP (Shan, 2012, Ullah et al., 2012), facial strips (Lee et al., 2010), Haar-like features (Viola and Jones, 2001), SIFT features (Wang et al., 2010), etc. These features can be extracted globally from complete face images or locally from defined subregions of the face images.

For the classification stage, SVMs and neural networks have been the most popular classifiers. SVMs with different kernels were investigated in (Moghaddam and Yang, 2000) and convolutional neural networks were adopted by (Tivive and Bouzerdoum, 2006) and (Phung and Bouzerdoum, 2007) as gender classifiers.

Following the four major stages, a number of approaches have reported relatively high classification rate on publicly available datasets. Some representative works on gender classification are reviewed here, with a detailed comparison to the proposed method in Section 3.

A decision-fusion based method is presented by (Alexandre, 2010) that utilises multiple SVMs to classify intensity values, local binary patterns and histogram of edge directions as features. The three types of features are extracted from images of various sizes. Finally all classification results are integrated to make the final decision by means of majority voting, leading to 99.07% accuracy. However this result is obtained from a small subset of the FERET database and their validation method is not sophisticated enough to reflect the performance of their approach objectively. In addition, only controlled databases are used for training and testing in this research so that the applicability of this approach to dynamic environments remains unevaluated. Similarly, another study (Lee et al., 2010) employs 10 regression functions to conduct region-based classifications and feed the vector of classification results into an SVM to generate the final decision. Despite the 98.8% accuracy with the 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 in both the training and testing data 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%. This is an inherent limitation of conventional region-based approaches where defined facial regions have to be perfectly aligned. A fusion-based method (Hu et al., 2010) proposed to integrate different facial regions for gender recognition using the 'matcher weighing fusion' method. The facial landmarks for segmenting the face into its subregions are detected by a profile-based method and a curvature based method. Interestingly they prove experimentally that the fusion of multiple facial sub-regions is superior to the complete face region alone and that the upper face contains more discrimination ability regarding gender classification. Apart from classification fusion, feature fusion provides an alternative way to boost gender classification rates. In (Wang and Kambhamettu, 2013), 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%.

With individual works reviewed, we draw conclusions from the literature regarding the preferences in gender classification. 1) SVM 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 face alignment in the pre-processing stage. 6) Most works employ the 5-fold cross validation for accuracy estimation.

In most gender classification studies, limitations and complications are seen as both intrinsic factors and extrinsic factors. The former type is mainly the consequence of large amount of variation in facial appearance due to aging, makeup, facial occlusion, ethnicity and accessories. The latter type is largely due to environmental variations such as camera viewpoint (head pose), illumination condition, etc. Variations incurred by undesirable illumination conditions in particular are difficult to address by employment of powerful classifiers, but rather should be tackled by seeking for more robust and reliable facial features. This has defined the trend for gender recognition researches – the exploitation of 3D features that are independent of lighting conditions. 3D features have been employed individually (Hu et al., 2010, Fagertun et al., 2012) or fused (Wang and Kambhamettu, 2013, Huynh et al., 2012) with 2D features to better characterise face shapes and textures. This will

be reflected in our future works aiming to extend the proposed gender recognition algorithm by incorporating 3D features.

2.2. Eye/gaze analysis

Eye/gaze analysis is receiving an increasing amount of attention for implementing HCI by utilising the visual modality. Compared to gender and age recognition, it reveals more personal information by estimating the attention and intention of an individual. With eye/gaze analysis, a HCI system can not only observe its user passively, but it allows its user to take control of the system actively with eye movement. Therefore eye/gaze analysis excels in remote and contactless interaction and provides an ideal channel for elderly people and those with motor disabilities to access HCI systems.

According to the features extracted, eye centre localisation methods fall into two main categories: inherent feature based methods and additive feature based methods. An additive feature based method actively projects infrared illumination toward the eyes that result in reflections on the corneas, which are referred to as 'glints' in the literature (Zhu and Ji, 2005). Being highly reliant on dedicated devices, this method essentially alters the primary task of eye centre detection into corneal reflection detection as a simplified detection task. A passive inherent feature based method is more generalizable since it employs characteristic features from the eye region itself and therefore becomes the method we explore in this paper. It can be further divided into 1) eye geometry or morphology based methods that utilise gradient, isophote or curvature features to estimate the eye centre that comply with geometrical or morphological constraints, 2) model and machine learning based methods where distinct features are extracted to train a model to search for the eye region that best matches the model representation, and 3) hybrid methods which normally follow a multi-stage scheme that comprises the previously summarised. While several methods have achieved interesting results, they have also exhibited their respective limitations.

One geometrical feature based method (Timm and Barth, 2011) localises eye centres by means of gradients. In this approach, the iris centre obtains the maximised value in the objective function that peaks at the centre of a circular object. Its performance declines in the presence of strong gradients from eyelids, eyebrows, shadows and occluded pupils that overshadow iris contours. Another unsupervised method employing geometrical features that was investigated is Self-Similarity Space. Here image regions that can maintain peculiar characteristics under geometric transformations receive high self-similarity scores (Leo et al., 2014).

Regarding model and machine learning based methods, for all algorithms that utilise extracted features to train a model, it holds that the training data are of critical influence on the performance of the algorithms (Zhu and Ramanan, 2012). More specifically, variations posed by illumination and head rotation have a huge impact on the accuracy and robustness of the algorithm for most types of features. Inspired by Fisher Linear Discriminant (FLD) (Duda et al., 2012), (Kroon et al., 2008) designed a linear filter trained by the image patches extracted from normalized face images. This method not only considers the high response from the filtered image, but also examines a rectangular neighbourhood around the estimated eye centre positions. This is based on the observation that a pupil in an image is formed by a collection of dark pixels within a small region. Another machine learning based method (Niu et al., 2006) focuses on the design of a novel classifier rather than the extraction of representative features. This method introduces a 2D cascade AdaBoost classifier that combines bootstrapping positive samples and bootstrapping negative samples (Viola and Jones, 2001). The final localisation of an eye can be achieved either by weighting all classifier results for high precision or by cascading all classifiers (i.e. only adopting the result from the first classifier that detects an eye window) for increased efficiency. In addition, a number of studies are only effective for frontal faces. For example, (Asadifard and Shanbezadeh, 2010) employed a cumulative distributed function (CDF) for adaptive centre of pupil detection on frontal face images. Their approach firstly extracts the top-left and top-right quarters of a face image as the regions of interest and then filters each region of interest with a CDF. An absolute threshold is defined for the filtering process given the fact that the pixels in the pupil region are darker than the rest of the eye region. Another study on frontal faces (Türkan et al., 2007) explored edge projections for eye localisation. With a face image available, their method firstly defines a rough horizontal position for the eye region according to facial anthropometric relations. After the eye band is cropped, it gathers eye candidate points that are extracted by a high-pass filter of a wavelet transform. A Support Vector Machine (SVM) based classifier (Chang and Lin, 2011) is then used to estimate the probability value for every eye candidate. This type of method normally requires that all face images are perfectly aligned so that the facial geometry agrees with facial anthropometric relations as the prior knowledge.

Although recent studies have shown promising results in accurately localising the eye centres, the estimation error increases at relatively long distances and is also affected by shadows and specularities. (Drewes et al., 2007) carried out a study on eye-gaze interaction for mobile phone use following two methods, the standard dwell-time based method and the gaze gesture method. This study concludes that gaze gesture is robust to head movement since it only captures relative eye movement rather than absolute eye fixation points. Calibration is also unnecessary and this therefore makes eye gesture more suitable for real-world applications. The two interaction methods are further compared by (Hyrskykari et al., 2012) which suggested that "gaze gestures are not only a feasible means of issuing commands in the course of game play, but they also exhibited performance that was at least as good as or better than dwell selections". Another study (Rozado et al., 2012) achieved gaze gesture recognition for HCI under more general circumstances. It employs the hierarchical temporal memory pattern recognition algorithm to recognise predefined gaze gesture patterns. 98% accuracy is achieved for 10 different intentional gaze gesture patterns. Some other works on gaze gestures dedicated to HCI have similar limitations. Firstly, they all depend on active NIR lighting for eye centre localisation. Secondly, the eye centre localisation algorithms work at only relatively short distance.

3. Fisher Vector for gender recognition

As seen from the literature, most methods regarding gender recognition and gaze analysis lack robustness and suffer from various limitations in real-world scenarios. To bridge these gaps, we explore novel methods that can assist HCI implementations robustly and efficiently while maintaining high accuracy. In this section, we introduce a gender recognition method that utilises Fisher Vectors as discriminative features.

3.1. 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) (Reynolds, 2009)

$$p(x|\lambda) = \sum_{i=1}^{B} \omega_i \, \mathcal{N}(x|\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) \tag{1}$$

where $\boldsymbol{x}$ is a L-dimensional data vector, $\lambda = \{\omega_i, \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i, i = 1, 2, \ldots, B\}$ 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. B is the number of Gaussians. The component $\mathcal{N}(x|\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)$ is further described in Eq. (2).

$$\mathcal{N}(x|\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) = \frac{e^{\left\{-\frac{1}{2}(x-\mu_i)'\sigma_i^{-1}(x-\mu_i)\right\}}}{(2\pi)^{L/2}|\boldsymbol{\sigma}_i|^{1/2}} \tag{2}$$

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

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

The covariance matrices are assumed to be diagonal since any distribution can be decomposed into a number of weighted Gaussians with diagonal covariances.

Let $\boldsymbol{X} = \{\boldsymbol{X_t}, t = 1, 2, \ldots, T\}$ 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:

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

A natural kernel on these gradients is:

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

where $\mathcal{F}_\lambda = \mathcal{L}_\lambda{}' \mathcal{L}_\lambda$ is the Fisher information matrix and $\psi_\lambda^X = \mathcal{L}_\lambda \, \psi_\lambda^X$ is referred to as the Fisher Vector of $\boldsymbol{X}$.

Let $\gamma_t(i)$ denotes the soft assignment of descriptor $\boldsymbol{X_t}$ to the Gaussian component i:

$$\gamma_t(i) = p(i|x_t, \lambda) = \frac{\omega_i \mathcal{N}(x_t|\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)}{\sum_{j=1}^{B} \omega_j \mathcal{N}\left(x_t|\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j\right)} \tag{7}$$

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

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

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

Finally a FV is represented as:

$$\Phi = \left\{\Psi_{\mu,\,1}^X, \; \Psi_{\sigma,\,1}^X, \; \ldots, \; \Psi_{\mu,\,N}^X, \; \Psi_{\sigma,\,N}^X\right\} \tag{10}$$

Therefore, when images from a certain database produce a large number of feature descriptors, a GMM can be trained by them using the Maximum Likelihood (ML) estimation. The difference between every individual descriptor and every Gaussian distribution is captured by Eq. (8) and (9), which is further stacked into a FV by Eq. (10). Following this manner, a FV can be computed for every descriptor, which encodes the difference between a single descriptor and all descriptors described by the GMM. This means that the difference between an image and the entire training dataset is preserved. As a result, a FV is provided with contextual definition and enhanced saliency for classification.

3.2. Fisher Vector encoding for gender recognition

Fisher Vectors have been used for face recognition and have proved to be an excellent encoding method (Simonyan et al., 2013). The FV encoding approach consists of five main stages: 1) face pre-processing, 2) low-level feature and face descriptor computation, 3) dimensionality reduction/feature selection, 4) FV encoding and 5) classifier training. In more detail, we illustrate the five primary stages adopted by our method as follows.

(1) Face pre-processing.

The techniques experimented at this stage include face detection, image resizing, histogram equalisation and face alignment. In the experiment, the Viola-Jones face detector was used to obtain the face region in the first place. As well as aiming to achieve high recognition accuracy, this study intends to investigate the most discriminative facial parts, i.e. regions on a face that can best characterise and differentiate male and female groups. To this end, we reshaped the face regions obtained by the face detector so that they could incorporate the hair region and the chin. Practically, this was managed by excluding the background margins from the width by a ratio of rw and expanding the length by a ratio of rl. Let the side of a square region detected by the face detector be fs, the length of a reshaped face region be fl and the width be fw. A reshaped face region can be defined as:

$$\begin{cases} fw = fs \times (1 - rw) \\ fl = fw \times (1 + rl) \end{cases} \tag{11}$$

where $rw = 0.12$ and $rl = 0.33$ in our experiments. This ensures that the reshaped face region has a uniform aspect ratio (i.e. 3:4), while keeping its size relative to that originally detected by the face detector. Facial regions were further resized to the same size in the next step with or without histogram equalisation. Face alignment was also experimented with and its impact was investigated.

An optional step at this stage is face alignment, which is intended to compensate for different head poses that are likely to sabotage most gender recognition algorithms. To evaluate the robustness of our method against different head poses, we conducted experiments (see Section 3.3) with or without face alignments. Eye centre coordinates obtained by the proposed eye centre localisation method (introduced in Section 4) were employed as facial landmarks for the alignment. However, eye centres accurately localised by other methods or other types of facial landmarks (e.g. eye corners and nose tip) should also suffice to perform the alignment. Fig. 1 further illustrates the face pre-processing stage with an example.

