# Multi-View Face Detection and Registration Requiring Minimal Manual Intervention

*Seyed Mohammad Hassan ANVAR, Student Member, IEEE, Wei-Yun YAU, Senior Member, IEEE, and Eam Khwang TEOH, Member, IEEE*

**Abstract**— Most face recognition systems require faces to be detected and localized a priori. In this paper, an approach to simultaneously detect and localize multiple faces having arbitrary views and different scales is proposed. The main contribution of this paper is the introduction of a face constellation, which enables multi-view face detection and localization. In contrast to other multi-view approaches that require many manually labeled images for training, the proposed face constellation requires only a single reference image of a face containing two manually indicated reference points for initialization. Subsequent training face images from arbitrary views are automatically added to the constellation (registered to the reference image) based on finding the correspondences between distinctive local features. Thus, the key advantage of the proposed scheme is the minimal manual intervention required to train the face constellation. We also propose an approach to identify distinctive correspondence points between pairs of face images in the presence of a large amount of false matches. To detect and localize multiple faces with arbitrary views, we then propose a probabilistic classifier based formulation to evaluate whether a local feature cluster corresponds to a face. Experimental results conducted on the FERET, CMU and FDDB datasets show that our proposed approach has better performance compared to the state-of-the-art approaches for detecting faces with arbitrary pose.

**Index Terms**— Multi-view, Image registration, Face constellation, Simultaneous face detection and localization.

## 1 Introduction

FACE detection and localization is a critical requirement for most face recognition systems. This is because it segments out the appropriate face region to the face recognition module for processing and as such, the performance of this module will significantly affect the overall result of the recognition task [1]. Almost all recognition algorithms [2-4] assume that the input face has been correctly detected and cropped during the pre-processing stage. However, such an assumption is often not valid for real world applications such as video surveillance as the faces are in various poses and some of the faces are occluded. Commonly used window-based face detection techniques such as Viola and Jones [5], Schneiderman and Kanade [6], Li and Zhenqiu [7], and Huang et al. [8] only yield coarse face detection results. Subsequently, fine tuning algorithms are required to crop and align the detected faces to meet the need of face recognition module [9]. Most fine tuning algorithms require locating the position of the eyes [10] or regularization with template matching [11-13]. These algorithms are sensitive to variations in face pose, facial expression, illumination, and background scene [14, 15]. To address the face alignment problems, patch based algorithms have been proposed [16, 17]. However, these algorithms require a face to be detected a priori. Furthermore, the existing algorithms do not perform well with non-frontal faces or in the presence of occlusion. Moreover, window-based face detection techniques [5, 7, 8] and probabilistic frameworks [6, 18] require a large amount of manual intervention to prepare a training set. Usually, these consist of thousands of manually modified images that must be cropped and labeled. Thus preparing the training images is a very time consuming and costly process.

To address these problems, in this paper we introduce an approach that combines face detection, localization, and feature selection tasks into a single framework. Our approach is based on establishing a face constellation, which is essentially a *reference map* indicating how all the other face images in the training set are registered with respect to a reference image. A key feature of this work is the formulation of links to establish a match of a given face with a reference image and to detect and localize faces from arbitrary views at different scales. The purpose of the training images is to provide enough diversity to the constellation such that it can provide sufficient links that connect various registered faces from different poses, illumination, background clutter, etc.

Another contribution of this work is the feature selection technique based on local descriptors and


- *S.M.H. ANVAR is with the school of Electrical and Electronic Engineering, Nanyang Technological University and the Institute for Infocomm Research, Singapore. E-mail: seye0005@e.ntu.edu.sg.*
- *W.Y. YAU is with the Institute for Infocomm Research, A*STAR, Singapore. E-mail: wyyau@i2r.a-star.edu.sg.*
- *E.K. TEOH is with the school of Electrical and Electronic Engineering http://coe.ntu.edu.sg/TheCollege/SchoolsunderCollege/Pages/SchoolofElectricalandElectronicEngineering.aspx, Nanyang Technological University, Singapore. E-mail: eekteoh@ntu.edu.sg.*

geometric constraints to select distinctive correspondence points between two faces from among many false or unreliable matches in the registration stage. A method to detect and correct erroneous matches is also included in order to avoid error propagation. Unlike [5-8, 18], our proposed approach requires only manual input of two *reference points* in just one reference image during the training stage. Subsequent processing is then fully automated. As such, the proposed approach is not purely a face detector (e.g. [5, 7]) or face alignment algorithm (e.g. [16, 17, 19]), but a learning technique for training multi-view face classifiers with minimum supervision or labeling. Experimental results show that the proposed method is comparable to state-of-the-art face detectors and can localize faces accurately.