(2) Low-level feature and face descriptor computation.

Conventionally, one face image produces only one feature vector, namely a descriptor (e.g. the LBP features) or a few descriptors around keypoints (e.g. the SIFT features). In both cases, feature descriptors are sparsely extracted. Different from both methods, 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 $r \times r$ window across an image horizontally and vertically with a predefined stride s ($s \in \mathbb{Z}$). One descriptor per patch rather than one descriptor per image is calculated. The geometry of the patch-based descriptors is shown in Fig. 2. For example, vector (px_c, py_c) records the centre position of the a^{th} ($a \in \mathbb{N}$, $a \leq pn$) patch in the image. The centre position of the first patch is therefore $(r/2, r/2)$. For an $m \times n$. image, the total number of patches is:

$$pn = \frac{m - r + 1}{s} \times \frac{n - r + 1}{s}, \qquad \begin{matrix} r < \min(m, n) \\ 1 \leq s \leq \min(m, n) \end{matrix} \tag{12}$$

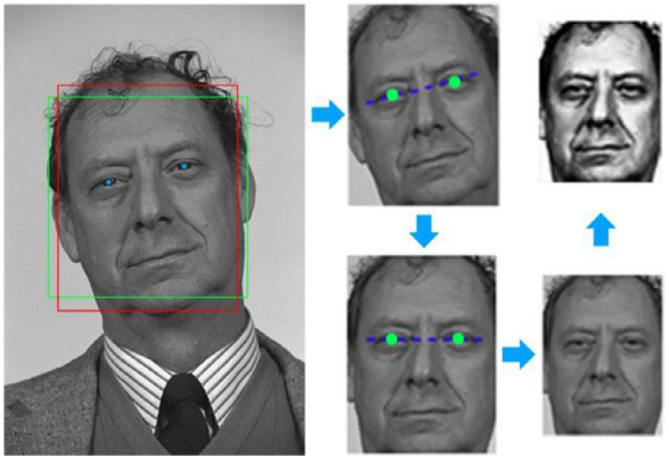


Fig. 1. An illustration of face pre-processing, including face detection, face alignment (optional), and histogram equalisation (optional). The two boxes in solid lines around the face area represent the detected facial region (square) and the reshaped facial region (rectangular), respectively.

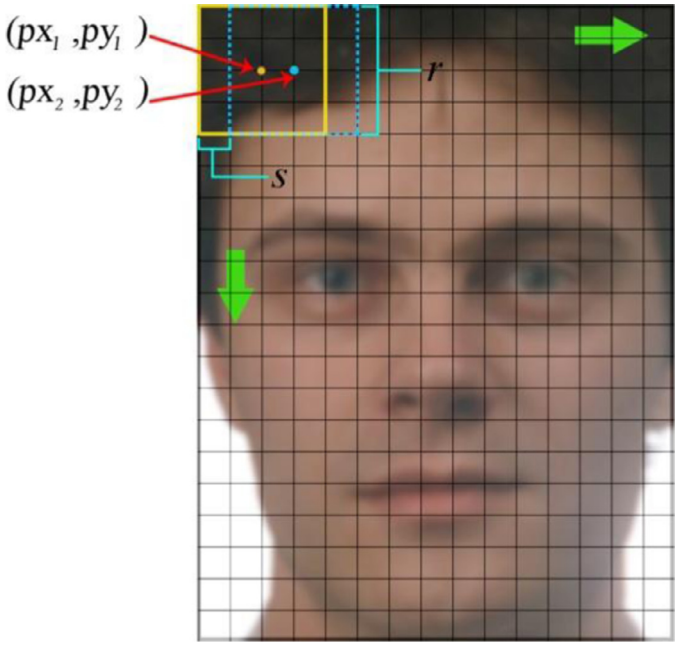


Fig. 2. Geometry of patch-based dense feature descriptors.

Although a number of pixels at the margins of an image cannot be the centres of a sliding window, they are incorporated by at least one image patch and are therefore involved in the formation of image descriptors. Low-level dense features can be visualised by mapping the top three principal components to the RGB space. A similar process can be found in (Liu et al., 2011). As an example, dense SIFT features extracted with different parameters can be seen in Fig. 3.

The visualisation implies that facial structures can be captured by densely sampled local features. It should be noted that this visualisation only reflects the top three principal components of the SIFT vectors which originally had 128 dimensions. Visualisation of low-level dense features under a particular colour space is only a feature-to-colour mapping process and therefore has no impact on the actual feature used for gender recognition.

(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 presentations.

In our experiments, Principal Component Analysis (PCA) was implemented to reduce the dimensionality to 64 features per descriptor. As Fig. 3 demonstrates that the top 3 principal components of SIFT features can capture the main facial structure, 64 principal components should be able to reflect sharp edges and fine facial structures. By reducing the dimensionality of facial descriptors, highly correlated information that does not contributed to discriminability could be removed. In order that a descriptor can be located on the original image after being encoded into a FV, the centre position of a patch was appended to the end of a corresponding compressed descriptor in the form of a 2D vector (Simonyan et al., 2013), increasing the dimensionality to 66 from 64. This 2D vector is rescaled to [−0.5, 0.5] so that it would have minimal effect compared to the other 64 features. As a result, although spatial information is embedded into feature vectors, the overall features are still relatively independent of global facial geometry and are therefore robust to head pose variation.

(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 (Vedaldi and Fulkerson, 2010) for GMM training and SIFT feature extraction. Among GMM parameters, different numbers of Gaussian components were experimented with, amongst which 512 was deemed a suitable value (see results in Section 3.3 for more details). It can be calculated from Eq. (10) that the dimension of a FV is $2 \times B \times L = 2 \times 512 \times 66 = 67584$. In this stage, decomposed patches are reunited to characterise a complete image, in the form of derivatives of all Gaussian components. In the computation of FVs, ℓ_2 normalisation and power normalisation are applied to the vectors since they were reported to improve classification performance (Perronnin et al., 2010).

(5) Classifier training

Our algorithm learns a SVM classifier from all training FVs. In our experiments (see Section 3.3), a linear SVM and a SVM with a Radial Basis Function (RBF) kernel were both tested. By comparing their respective classification rate, we show that the FV encoding method only requires a linear SVM to function accurately and robustly. The employment of a SVM classifier has a twofold purpose. Firstly, it naturally fulfils the classification task by defining a hyper-plane, with or without a kernel. Its output gives the predicted labels for all testing images. Secondly, when a linear SVM is concerned, it reveals the discriminative power of each variable/feature in FVs. From the perspective of a SVM classifier, a SVM learns a hyper-plane that separates two classes with maximum margin. As the hyper-plane is defined by a decision hyper-plane normal vector $\vec{\boldsymbol{w}}$ that is perpendicular to the hyper-plane, as well as an intercept term b, the absolute values of the elements in $\vec{\boldsymbol{w}}$ imply the significance of the corresponding elements in the FVs. From the perspective of metric learning, the diagonal linear transformation matrix $\boldsymbol{W}$ to be learnt has its diagonal values as in $\vec{\boldsymbol{w}}$. It can be considered as projecting the original data so that they are located on each side of a *fixed* hyper-plane, different from the former perspective where the data are fixed and the hyper-plane is unknown (Do et al., 2012). Both methods state that $\vec{\boldsymbol{w}}$, also known as the weight vector, reflects the discriminative power of individual features.

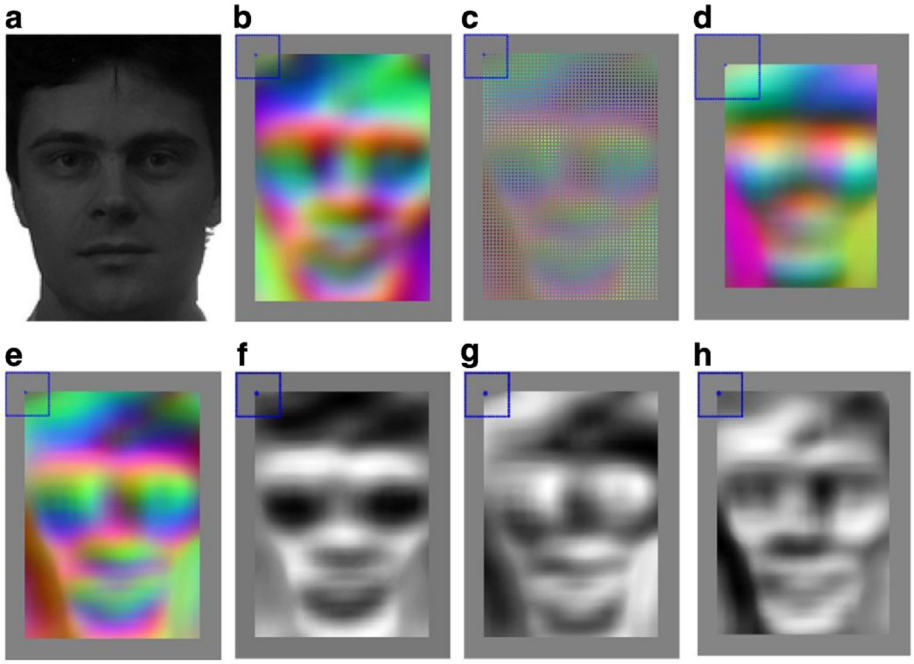


Fig. 3. Dense SIFT feature visualisation in the RGB colour space. The square on the top left corner denotes the size and centre of the first descriptor patch. (a) The original facial image. (b) Visualisation for $\mathbf{r} = 24$ and $\mathbf{s} = 1$, with histogram equalisation and root SIFT applied. (c) Visualisation for $\mathbf{r} = 24$ and $\mathbf{s} = 2$. (d) Visualisation for $\mathbf{r} = 36$ and $\mathbf{s} = 1$. (e) Visualisation for $\mathbf{r} = 24$, $\mathbf{s} = 1$. (f) Visualisation of the first principal component for $\mathbf{r} = 24$, $\mathbf{s} = 1$. (g) Visualisation of the second principal component for $\mathbf{r} = 24$, $\mathbf{s} = 1$. (h) Visualisation of the third principal component for $\mathbf{r} = 24$, $\mathbf{s} = 1$.

From the manner in which a FV is constructed (Eq. (10)), a mapping can be obtained between the features and the Gaussian components. Therefore the relationship between the discriminative power and each Gaussian component can be established. When visualised with regard to its spatial location, a Gaussian component can be used to signify the discriminability of a corresponding facial region.

3.3. Gender classification experiments and results

In order to evaluate the performance of our gender recognition algorithm under controlled environments and real-world conditions respectively, the Grey FERET database (Phillips et al., 1998) (referred to as the FERET database in the rest of the paper), the Labelled Face in the Wild (LFW) database (Huang et al., 2007) and the FRGCv2 database (Phillips et al., 2005) are employed.

The FERET database consists of 14051 greyscale images of frontal and profile face images. It is further divided into several partitions where the *fa* and *fb* partitions contain near-frontal facial images. Since the two partitions significantly overlap, only the *fa* partition of 1152 male patterns and 610 female patterns were employed. Although captured under controlled environment, the FERET database still poses great challenge as it accommodates different ethnicities, facial expressions, facial accessories, facial makeup and illumination conditions. The LFW database is considered one of the most challenging databases and has become the evaluation benchmark for face recognition under unconstrained environment. The 13233 colour facial images of 5749 subjects collected from the web include all types of variation and interference (illumination, head poses, occlusion, image blur, chromatic distortion, etc.), and come in inconsistent image quality. Only one image per subject (the first image) was used in the experiment so that the same subject could not appear in both the training set and the testing set. The FRGCv2 database includes 4007 depth images belonging to 466 subjects. These data also include different ethnicities and age groups. Some sample images from the three databases are shown in Fig. 4.

Fig. 4. Sample images from (a) the FERET database ***fa*** partition (greyscale), (b) the LFW database (converted to greyscale) and (c) the FRGCv2 database (depth).

The 5-fold cross validation technique was adopted for evaluation. More specifically, the database was partitioned into five splits of similar size, four of which were used for training in each repetition with the remaining 1 split for testing. After 5 repetitions, the average classification rate was calculated as the final result.

In the first group of experiments, different types of features are explored including the greyscale values, the SIFT features, the LBP features (extracted using the circularly symmetric neighbour sets (Ojala et al., 2000) with 8 neighbouring pixels and radius of 1),

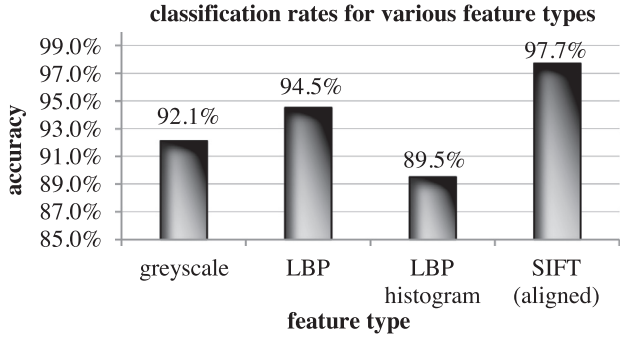


Fig. 5. Gender classification rates for various feature types, tested on the FERET database.

and LBP histogram with uniform pattern. Evaluated on the FERET database, their respective performances are summarised in Fig. 5.

The size of facial images is 160 × 120, the sampling stride is four pixels, the window size for local patches is 24 × 24 and the number of Gaussian components in the GMM is 512. The classification results suggest that dense SIFT features perform the best. Therefore this feature type was selected for a further inspection regarding a number of parameters. Note that our dense SIFT features were only extracted at one scale since we did not observe higher accuracy when multiple scales were used. In addition, we only employed a linear SVM since the RBF kernel in the experiments did not contribute to higher classification rate.

In the second group of experiments, parameters inspected concerned image size, size of the sliding window, sampling stride and Gaussian number. One of these parameters was altered at a time while the others remained fixed as in the last group of experiments (if not otherwise specified). Misaligned and non-normalised face images were used in this group of experiments.

(1) Image size

The maximum size of selected face regions from the FERET database is 160 × 120. With the aspect ratio fixed, each face region was resized to between 30% and 90% of the maximum size, with an interval of 10%. Fig. 6(a) reflects the declined performance as the size decreases. The results suggest that larger images produce higher classification rates. This coincides with the intuition that more details reside in larger images which should generate more effective features for classification.

(2) Sampling stride for local patches

Increasing the sampling stride in extracting local descriptors is comparable to decreasing the sampling rate in sub-sampling an image. Fig. 6(b) summarises the impact of sampling stride. To avoid high memory usage, we firstly employed all the facial images resized to 96 × 72. Strides from 2 to 7 were experimented with in this setting. We then employed the first 500 male and female images (1000 in total) and further resized them to 80 × 60, in order that experiments with stride of 1 could be conducted. It can be seen from Fig. 6(b) that although a smaller stride tends to produces a higher classification rate in general, it will only have a dramatic influence when it goes beyond four pixels. Therefore, a stride of 3 or 4 can be deemed an appropriate value without incurring high memory usage at the training stage.

(3) Window size of local patches

Our algorithm is appearance-based and densely extracts descriptors at local level. Broadly speaking, a larger window implicitly maintains more global and geometric information. An extreme case concerns a window as large as the entire image which removes the advantage of using local features. Hence there is a critical need for the window to be defined with an appropriate size that allows sufficient tolerance to global variations. The window size ranged from 12 × 12 to 40 × 40 pixels in the experiments, with an interval of four pixels. The respective classification rate is summarised in Fig. 6(c). It can be concluded that a window 15% to 20% the size of the entire image is optimal.

(4) GMM component number

Fitting a generative model, the GMM, to the features is a key step in this method. Fig. 6(d) shows that the classification rate fluctuates as the number of Gaussians changes. The size of the facial images used was 128 × 96 (80% of the maximum size). The result agrees with the statement (Sánchez et al., 2012) that an appropriate number of Gaussian components is needed since too many components result in very few 'per Gaussian' statistics that are pooled together, i.e. sparse Gaussian representation. Too few Gaussians, on the other hand, are not sufficient enough to capture the uniqueness and reflect the separability of local descriptors.

(5) Face alignment, histogram equalisation and root SIFT

Face alignment has been reported in the literature to have a substantial impact on classification rates. Without alignment, one may experience a drop in classification rates as much as 6% (Lee et al., 2010). In addition, applying histogram equalisation is a common pre-processing procedure that may increase classification rate. When the SIFT features are concerned, a variation of the SIFT, the root SIFT is recommended by (Arandjelovic and Zisserman, 2012), which uses the Hellinger kernel instead of the standard Euclidean distance to measure the similarity between SIFT descriptors. The impacts of the three factors were explored in this experiment, summarised in Fig. 7.

It should be noted that each time only one type of adjustment was made so that their impacts could be evaluated independently from other factors. As seen from Fig. 7, only face alignment plays a positive role while the other two types of adjustment decrease the classification rate. It can be claimed that the proposed algorithm is relatively robust against misalignment which caused only a 0.6% drop in the classification rate. In addition, the impact of PCA was also explored by reducing the original facial descriptors to dimensionality of 32, 64 and 96, respectively. When the first 64 principal components are employed, the highest classification rate (97.7%) can be achieved, in comparison to 96.4%, 97.2% and 95.7% for 32, 96 and 128 principal components, respectively.

As SIFT features yielded the highest classification accuracy, it was further tested on the LFW database (real data) and the FRGCv2 database (3D data). The optimal parameters identified by the previous experiments were employed in an attempt to achieve the highest classification result, except that the facial images were resized to 112 × 84 for the LFW database and 240 × 180 for the FRGCv2 database. The classification rates for aligned and misaligned images in the LFW database are 92.5% and 92.3%, respectively; the classification rate for misaligned images in the FRGCv2 database is 96.7%.

The endeavour to align faces in the LFW database sees very limited benefit (0.2% increase) and is therefore not worthwhile, since in this case and in general cases, it incurs large amount of computation. The reason behind the reduced improvement brought by face alignment on this database, compared to the 0.6% increase on the FERET database, is possibly the large extent of out-of-plane head rotation that cannot be compensated by conventional alignment algorithms.

It is stated in Section 3.2 that during the FV encoding process, spatial information of every image patch is implicitly embedded in

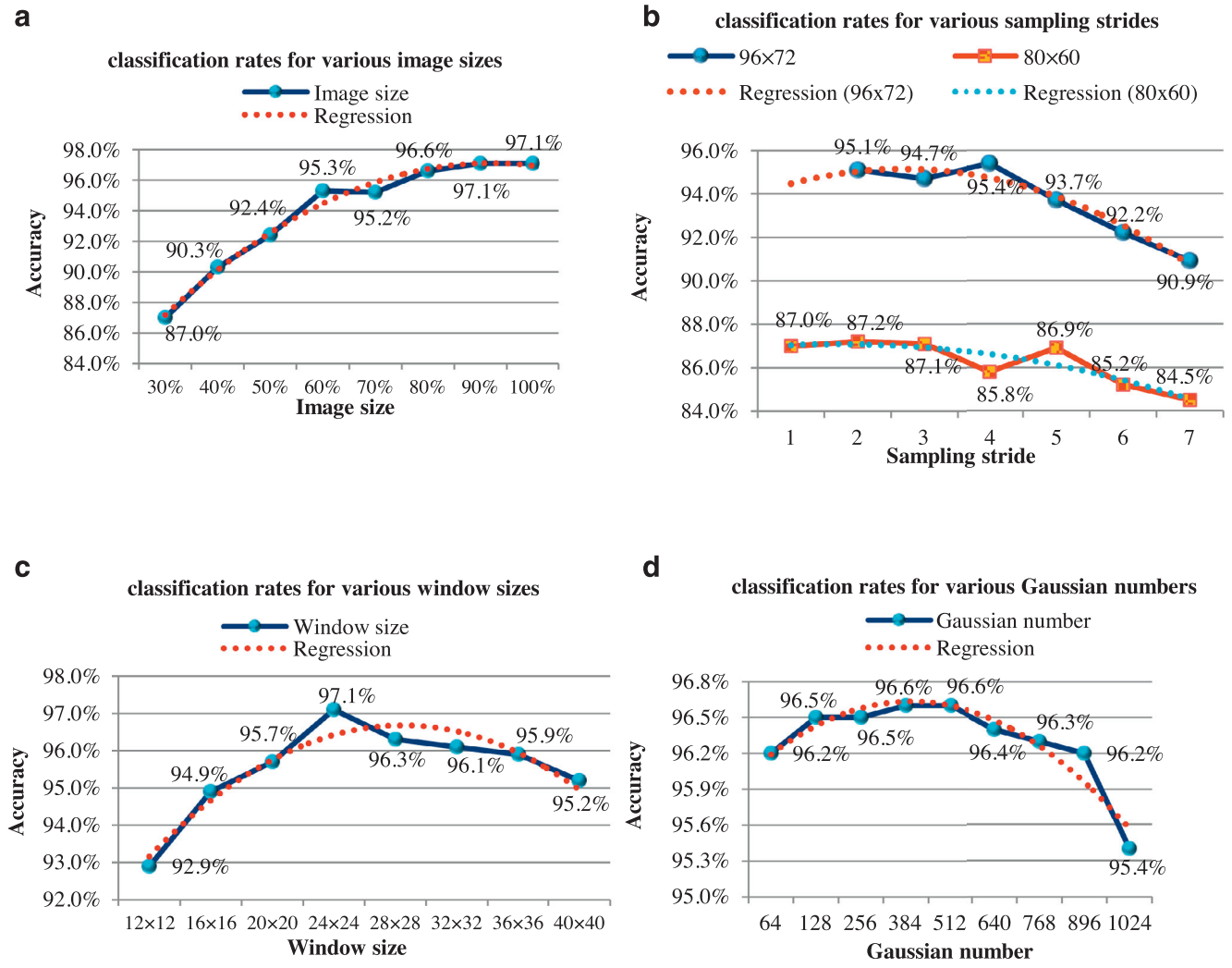


Fig. 6. Gender classification rates for various algorithmic parameters, tested on the FERET database. The fluctuation trend of the classification accuracy in each subfigure is described by a quadratic polynomial regression fit, represented by the dotted line. (a) Gender classification rates for various image sizes. (b) Gender classification rates for various sampling strides. (c) Gender classification rates for various window sizes. (d) Classification rates for various Gaussian numbers.

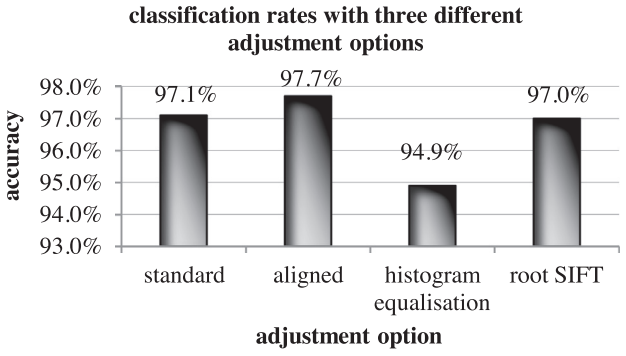


Fig. 7. Gender classification rates with or without face alignment, histogram equalisation and root SIFT implementation, tested on the FERET database.

a GMM. Therefore it is possible to restore the spatial coordinates from the Gaussian means (where the last two variables stand for the x and y coordinates). Similarly, the spatial variances can be restored from the Gaussian covariances, which indicate how well the spatial coordinates can represent individual patch locations. By visualising the Gaussians that correspond to the most discriminative image patches, it is possible to localise the facial regions that are most powerful in distinguishing male and female faces. Visualisation of facial discriminability is shown in Fig. 8. Note that the face image displayed is only a generic representation of facial geometry and therefore is not gender-specific.

The first visualisation format (Fig. 8(a)) shows that the most discriminative Gaussians agree with intuitive feature points (e.g. edges and corners), and therefore are deemed a suitable representation of face appearance. The Gaussians are further visualised in the form of dense image patches whose rankings in the discriminative power are indicated by the numbers centred at each patch (Fig. 8(b)). It can be noticed that the patches overlap significantly, making it difficult to distinguish the most discriminative ones. One simple solution is to construct an energy map for all pixel locations, stacking up all the patches in the previous format. The energy map agrees with the following two rules: 1) image patches with higher rankings hold more energy; 2) A pixel location where overlap exists draws energy from all the patches that cause this overlap. It can be seen that the mouth region (where male adults may have beard and moustache), the nasolabial furrows and the forehead region (where females are more likely to have bangs) are the most discriminative. This result provides insight into region-based facial discriminability so that future

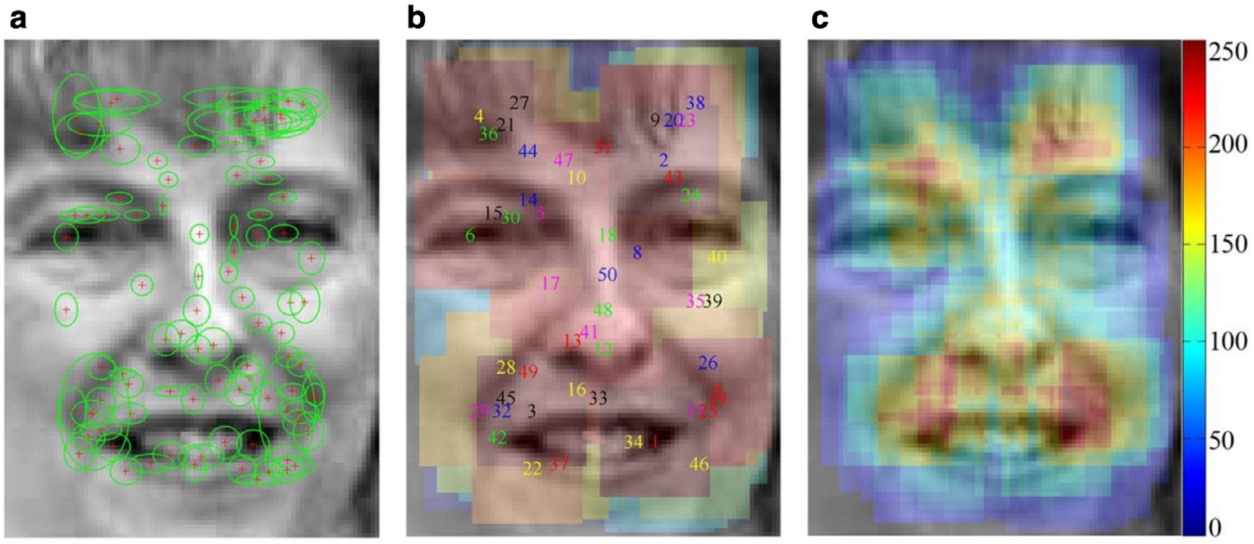


Fig. 8. Visualisation of facial discriminability with regard to gender classification. (a) The top 128 Gaussians. (b) Facial patches at top 50 Gaussian locations. (c) The energy map for facial discriminability.

Table 1
Comparison of the proposed method to eight other state-of-the-art methods. The ∗ notation indicates that training images and testing images may contain the same subjects.

Method in comparison	Features and classifiers employed	Database(s)	Evaluation method	Accuracy	Limitation(s)
proposed	*FVs; linear SVM*	*FERET fa 1762*	*5-CV*	*97.7%*	*high memory consumption at training stage*
		LFW 5749		*92.5%*	
		FRGCv2 depth all 466 subjects		*96.7%*	
(Mäkinen and R., 2008)	greyscale & LBP; classification fusion	FERET fa+fb 900	5-CV	92.9%	small database size
(Moghaddam and Yang, 2000)	greyscale; SVM (Gaussian RBF)	FERET thumbnails 1855	5-CV	96.6%	only intensity values are used as features ∗
(Shan, 2012)	refined LBP histogram; SVM (RBF)	LFW 7443	5-CV	94.8%	challenging images (about half the size of the database) are manually removed
(Ullah et al., 2012)	dyadic wavelet transform & LBP;	FERET fa+fb 2400	falf images for training and half for testing	99.3%	∗
	Minimum-distance classifier	Multi-PIE 1990		99.1%	manual data selection
(Lee et al., 2010)	Facial strips; ε-SVR & C-SVC	FERET fa 1763	not specified	98.8%	11 classifiers needed; perfect face alignment needed; 92.8% accuracy for misaligned faces; evaluation method not specified
(Tivive and Bouzerdoum, 2006)	Greyscale;	FERET fa 1762	5-CV	97.2%	∗
	convolutional neural networks	Web		83.7%	
(Phung and Bouzerdoum, 2007)	greyscale;	FERET fa 1762	5-CV	96.4%	∗
	convolutional neural networks				
(Wang and Kambhamettu, 2013)	LBP & Shape Index; SVM (Gaussian RBF)	FRGCv2 depth all 466 subjects	5-CV	93.7%	∗

research on gender recognition can better target on facial regions with high discriminative power. To further validate our method, we compare the accuracy of our method to eight other state-of-the-art methods with an evaluation of their respective limitations, detailed in Table 1.

The accuracy of the proposed method is only slightly lower than (Ullah et al., 2012) and (Lee et al., 2010) on the FERET database and (Shan, 2012) on the LFW database, but outperforms the others in comparison. However it should be noted that (Ullah et al., 2012) used both *fa* and *fb* subsets of the FERET database containing replication of most subjects, while (Lee et al., 2010) used a large number of classifiers and did not specify the evaluation method. Using images of the same subjects for both training and testing should be avoided in an objective evaluation process. On the other hand, (Shan, 2012) removed images with large head rotation and those with ambiguous ground truth. This imposed

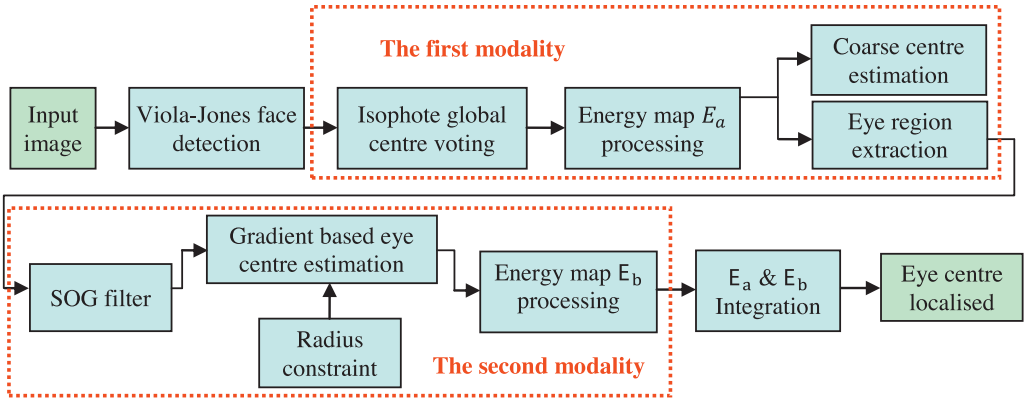


Fig. 9. An overview of the eye centre localisation algorithm chain.

manual intervention on the database and selected only the 'good' data that were more frontal and easier to classify. It can be therefore concluded that the accuracy of the proposed gender classificaiton algorithm is among one of the highest in the literature, tested on both *controlled* and *uncontrolled* databases. The high accuracy of gender classification brought by our study allows a HCI system to understand human charactersitics. As a result, assistive systems are enabled to create user-centred HCI environments that bring more comfort and naturalness to users. In addition, the high robustness and efficiency ensure that it is applicable to practical applications rather than being limited to theoretical studies.

4. Eye centre localisation – a modular approach

While our gender recognition method accurately and robustly gathers demographic data such that assistive HCI systems can better understand human characteristics, a hybrid method is further proposed that can perform accurate and efficient localisation of eye centres in low-resolution images and videos in real time. Therefore demographic data from gender recognition and behavioural data from eye/gaze analysis can be fused by assistive technologies in order to gain higher intelligence and interactivity. Our algorithm is summarised in Fig. 9 as an overview of the eye centre localisation chain.

The algorithm includes two modalities. The first modality performs a global estimation of eye centres over a face region detected by the Viola-Jones face detector (Viola and Jones, 2004) and extracts corresponding eye regions. Results from the first modality are fed into the second modality as prior knowledge and lead to a local and more precise estimation of eye centres. The two energy maps generated by the two modalities are fused to produce final estimation of eye centres.

4.1. Isophote based global centre voting and eye detection

Human eyes can be characterised as radially symmetrical patterns which can be represented by contours of equal intensity values in an image, i.e. isophotes (Lichtenauer et al., 2005). Due to the large contrast between the iris and the sclera as well as that between the iris and the pupil, the isophotes that follow the edges of the iris and the pupil reflect the geometrical properties of the eye. Therefore centre of these isophotes will be able to represent estimated eye centres. One study (Valenti and Gevers, 2008) proposed an isophote based algorithm that enables pixels to vote for isophote centres they belong to. The displacement vector pointing from a pixel to its isophote centre follows Eq. (13).

$$\{D_x,\ D_y\} = -\frac{\{I_x,\ I_y\}\left(I_x^2 + I_y^2\right)}{I_y^2 f_{xx} + 2I_x I_{xy} I_y + I_x^2 I_{yy}} \tag{13}$$

where I_x and I_y are first-order derivatives of the luminance function $I(x, y)$ in the x and y directions. I_{xx}, I_{xy} and I_{yy} are the second-order partial derivatives. The weight of each vote is indicated by the curvedness of an isophote since the iris and pupil edges that are circular obtain high curvedness values as opposed to flat isophotes. The curvedness (Koenderink and Doorn, 1992) is calculated as:

$$cd(x, y) = \sqrt{I_{xx}^2 + 2 \times I_{xy}^2 + I_{yy}^2} \tag{14}$$

We also consider the brightness of the isophote centre in the voting process based on the fact that the pupil is normally darker than the iris and the sclera. Therefore, an energy map $E_a(x, y)$ is constructed that collects all the votes to reflect the eye centre position following Eq. (15), where α is the maximum greyscale in the image ($\alpha = 255$ in the experiments).

$$E_a(x + D_x,\ y + D_y) = [\alpha - I(x + D_x,\ y + D_y)] \times cd(x,\ y) \tag{15}$$

Though isophotes have been employed in the literature, related works have only extracted isophote features from eye regions which are either cropped according to anthropometric relations (which are interrupted by head poses) or found by an eye detector (which largely increases the complexity of algorithm). Our method is different, in that it extracts isophote features for a *whole face* and constructs a global energy map $E_a(x, y)$. The energy points below 30% of the maximum value are removed. The remaining energy points therefore become the new eye centre candidates that are fed to our second modality for further analysis. $E_a(x, y)$ is then split into the left half $E_{aul}(x, y)$ and the right half $E_{aur}(x, y)$, corresponding to the left and right half of the face, where the mouth region (the lower half of the energy map) is simply removed since it is unlikely to concern any eye information regardless of normal head poses. We further calculate the energy centre, i.e. the first moment of the energy map divided by the total energy, which is selected instead as the optimal eye centre. Taking $E_{aul}(x, y)$ as an example, this can be formulated as Eq. (16)

$$\{\mathrm{cx}_{aul},\ \mathrm{cy}_{aul}\} = \frac{\sum_{x=1}^{m}\sum_{y=1}^{n}\{x,\ y\} \cdot E_{aul}(x,\ y)}{\sum_{x=1}^{m}\sum_{y=1}^{n} E_{aul}(x,\ y)} \tag{16}$$

where $C_{aul} = \{\mathrm{cx}_{aul},\ \mathrm{cy}_{aul}\}$ is the optimal estimation of the left eye centre, m and n are the maximum row and column number in E_{aul}. The eye region to be analysed by our second modality is then selected, which centres at the optimal eye centre estimation (its width being 1/10 of the face size and its height being 1/15 of

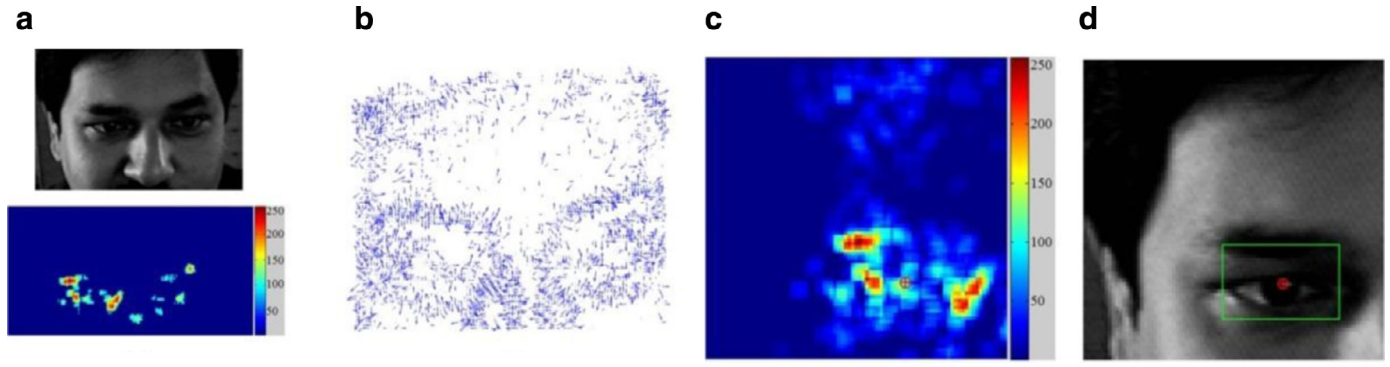


Fig. 10. An illustration of the first modality for eye centre estimation and eye detection. (a) An image of an upper face (BioID., 2014) and the corresponding global energy map calculated by Eq. (15). (b) Displacement vectors calculated by Eq. (13). (c) An enlarged view of the top-left quarter of the energy map, where the energy centre is found by Eq. (16). (d) The optimal eye centre position found and the eye region selected.

the face size). As a result, our method does not require an eye detector and is robust to head rotations since global isophotes are investigated. This process is shown in Fig. 10.

4.2. Gradient based eye centre estimation

The first modality performs the initial eye centre estimation, filtering eye centre candidates and selecting local eye regions for the second modality. Based on an objective function (Eq. (17)), we further design a radius constraint and a Selective Oriented Gradient (SOG) filter to re-estimate the eye centre positions with enhanced accuracy and reliability.

The radially symmetrical patterns of an eye generate isophotes around the iris and pupil edges that can effectively vote for the eye centre. When simply modelled as circular objects, the iris and the pupil can produce gradient features that give an accurate estimation of the eye centre. This is based on the idea that the prominent gradient vectors on circular iris/pupil boundaries should agree with the radial directions and therefore the dot product of each gradient vector with its corresponding radial vector is maximised.

This is formulated by (Timm and Barth, 2011) as an objective function:

$$\boldsymbol{c}^* = \underset{\boldsymbol{c}}{\arg max} \left\{ \frac{1}{N} \sum_{x=1}^{m} \sum_{y=1}^{n} I_{\boldsymbol{c}}(x,\ y) \cdot (\boldsymbol{d}^t(x,\ y) \cdot \nabla I_{\boldsymbol{c}}(x,\ y))^2 \right\} \tag{17}$$

$$\boldsymbol{d}(x,\ y) = \frac{\boldsymbol{p}(x,\ y) - \boldsymbol{c}}{\| \boldsymbol{p}(x,\ y) - \boldsymbol{c}_2 \|}, \quad \forall x \forall y : \| \boldsymbol{d}(x,\ y) \|_2 = 1, \quad \| \nabla I_{\boldsymbol{c}}(x,\ y) \|_2 = 1 \tag{18}$$

where $\boldsymbol{c}$ is a centre candidate that remained from the isophote based modality, $\boldsymbol{c}*$ is the optimal centre to be calculated, N is the total number of pixels in an eye region to be analysed, $\boldsymbol{d}(x,y)$ is the displacement vector connecting a centre candidate $\boldsymbol{c}$ and a different pixel $\boldsymbol{p}(x,y)$. $\nabla I_{\boldsymbol{c}}(x,y)$ is the gradient vector and $I_{\boldsymbol{c}}$ is the intensity value for an eye centre candidate, meaning that the objective function only considers pixels that have been previously selected as eye centre candidates. The displacement vectors and gradient vectors are normalised to unit vectors. This objective function is modified such that a gradient vector is only considered if its direction is reverse to the displacement vector. This is based on the observation that the pupil is always darker than its neighbouring regions and thus generates outward gradients. A sample implementation of this approach on an eye image is illustrated in Fig. 11. We tested this modality on a number of images with varied head poses, gaze directions and partial eye/pupil occlusions. The

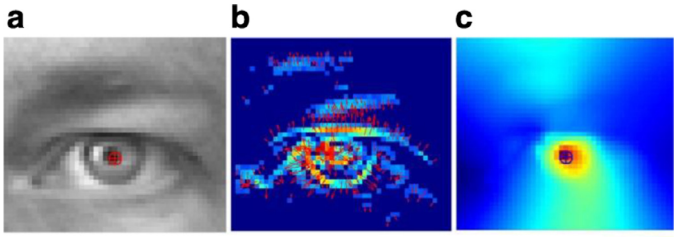


Fig. 11. A sample implementation of the second modality using local gradient features. (a) An eye image where the localised eye centre is displayed. (b) The gradient magnitude image where gradient directions are represented by arrows. Gradients with magnitude below 70% of the maximum are removed (c) The resulting energy map for eye centre candidates.

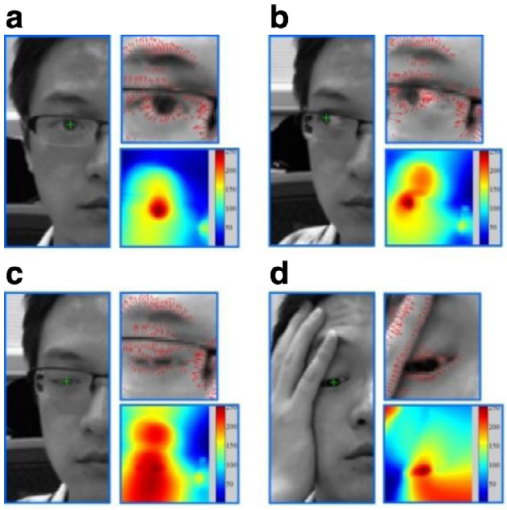


Fig. 12. Sample results generated by the second modality. (a) Results for a frontal face. (b) Results for a rotated face with an extreme pupil position. (c) Results for a half-closed eye. (d) Results partially occluded eye.

raw face images (right halves), gradients from the eye regions, and the corresponding energy maps are shown in Fig. 12.

While this method provides an effective solution to eye centre estimation, it has a number of inherent limitations that would incur error or even failure in the estimation. First of all, the objective function is formulated given that the pupil and the iris are circular objects. When the edges around eye corners and shadows surpass the pupil and the iris in circularity measure, they will cause eye centre estimates to be located on themselves rather than the centre of the pupil. Secondly, gradients on eyelids, eye corners and

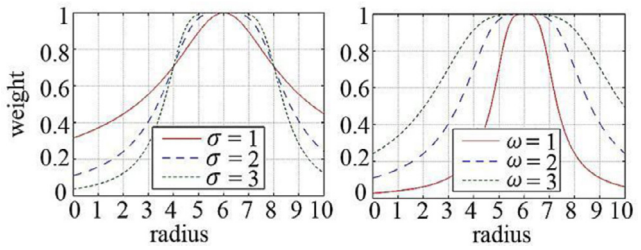


Fig. 13. The radius weight function curves emulating the frequency responses of a Butterworth low pass filter. (a) Curves with varying $\boldsymbol{\sigma}$ ($\boldsymbol{\sigma} = 1,\ 2,\ 3$) and constant $\boldsymbol{\omega}$ ($\boldsymbol{\omega} = 2$). (b) Curves with varying $\boldsymbol{\omega}$ ($\boldsymbol{\omega} = 1,\ 2,\ 3$) and constant $\boldsymbol{\sigma}$ ($\boldsymbol{\sigma} = 2$).

eyebrows will also vote for eye centre estimations. We define these as 'non-effective gradients', as opposed to 'effective gradients' that are located on edges of the iris and pupil. In more severe cases where makeup and shadows are prominent, the iris and the pupil, by contrast, generate weak 'effective gradients' such that the energy map is prone to erroneous energy response. This will further escalate estimation errors.

To resolve the above problems shared by most methods that utilise geometrical features for eye centre localisation, we introduce a radius constraint and design a SOG filter that can effectively deal with circularity measure and can resolve problems posed by eye corners, eyebrows, eyelids and shadows.

4.3. Iris radius constraint

A radius constraint is introduced such that the Euclidean norms of displacement vectors, which are related to estimated iris radius, have more influence on the calculation of an eye centre location. This is based on the assumption that shadows and eyebrow segments have random radius values, while the iris radii are more constant relative to the size of a face. This provides a way to differentiate circular clusters of various radiuses and to determine their weights in energy map accumulation. The function for the significance measure emulates the frequency response of a Butterworth low pass filter:

$$w_r(x,\ y) = \sqrt[2]{\frac{1}{1+\left(\frac{\|\boldsymbol{d}(x,\ y)\|_2-\rho}{\omega}\right)^{2\sigma}}} \tag{19}$$

where $\boldsymbol{d}(x,y)_2$ is the L_2 norm of a displacement vector without being normalised to unit vector. ρ is the estimated radius of the iris. σ and ω correspond to the order and the cutoff frequency of the filter. The curves corresponding to varying σ and ω following Eq. (19) are shown in Fig. 13. It should be noted that in each subfigure only one parameter is variable while the other remains constant.

The radius weight function is maximally flat around the estimated centre ρ and drops rapidly when the radius is out of the flatness band whose range is controlled by σ. The roll-off rate is controlled by ω, indicating the decreasing rate in weight. Increasing ω while decreasing σ will enhance the rigidity of the constraint which could be assumed for circumstances where strong shadows are present. The value for ρ can be set according to the size of the face in an image. Consequently, this constraint can effectively alleviate problematic issues posed by dark and circular pixel clusters.

4.4. Selective oriented gradient filter

With the inspiration drawn from the histogram of oriented gradients (HOG) feature descriptor (Dalal and Triggs, 2005), a SOG filter is introduced that discriminates gradients of rapid change in orientation from those of less change. We specifically design this novel SOG filter and introduce it into the modular eye centre localisation scheme so that it is perfectly tailored to reinforce the two main modalities despite its versatile applicability. The basic idea takes the form of a statistical analysis of the gradient orientations within a window centred at a pixel position. For each $K_x \times K_y$ window (12×8 pixels in the experiment) centred at pixel i, the gradients in x and y directions are calculated, whose orientations follow:

$$\theta = \tan^{-1}\left(\frac{I_y}{I_x}\right) \cdot \frac{180^\circ}{\pi} \tag{20}$$

The gradient orientations are then accumulated into $k (k < 360)$ orientation bins, where each bin contains the count of orientations from $b \cdot \frac{360^\circ}{k}$ to $(b+1) \cdot \frac{360^\circ}{k}$ $(0 \le b \le k-1)$ within the window. If the recorded count in a bin exceeds a threshold, the corresponding pixels that accumulate the bin will have their gradient vectors halved, i.e. their weights reduced. As a result, the objective function becomes:

$$\boldsymbol{c}^* = \underset{\boldsymbol{c}}{\arg max} \left\{ \frac{1}{N}\sum_{x=1}^{m}\sum_{y=1}^{n} w_g(x,y)\cdot w_r(x,\ y)\cdot[\alpha - I(x,\ y)] \cdot(\boldsymbol{d}^{\boldsymbol{t}}(x,\ y)\cdot\nabla I_{\boldsymbol{c}}(x,\ y))^2 \right\} \tag{21}$$

where $w_g(x,\ y)$ *is the weight of a gradient adjusted by the SOG filter (i.e.* $w_g(x,y) = 0.5$ *to halve the gradient vectors and* $w_g(x,y) = 1$ *otherwise).* The threshold for the counts is determined by an absolute value (6 in the experiment) as well as a value (40% in the experiment) relative to the number of pixels with non-zero gradients within the window. As a result, pixels that maintain similar gradient orientations to their neighbours will have their weights reduced and they are referred to as 'impaired pixels' in the rest of the paper. When a curve has low curvature, it comprises more 'impaired pixels'. Therefore the SOG filter can be used for general curvature discrimination tasks. It has the advantage that it does not require an explicit function for the curve and that it is effective in dealing with curves forming irregular shapes. Fig. 14 demonstrates the effectiveness of a SOG filter applied to an image containing irregular curves and an image of an eye region.

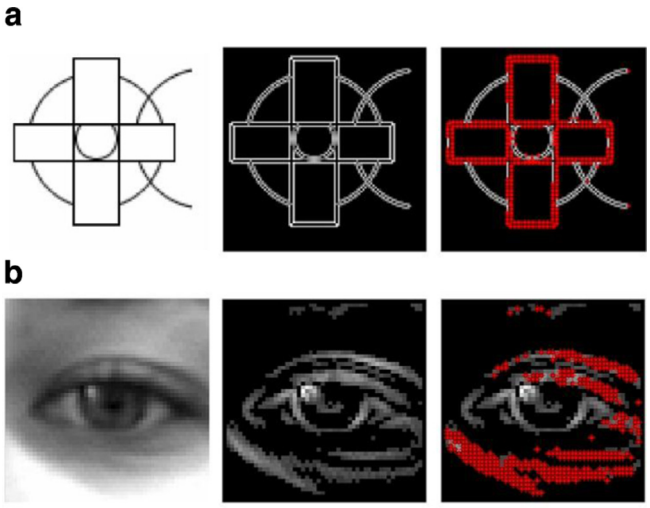


Fig. 14. Gradient filtering using a SOG filter. (a) An example of curved shape detection using a SOG filter where the 'impaired pixels' are detected. (b) An example where a SOG filter is applied to an eye image. The edges on the eye pouch and eyelids are successfully detected.

It is shown in Fig. 14 that the SOG filter has successfully distinguished curves with low and high curvatures and that it is effective in dealing with intersected and occluded curves or curve segments.

This approach allows both magnitude and orientation of gradients to serve for gradient filtering. It resolves the challenges brought by shadows, facial makeup and edges on the eyelids, eyebrows and other facial parts outside the iris that are most interfering in geometrical feature based eye centre localisation approaches.

4.5. Energy map integration

In the final stage, two energy maps E_a and E_b are integrated into $E_f(x,y)$ so that they both contribute to election of an eye centre. It is critical, prior to the integration, to determine the confidence of each modality, to estimate the complexity of the eye image, and thus to determine their weights in the fusion mechanism.

The left eye region is taken as an example to illustrate the fusion mechanism. If the equivalent centroid C_{aul} calculated by Eq. (16) is close to the pixel position C_{maxl} that has the maximum value in the first energy map $E_{aul}(x,y)$. , C_{aul} is considered confident since the isophote centre and the equivalent centroid coincide. In this case, more 'effective gradients' are present, allowing the second modality to be more robust and precise. The two modalities are then utilised and fused following Eq. (22). When C_{aul} and C_{maxl} disagree and have a large Euclidean distance, the first energy map will have high energy clusters sparsely distributed, potentially caused by severe shadows and specularities. The second modality will be fluenced by 'impaired pixels' and produce erroneous centre estimates. Therefore only the equivalent centroids C_{aul} and C_{aur} are selected as final eye centres.

$$E_f(x,\ y) = \frac{1}{\|C_{aul} - C_{Eamax}\|_2} \cdot E_a(x,\ y) + E_b(x,\ y) \tag{22}$$

where ϵ takes a value relative to the width of the eye region ϵ_f and $0 < \| C_{aul} - C_{Eamax} \|_2 \leq \epsilon$. In our experiments, $\epsilon = 0.3\epsilon_f$ pixels. The maximum response in the final energy map will represent the estimated eye centre. The estimate for the final right eye centre follows the same approach.

4.6. Eye centre localisation experiments and results

The public dataset used is the BioID database (BioID., 2014), consisting of 1520 images. It has been popular in the literature for evaluations of other eye centre localisation algorithms. The variations in the database include illumination, face scale, head pose and presence of glasses. Applied to this database, the proposed algorithm was tested against the others in the literature in resolving challenges introduced by the wide range of variations. The relative error measure proposed by (Jesorsky et al., 2001) was used to evaluate the accuracy of the algorithm. This method firstly calculates the absolute error, i.e. the Euclidian distance between the centre estimates and the ground truth provided by the database; it then normalises the Euclidian distance relative to the pupillary distance. This is formulated by Eq. (23).

$$e = \frac{\max\left(d_{left},\ d_{right}\right)}{\varepsilon} \tag{23}$$

where d_{left} and d_{right} are the absolute errors for the eye pair, and ε is distance between the left and the right eye, i.e. the pupillary distance, measured in pixels. The maximum of d_{left} and d_{right} after normalisation is defined as 'max normalised error' e_{max}. Similarly, the accuracy curve for the minimum normalised error e_{min} and the average normalised error e_{avg} are calculated. A relative distance of $e = 0.25$ corresponds to half the width of an eye. The accuracy of the proposed algorithm is evaluated by this error measure technique. Its accuracy curves are shown in Fig. 15. The proposed algorithm is further compared with 10 state-of-the-art methods in the literature, summarised in Table 2.

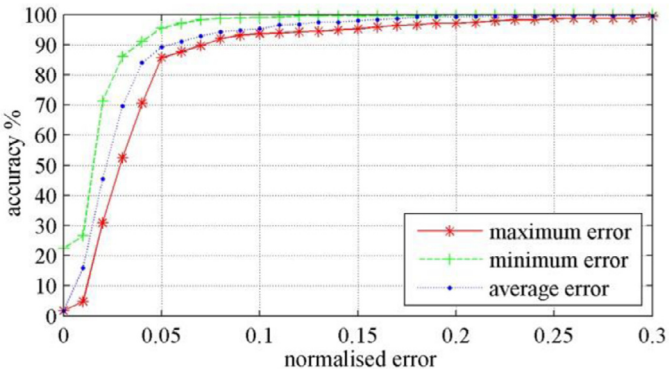


Fig. 15. Accuracy curves of the proposed eye centre localisation method.

The proposed method gains the best results for the accuracy measure $e_{min} \leq 0.10$ as well as $e_{max} \leq 0.25$, and the second best for $e_{max} \leq 0.05$ and $e_{max} \leq 0.10$. Except for the accuracy measure for $e_{min} \leq 0.25$ where very similar results are achieved, a score of 2 is assigned to every first rank and a score of 1 is assigned to every second rank. The proposed method gains a total score of 6, outperforming all the other methods when comparing the classification accuracy.

The simplicity and efficiency of the proposed method in computation is demonstrated by comparing it to (Timm and Barth, 2011) which claims to have achieved excellent real-time performance as one of its key features. Take an image containing a 41×47 eye region, i.e. 1927 pixels, as an example (Fig. 11), the method in comparison performs per-pixel estimation of the eye centre, assuming that every pixel is an eye centre candidate. Therefore 1927 iterations are needed before the optimal candidate is selected. The proposed method, on the other hand, resolves the problem by utilising prior knowledge drawn from the first modality which, through an initial estimation, avoids the per-pixel candidate assumption. The removal of the low-energy pixels in the first modality largely reduced the number of candidates, i.e. number of iterations in the second modality. In the same sample image of the 41×47 eye region, the iterations are decreased to only 67 from 1927, making our algorithm 29 times faster.

In summary, our eye centre localisation algorithm follows a coarse-to-fine, global-to-regional approach, making use of both isophote and gradient features. The high accuracy and efficiency of the proposed algorithm is achieved by two complementary modalities with a SOG filter designed to deal with the strong edges and shadows in the eye regions. The proposed method does not need any training or learning stages and is easy to implement.

5. Gaze gesture control for assistive HCI

5.1. Gaze gesture recognition

Gaze gestures are predefined sequences of eye movements which hold great potential in HCI. It has been proposed in the literature for disability assistance and other HCI purposes. Gaze gestures are derived from consecutive image frames where eye centre coordinates are collected. Therefore, an efficient and robust eye centre localisation algorithm becomes the key to successful gaze gesture recognition and real-time assistive technologies. Unfortunately, as reviewed previously, most eye centre localisation algorithms require dedicated devices (e.g. NIR illuminator for active methods or head-mounted devices) which render the resulting systems unrealistic for extensive use due to the high cost and inconvenience. Others, often being short-ranged, lack accuracy and robustness in real-world scenes. In comparison, our eye centre

Table 2
Comparison of the accuracy for eye centre localisation on the BioID database. Those with a '∗' notation are not explicitly provided by the author(s) but are measured from the accuracy curves available. Those with a '\' notation are neither explicitly nor implicitly provided by the authors. The numbers in bold are the highest accuracy in their corresponding ranges and those underlined the second highest.

Method	Accuracy under minimum and maximum normalised error						Score
	$e_{max} \leq 0.05$	$e_{min} \leq 0.05$	$e_{max} \leq 0.10$	$e_{min} \leq 0.10$	$e_{max} \leq 0.25$	$e_{min} \leq 0.25$	
The proposed method	85.66%	95.46%	93.68%	**99.06%**	**99.21%**	99.93%	**6**
(Timm and Barth, 2011)	82.50%	93.50%∗	93.40%	98.50%∗	98.00%	100%∗	**1**
(Leo et al., 2014)	80.67%	\	87.31%	\	93.86%∗	\	**0**
(Zhu and Ramanan, 2012)	65.00%	\	87.00%	\	98.80%	\	**1**
(Duda et al., 2012)	75.10%∗	\	93.00%	\	96.30%∗	\	**0**
(Asadifard and Shanbezadeh, 2010)	47.00%	\	86.00%	\	96.00%	\	**0**
(Valenti and Gevers, 2008)	84.10%	**96.28%**	90.85%	97.94%	98.49%	100%∗	**2**
(Valenti and Gevers, 2012)	**86.09%**	96.07%	91.67%	97.87%	97.87%	100%∗	**3**
(Campadelli et al., 2006)	62.00%	\	85.20%	\	96.10%	\	**0**
(Hamouz et al., 2005)	58.00%∗	\	76.00%∗	\	90.80%∗	\	**0**
(Cristinacce et al., 2004)	57.00%∗	\	**96.00%**∗	\	97.10%∗	\	**2**

localisation method provides an efficient and accurate means for collecting eye centre coordinates whose patterns, namely gaze gestures, can be further analysed.

The realisation of remote control of a HCI system via gaze gestures is non-invasive, low-cost and efficient. We introduce our method as a four-stage process: 1) accurate eye centre localisation in spatial-temporal domain, 2) eye movement encoding (Hyrskykari et al., 2012), 3) gaze gesture recognition and 4) HCI event activation.

We first employ the algorithm introduced in the preceding subsection for accurate and robust eye centre localisation. Two eye centre positions $\boldsymbol{E_l}(f) = \{e_{lx},\ e_{ly}\}$ and $\boldsymbol{E_r}(f) = \{e_{rx},\ e_{ry}\}$ are estimated and recorded for each frame. The mean value of them is calculated as $\bar{\boldsymbol{E}}(f) = \{\bar{e}_x, \bar{e}_y\} = \{(e_{lx} + e_{rx})/2, (e_{ly} + e_{ry})/2\}$, where f is the frame number and the x and y notations stand for the horizontal and vertical components.

In the second stage, the first order derivatives of the vectors $\boldsymbol{E_l}$, $\boldsymbol{E_r}$ and $\bar{\boldsymbol{E}}$ are calculated as $\boldsymbol{G_l}(f) = \{e_{lx}',\ e_{ly}'\}$, $\boldsymbol{G_r}(f) = \{e_{rx}',\ e_{ry}'\}$ and $\bar{\boldsymbol{G}}(f) = \{\bar{e}_x',\ \bar{e}_y'\}$, respectively. A threshold (4.5 % in the experiments) is then set to remove any small values in the two vectors, which might be caused by unintentional saccadic movements. It should be noted that the threshold is normalised by the pupillary distance ε so that it is independent of user-to-camera distance. Additionally, the movements of the two eyes are compared to each other with regard to their magnitudes according to Eq. (24) and (25). This logarithm function accounts for the fact that the two eyes in natural behaviours always move together, i.e. the left and the right eye move toward the same direction and shift by similar amount.

$$R_g = \begin{cases} \log_{10}\left(\dfrac{\sqrt{e_{lx}'^2 + e_{ly}'^2}}{\sqrt{e_{rx}'^2 + e_{ry}'^2}}\right), & if \sqrt{e_{lx}'^2 + e_{ly}'^2} \geq 1 \,\&\, \sqrt{e_{lx}'^2 + e_{ly}'^2} \geq 1 \\ 0, & otherwise \end{cases} \tag{24}$$

$$\begin{aligned} \mathcal{D}_{gx} &= e_{lx} \cdot e_{rx} \\ \mathcal{D}_{gy} &= e_{ly} \cdot e_{ry} \end{aligned} \tag{25}$$

When both $\mathcal{D}_{gx}$ and $\mathcal{D}_{gy}$ take positive values, we consider the directions of eye movements consistent. Ideally, when the two eyes shift by the same amount, R_g should have value 0. To allow for a margin of error, we set a threshold of 0.6 and consider the eye movement consistent when the absolute value of R_g falls below the threshold. Only when these two conditions are satisfied are the eye centre positions updated by those from the subsequent frame.

To encode voluntary eye saccades, we decompose $\bar{\boldsymbol{G}}(f)$ into $\overline{\boldsymbol{G_X}}(f) = \bar{e}_x'$ and $\overline{\boldsymbol{G_Y}}(f) = \bar{e}_y'$. Then a positive value in $\overline{\boldsymbol{G_X}}(f)$ is denoted by '1', a negative value by '2'; a positive value in $\overline{\boldsymbol{G_Y}}(f)$ is denoted by '4' and a negative value by '7'. Therefore '1', '2', '4', '7' are encoded gaze shifts representing saccadic strokes 'left', 'right', 'up' and 'down'. We further summate the two gaze shift vectors $\overline{\boldsymbol{G_X}}(f)$ and $\overline{\boldsymbol{G_Y}}(f)$ and produce $\overline{\boldsymbol{G_s}}(f)$, the integrated gaze shift vector. As a result, voluntary gaze shifts are recorded as a combination and repetition of the four digits, while a '0' represents an unchanged eye position or an involuntary saccade. We define in the experiments for HCI seven types of gaze gestures, shown in Table 3. Other saccadic codes ('5'='1'+'4', '6'='2'+'4', '8'='1'+'7', '9'='2'+'7') denote diagonal saccadic strokes that are reserved for our future work.

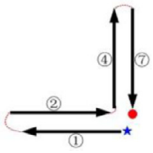
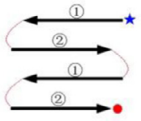
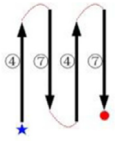
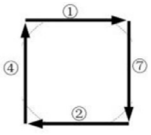
In the third stage, the gaze gesture patterns are recognised by searching for specific gesture sequences in a segment of $\overline{\boldsymbol{G_s}}(f)$ (every segment being a two-second time slot in our experiments), which will trigger pre-defined HCI events in the last stage.

Fig. 16. The structure of the experimental directed advertising system.

5.2. Development of an assistive HCI system

To validate our method from an applied perspective, we designed an experimental directed advertising system (Fig. 16) that consisted of a high-definition (HD) 47-inch display, a webcam operating at 640 × 480 resolution and a PC in the cabinet for camera control and data storage/processing. Digital signage systems have been prevalent for years in this digital era and the information age.

Table 3
Definition of seven types of gaze gestures for HCI. A star notation denotes the starting position of a gaze gesture and a dot denotes the end of a gaze gesture. The circled numbers represent the encoded gaze shifts. The arrows denote saccadic strokes. Type 7 gaze gesture can start from any position. Type 6 gaze gesture will only be recognised as a subsequent gesture of type 1, 2, 3 or 4.

Gesture No.	Gesture Sequence	Gesture Pattern	Gesture Name	HCI Event	Dependency
1	1 → 2 → 4 → 7		Top-left gaze	Bring the *top-left* thumbnail advertisement to the screen centre	N/A
2/3/4	2 → 1 → 4 → 7 1 → 2 → 7 → 4 2 → 1 → 7 → 4	Similar to Gesture No. 1	Top-right gaze/Bottom-left gaze/Bottom right gaze	Bring the *top-right/bottom-left/bottom-right* thumbnail advertisement to the screen centre	N/A
5	1 → 2 → 1 → 2		Reset gaze	*Reset* to default display	N/A
6	7 → 4 → 7 → 4		Zoom-in gaze	*Show details* of the selected advertisement (at the screen centre)	Gesture No. 1 or 2 or 3 or 4
7	2 → 4 → 1 → 7 4 → 1 → 7 → 2 1 → 7 → 2 → 4 7 → 2 → 4 → 1		Change-content gaze	Change all the advertisement thumbnails	N/A

Their use for advertising has become ubiquitous and can be found at venues such as restaurants, shopping malls, airports and other public spaces. Often referred to as digital out-of-home (DOOH) advertising (Lasinger and Bauer, 2013), this advertising format aims to extend the exposure and the effectiveness of marketing messages by engaging consumers to an increased extent, compared to conventional print based billboards. Following the idea of switching from print media to digital media, it is only intuitive to reform a conventional DOOH advertising system toward a HCI system for enhanced interactivity and adaptability. This system further plays assistive roles by allowing the elderly and the disabled to browse information remotely and creating a user-friendly atmosphere according to the user characteristics it gathers and predicts.

When a user approaches the system, his/her face images are captured by the camera. The FV encoding method then output the predicted gender label. As a result, advertisement thumbnails that are highly relevant to the predicted gender group will be displayed in replacement to previous advertising messages. The eye centre localisation algorithm then detects eye centres in every image frame and supplies the information to the gaze gesture recognition algorithm. When a gaze sequence matches one of the seven predefined gaze gesture patterns (see Table 3), the advertising messages displayed on the screen can be manipulated correspondingly. For example, when gaze gesture No. 2 is detected, the top-right advertisement thumbnail will be displayed at the screen centre. If gaze gesture No. 6 is detected as a subsequent gesture, an enlarged view of the centred thumbnail will become available. This puts the user in an active role for being able to receive the recommended advertisements from the system, as well as being able to browse or switch the advertisements oneself. In the tests, the system could robustly and accurately perform advertisement selection (gaze gesture type 1, 2, 3, 4, 7), reset (gaze gesture type 5) and zoom-in (gaze gesture type 6) operations. Any gaze gesture could be recognised as soon as the last eye saccade in a gaze gesture sequence was issued by a user. When no face appears in a frame, randomly selected advertisement thumbnails will be brought to circulation on the screen. Apart from enabling active advertisement browsing, the system can also passively construct an attentive energy map to reflect relative user attention by accumulating eye centre positions over time. Higher energy points correspond to directions where a user gazes at for longer time.

A video demonstration (see Supplementary Material directed_advertising.mp4) has been made to illustrate this HCI process where the display of advertisements responds to the gender and gaze of a user. A representative frame from the demonstration is displayed in Fig. 17. This demonstration shows that the proposed method can accurately localise eye centres on well-illuminated faces (in a normal indoor environment) and on poorly-illuminated

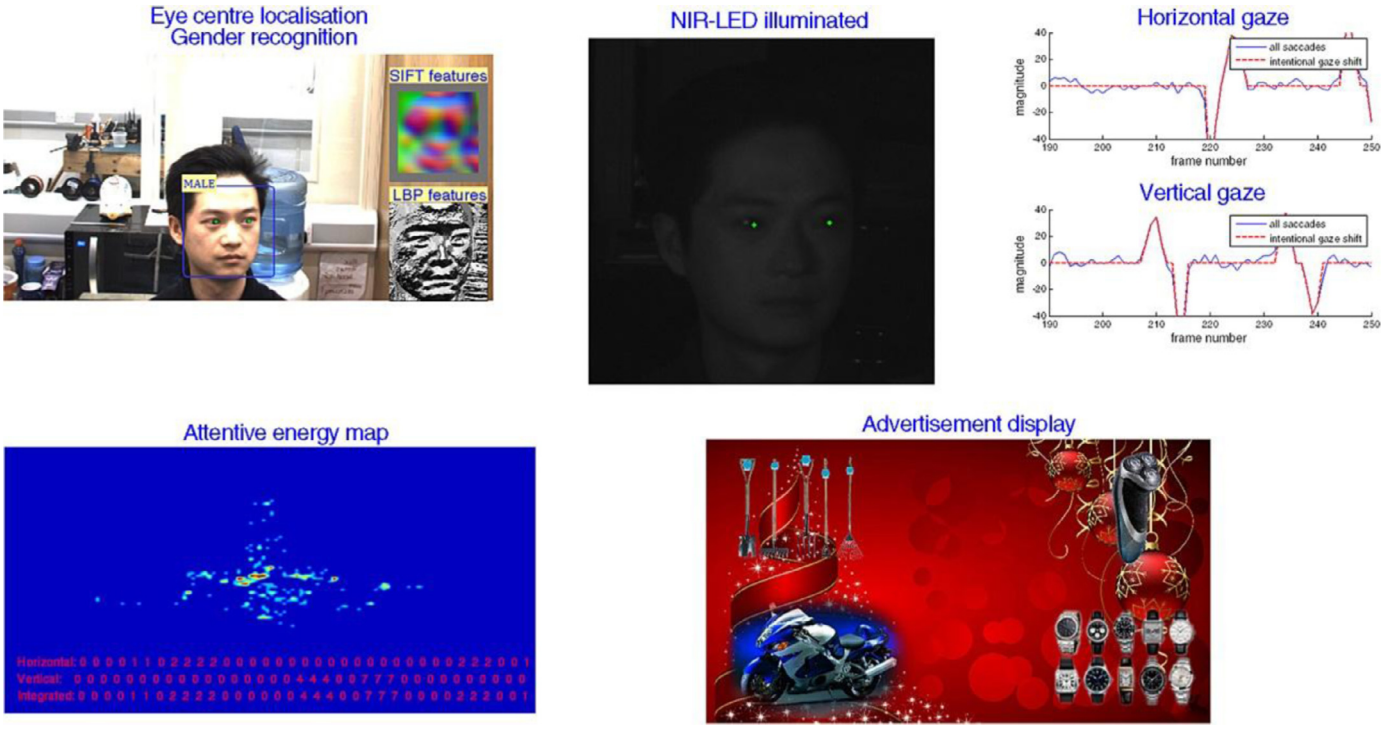


Fig. 17. A representative frame from the directed advertising demonstration. The top-left image shows the detected face, the predicted gender label, the localised eye centres and the extracted features on a well-illuminated face. The top-middle image shows localised eye centres on a poorly-illuminated face. The top-right image shows horizontal and vertical eye saccades (see Section 5.1). The bottom-left image displays the attentive energy map that accumulates eye centre coordinates over time. The bottom-right image shows the advertisements being delivered.

faces (only illuminated by two near-infrared (NIR) LEDs). The predicted gender label and the gaze gestures can be utilised collectively to change the advertisement display. Another video demonstration (see Supplementary Material eye_centre_localisation.mp4) has been made to show that eye centres can be accurately localised on over-exposed images of a face with different head poses and eye occlusions (by a glass frame and a finger). A representative frame of this demonstration is shown in Fig. 18.

It should be noted that although the proposed HCI strategy combines gender and gaze analysis, the two respects have been implemented as individual algorithms such that the results from automatic gender recognition and gaze gesture recognition are not correlated in general. The highly accurate and efficient gaze gesture recognition benefits from the two complementary modules in the eye centre localisation algorithm; the gender recognition method benefits from the encoded discriminant FVs and a simple linear classifier. However, the localised eye centres on a face image can be utilised for face alignment which increases gender recognition accuracy. The fused knowledge drawn from demographic data and behavioural data can enrich functionality of HCI systems while maintaining the merits of individual recognition algorithms.

Our directed advertising system has manifested excellent accuracy and robustness in the real-world tests. Tested with Microsoft Visual Studio 2012 on a computer with an Inter(R) Core(TM) i5-4570 CPU and 12GB RAM, the proposed gender and gaze gesture recognition method has exhibited excellent real-time performance with an average frame rate of 32 frames per second. It should be noted that during the face detection stage, we simply employed the Viola-Jones face detector which consumed nearly 65% of the computational time in our tests. This indicates that much higher frame rates can be achieved by employing a more efficient face detector.

User-camera distance was set at 0.5 metres, 1 metre and 1.5 metres, respectively. The alteration of user-camera distance saw negligible impact on the system performance. The reason behind this is that each face image is resized to a standard size before eye centre localisation, and that eye saccadic signals are relative values, normalised by the pupillary distance. The illumination in the tests was provided by overhead lamps (at around seven metres high), which could represent common indoor settings. NIR illumination was also employed to create a controllable under-illuminated environment. Outdoor environments with strong sunshine are simulated by over-exposed images.

The combined demographic recognition and behavioural recognition have brought higher functionality and usability to this case study HCI system. Without the gender recognition module, advertising messages cannot be tailored to user characteristics. On the other hand, without the gaze gesture recognition module, although personalised advertisements can be delivered to suit different gender groups, they cannot respond to the level of user satisfaction and attentiveness, but can only present digital content in a passive manner. More importantly, the functionality of the system is not restricted to a particular context but is applicable to the broader theme of assisting the elderly and the disabled by creating a user-centred HCI environment. Assistive systems can be designed to provide service for patients in hospitals, personal care for the elderly who live alone and the disabled in public venues, knowing their gender and knowing more about their attentions/intentions. Therefore the care and service delivered will be less general but more personalised.

6. Discussion and conclusions

This paper explores the two popular visual modalities in HCI – gender and gaze. Three novel algorithms have been proposed which enable HCI system to fulfil assistive roles in a wide range of scenarios.

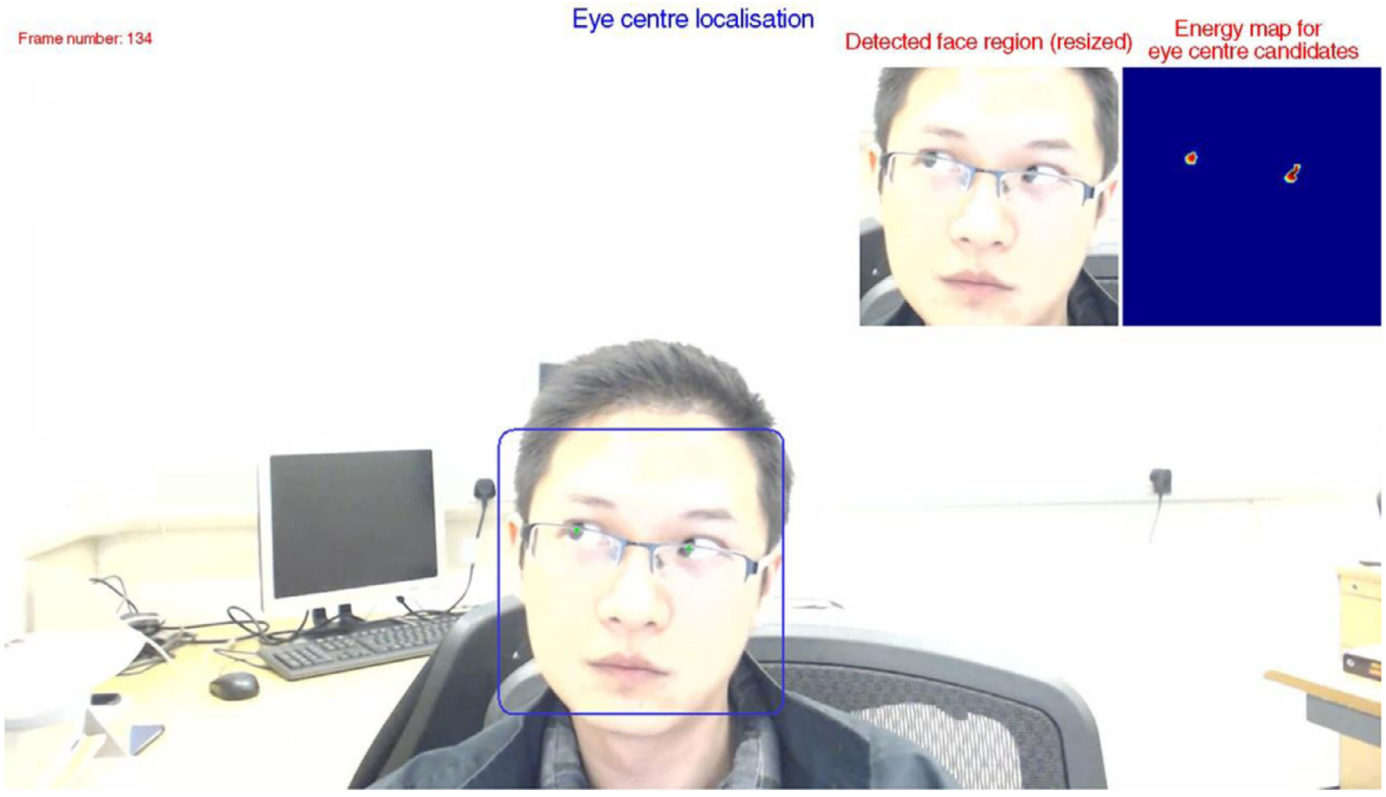


Fig. 18. A representative frame demonstrating eye centres localised on a face with head pose, eye movement, and occlusion caused by facial accessory. The energy map for eye centre is also displayed (low energy points removed, see Section 4.1).

We introduce a gender recognition algorithm that adopts a type of discriminative encoded feature, namely the Fisher Vectors, to reliably predict gender from facial images. Comprehensive tests have been carried out that evaluate: 1) the FVs encoded from four different types of low level features – greyscale, LBP, LBP histogram and SIFT, 2) the impacts of algorithm parameters – image size, sampling window size, sampling stride, GMM component number and principal component number 3) pre-processing techniques – histogram equalisation and face alignment and 4) algorithm performance on controlled image data and uncontrolled image data. As a result, we conclude that the SIFT feature yields the highest gender recognition rate, 97.7% on the FERET dataset, 92.5% on the LFW database and 96.7% on the FRGCv2 database. In addition, we compare our method with eight other state-of-the-art approaches and prove the superiority of our method. We further prove the robustness of our algorithm against head pose by showing that misaligned facial images have an insignificant negative impact on the recognition accuracy (less than 1% decrease). Another merit of our gender recognition method is its ability to identify the most discriminative facial region, i.e. the regions on a face that characterise male and female groups. As found by our algorithm, the mouth, the nasolabial furrows and the forehead regions are the most discriminative. The only disadvantage of our algorithm is high memory consumption in the phase of GMM training. This is incurred by low level features that are densely sampled such that the number of descriptors is relatively large. However the prolonged training phase will not have an impact on the classification phase since we only employ a linear SVM. Although the proposed FV encoding method has yielded promising results for gender recognition, further accuracy boost can be expected by incorporating more robust and intrinsically discriminative features. To this end, in our future works, novel 3D imaging systems, 3D reconstruction methods and 3D gender recognition strategies will be explored.

As for gaze analysis, we first propose an unsupervised modular approach for eye centre localisation as the preliminary stage. This approach utilises gradient and isophote features and follows a coarse-to-fine and global-to-regional scheme. We further design a SOG filter that specifically deals with the prominent gradients from the eyelids, eyebrows and shadows which sabotage most geometrical feature based methods. Our eye centre localisation method is free from classifier training and absolute facial anthropometric relations so that it is efficient and robust. This approach has been tested on the BioID dataset and compared to 10 other state-of-the-art methods in six accuracy measures. It outperforms all the other methods in comparison by gaining the highest accuracy measure score. Apart from its high accuracy, the algorithm exhibits superior real-time performance as the two modules interact with each other and largely reduce eye centre candidates.

Building on this algorithm, we design seven gaze gestures that can be used to control a HCI system in a remote and contactless manner. We tested the gaze gesture recognition algorithm by designing a directed advertising system that, upon receiving gaze gestures issued by a user, displays advertisements in different manners. This type of system combines demographic recognition (gender) and behaviour analysis (gaze) and therefore is able to create a user-centred HCI environment by better understanding the needs and intentions of its users. We consider that these capabilities can enable advanced functionality that would offer potential to commercially implement many new advertising applications. Regarding assistive roles, the system offers huge potential in various situations and is especially valuable for enabling assistance and communication for the elderly and people with motor disabilities.

In our future works, the Fisher Vector encoding method will be extended such that reconstructed 3D faces can be explored as the source of more robust features. Novel 3D reconstruction algorithms

will also be explored, accompanied by the development of other types of 2D and 3D based HCI systems suitable for use in real-world environments.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cviu.2016.03.014.

References

Grudin, J., 2011. Human-computer interaction. Annual Review of Information Science and Technology 45 (1), 367–430.

Karray, F., Alemzadeh, M., Saleh, J.A., Arab, M.N., 2008. Human-computer interaction: Overview on state of the art. Int. J. Smart Sens. Intell. Syst. 1 (1), 137–159.

Jaimes, A., Sebe, N., 2007. Multimodal human–computer interaction: A survey. Comput. Vis. Image Und. 108 (1), 116–134.

Tawari, A., Chen, K.H., Trivedi, M.M., 2014. Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation. In: 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 988–994.

Wachs, J.P., Stern, H.I., Edan, Y., Gillam, M., Handler, J., Feied, C., Smith, M., 2008. A gesture-based tool for sterile browsing of radiology images. J. Am. Med. Inform. Assoc. 15 (9), 321–323.

Kapoor, A., Burleson, W., Picard, R.W., 2007. Automatic prediction of frustration. Int. J. Hum. Comput. Stud. 65 (8), 724–736.

Ng, C.B., Tay, Y.H., Goi, B.M., 2012. Vision-based human gender recognition: A survey. Comput. Vis. Pattern Recognit. 1611.

Viola, P., Jones, M.J., 2004. Robust real-time face detection. Int. J. Comput. Vis. 57 (2), 137–154.

Viola, P., Jones, M.J., 2001. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition I-511–I-518.

Mäkinen, E., R., R., 2008. An experimental comparison of gender classification methods. Pattern Recognit. Lett. 29 (10), 1544–1556.

Mäkinen, E., Raisamo, R., 2008. Evaluation of gender classification methods with automatically detected and aligned faces. IEEE Trans. Pattern Anal. Mach. Intell. 30 (3), 541–547.

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

Shan, C., 2012. Learning local binary patterns for gender classification on real-world face images. Pattern Recognit. Lett. 33 (4), 431–437.

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

Lee, P.H., Huang, J.Y., Huang, Y.P., 2010. Automatic gender recognition using fusion of facial strips. In: in: 20th International Conference on Pattern Recognition, pp. 1140–1143.

Wang, J.G., Li, J., Yau, W.Y., Sung, E., 2010. Boosting dense SIFT descriptors and shape contexts of face images for gender recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 96–102.

Tivive, F.H.C., Bouzerdoum, A., 2006. A gender recognition system using shunting inhibitory convolutional neural networks. In: International Joint Conference on Neural Networks, pp. 5336–5341.

Phung, S.L., Bouzerdoum, A., 2007. A pyramidal neural network for visual pattern recognition. IEEE Trans. Neural Netw. 18 (2), 329–343.

Alexandre, L.A., 2010. Gender recognition: A multiscale decision fusion approach. Pattern Recognit. Lett. 31 (11), 1422–1427.

Hu, Y., Yan, J., Shi, P., 2010. A fusion-based method for 3D facial gender classification. In: The 2nd International Conference on Computer and Automation Engineering, pp. 369–372.

Wang, X., Kambhamettu, C., 2013. Gender classification of depth images based on shape and texture analysis. In: Proceedings of Global Conference on Signal and Information Processing, pp. 1077–1080.

Fagertun, J., Andersen, T., Paulsen, R.R., 2012. Gender recognition using cognitive modelling. In: Proceedings of Computer Vision–ECCV, pp. 300–308.

Huynh, T., Min, R., Dugelay, J.L., 2012. An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data. In: Proceedings of Computer Vision–ECCV, pp. 133–145.

Zhu, Z., Ji, Q., 2005. Robust real-time eye detection and tracking under variable lighting conditions and various face orientations. Comput. Vis. Image Und. 98 (1), 124–154.

Timm, F., Barth, E., 2011. Accurate Eye Centre Localisation by Means of Gradients. In: International Conference on Computer Vision Theory and Applications, pp. 125–130.

Leo, M., Cazzato, D., De Marco, T., Distante, C., 2014. Unsupervised Eye Pupil Localization through Differential Geometry and Local Self-Similarity Matching. PLoS One 9 (8).

Zhu, X., Ramanan, D., 2012. Face detection, pose estimation, and landmark localization in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2879–2886.

Duda, R.O., Hart, P.E., Stork, D.G., 2012. Pattern classification, second ed. John Wiley & Sons, New York.

Kroon, B., Hanjalic, A., Maas, S.M., 2008. Eye localization for face matching: is it always useful and under what conditions? In: Proceedings of the 2008 international conference on Content-based image and video retrieval, pp. 379–388.

Niu, Z., Shan, S., Yan, S., Chen, X., Gao, W., 2006. 2d cascaded adaboost for eye localization. In: 18th International Conference on Pattern Recognition, pp. 1216–1219.

Asadifard, M., Shanbezadeh, J., 2010. Automatic adaptive center of pupil detection using face detection and cdf analysis. In: Proceedings of the International MultiConference of Engineers and Computer Scientists, p. 3.

Türkan, M., Pardas, M., Enis, C.A., 2007. Human eye localization using edge projections. In: International Conference on Computer Vision Theory and Applications, pp. 410–415.

Chang, C.C., Lin, C.J., 2011. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 (3) 27.

Drewes, H., Luca, A.D., Schmidt, A., 2007. Eye-gaze interaction for mobile phones. In: Proceeding of the 4th international conference on mobile technology, applications, and systems and the 1st international symposium on Computer human interaction in mobile technology, pp. 364–371.

Hyrskykari, A., Istance, H., Vickers, S., 2012. Gaze gestures or dwell-based interaction? In: Proceedings of the Symposium on Eye Tracking Research and Applications, pp. 229–232.

Rozado, D., Rodriguez, F.B., Varona, P., 2012. Low cost remote gaze gesture recognition in real time. Appl. Soft Comput. 12 (8), 2072–2084.

Reynolds, D., 2009. Gaussian Mixture Models. Encyclopedia of Biometrics. Springer, US, pp. 659–663.

Simonyan, K., Parkhi, O., Vedaldi, A., Zisserman, A., 2013. Fisher Vector Faces in the Wild. In: Proceedings of British Machine Vision Conference, 8, pp. 1–8. 12.

Liu, C., Yuen, J., Torralba, A., 2011. SIFT flow: dense correspondence across scenes and its applications. IEEE Trans. Pattern Anal. Mach. Intell. 33 (5), 978–994.

Vedaldi, A., Fulkerson, B., 2010. VLFeat: An open and portable library of computer vision algorithms. In: Proceedings of the international conference on Multimedia, pp. 1469–1472.

Perronnin, F., Sánchez, J., Mensink, T., 2010. Improving the fisher kernel for large-scale image classification. Computer Vision–ECCV. Springer, Berlin Heidelberg, pp. 143–156.

Do, H., Kalousis, A., Wang, J., Woznica, A., 2012. A metric learning perspective of SVM: on the relation of LMNN and SVM. In: International Conference on Artificial Intelligence and Statistics, pp. 308–317.

Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.J., 1998. The FERET database and evaluation procedure for face-recognition algorithms. Image Vis. Comput. 16 (5), 295–306.

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

Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W., 2005. Overview of the face recognition grand challenge. In: Proceedings of IEEE Computer Society Conference on CVPR, pp. 947–954.

Ojala, T., Pietikäinen, M., Mäenpää, T., 2000. Gray scale and rotation invariant texture classification with local binary patterns. In: Proceedings of Computer Vision–ECCV, pp. 404–420.

Sánchez, J., Perronnin, F., Campos, T.D., 2012. Modeling the spatial layout of images beyond spatial pyramids. Pattern Recognit. Lett. 33 (16), 2216–2223.

Arandjelovic, R., Zisserman, A., 2012. Three things everyone should know to improve object retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2911–2918.

Lichtenauer, J., Hendriks, E., Reinders, M., 2005. Isophote properties as features for object detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 649–654.

Valenti, R., Gevers, T., 2008. Accurate eye center location and tracking using isophote curvature. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8.

Koenderink, J.J., Doorn, A.J.V., 1992. Surface shape and curvature scales. Image Vis. Comput. 10 (8), 557–564.

The BioID. Face database. https://www.bioid.com/About/BioID-Face-Database, 2014

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 886–893.

Jesorsky, O., Kirchberg, K.J., R.W., 2001. Robust face detection using the hausdorff distance. Audio-and Video-Based Biometric Person Authentication. Springer, Berlin Heidelberg, pp. 90–95.

Valenti, R., Gevers, T., 2012. Accurate eye center location through invariant isocentric patterns. IEEE Trans. Pattern Anal. Mach. Intell. 34 (9), 1785–1798.

Campadelli, P., Lanzarotti, R., Lipori, G., 2006. Precise eye localization through a general-to-specific model definition. In: Proceedings of British Machine Vision Conference, pp. 187–196.

Hamouz, M., Kittler, J., Kamarainen, J.K., Paalanen, P., Kalviainen, H., Matas, J., 2005. Feature-based affine-invariant localization of faces. IEEE Trans. Pattern Anal. Mach. Intell. 27, 1490–1495.

Cristinacce, D., Cootes, T.F., Scott, I.M., 2004. A multi-stage approach to facial feature detection. In: Proceedings of British Machine Vision Conference, pp. 1–10.

Lasinger, P., Bauer, C., 2013. Situationalization, The new road to adaptive digital-out-of-home advertising. In: Proceedings of IADIS International Conference e-Society, pp. 162–169.