The organization of this paper is as follows: section 2 presents a literature review of face detection techniques. Section 3 introduces the framework for constructing a multi-view face constellation by finding distinctive correspondence points among the images and describes a method to register arbitrary views of faces from different persons and at different scales. Section 4 then describes a probabilistic classification algorithm to locate the registered faces having multiple views and scales in highly cluttered background scenes. Section 5 provides the experimental results for each stage of the proposed method and finally section 6 concludes the paper.

## 2 Related Work

In this section, we review some of the popular feature extraction and matching techniques used in face detection and summarize the state-of-the-art multi-view face detection methods.

### 2.1 Invariant Feature Extraction

Objects are usually recognized by their unique features. Many feature extraction methods have been proposed in the literature. Some methods such as Haar-like features [20] and Local Binary Pattern features (LBP) [21] are computed for every pixel in an image. However, these features are not scale invariant. Alternatively, other methods compute the salient points in images to both decrease the processing time and increase the performance. Mikolajczyk and Schmid [22] introduced the Harris-Laplace detector to detect corners and salient points in images which is invariant to scale change. Lowe [23] used Difference of Gaussian operations at multiple scales to find the salient points and defined an appearance descriptor for these points based on the gradient histogram in the local neighborhood. These features are invariant to scale changes and robust to small rotations and are commonly referred to as Scale Invariant Feature Transform (SIFT). Some variants of SIFT have been proposed recently, such as GLOH [24], SURF [25] and SCARF [26].

Solutions for detecting objects or faces with arbitrary views require a feature extraction method that would be able to cope with illumination and pose variation, object deformation and partial occlusion. A comprehensive survey was carried out by Mikolajczyk and Schmid [24] to evaluate the performance of different feature extraction algorithms in images under various transformation such as rotation, zoom, viewpoint change, blurring, JPEG compression and illumination variation. They concluded that SIFT features and their improved version (GLOH) have better performance compared to the other features. Therefore, in this paper, we adopt the SIFT features to represent salient points in face images.

### 2.2 Matching Techniques

A matching approach was proposed by Mikolajczyk and Schmid [27] to find the correspondence points between two images based on the scale and affine invariant interest points identified using a Harris detector [22]. They computed Gaussian derivatives up to the $4^{th}$ order around the interest points to generate a 12-dimension descriptor. They then calculate the similarity between the descriptors in the first image and the second image using the Mahalanobis distance. A potential correct match is found if the distance falls below a preset threshold. Subsequently, cross correlation is performed to eliminate some false matches.

Lowe [23] proposed another matching method based on SIFT salient points. The similarities between salient points in the first and second images were found using minimum Euclidean distance. Post-processing was performed to remove any false correspondences. The final correspondence points are obtained by estimating the transformation between the two images using the random sample consensus (RANSAC) method [28]. RANSAC works well when the number of false correspondences is small compared to the overall matches found. However, when the number of real correspondence points is small compared to the number of outliers, RANSAC will fail [27].

### 2.3 Multi-View Face Detection Approaches

Several approaches have been proposed to detect objects and faces in multi-view [6, 8, 18, 29]. Some schemes [6, 8, 29] used an ensemble of similar detectors, with each detector trained to detect the face in a predefined pose. The training process requires manually labeled images of all poses. A sliding window at different scales exhaustively searches the given image to determine the existence of a face. This procedure is repeated for every pose of the face by running it through the ensemble of classifiers. Such approaches are slow and work well only on the finite poses being trained. Generalization is required for the other poses. To improve the accuracy, more poses have to be trained, requiring more detectors as well as increasing the processing time. Thus, a compromise has to be reached on the number of poses to be trained versus the desired computational speed. To improve the speed, Toews and Arbel [18] proposed a non-window approach using Bayesian classifier. To train their detector, some control points in all the training images are manually selected. The control points are

chosen such that they would be invariant to view changes. Although the performance of this method for frontal face is not as good as the ensemble of classifiers, the result for the non-trained views is very promising. In all the face detection methods above, the training images have to satisfy some constraints or assumptions such as face images are manually cropped and aligned, have a plain background, contain only frontal and near frontal views or have minimum rotation at the control points chosen in the training images.

In this paper, we propose a multi-view face detection approach that requires no prior assumption about the training images. Our approach starts with an image of a face with two known reference points and then enriches the classifier with some training data in various poses. With sufficient data, a constellation similar to the projection of a 3D model onto 2D space is formed for the face, comprising links to various faces in different poses. This model is then used to detect and localize faces in arbitrary poses and other challenging scenarios such as image with multiple faces in different views.

## 3 Multi-view Single Face Detection and Registration with Minimal Manual Intervention

In this section, we describe our proposed approach to detecting a single face in an arbitrary view. The next section will then extend it to multiple faces having arbitrary views. The basic concept of our proposed approach is to detect correspondence points between image pairs. Although pairs of face images with significantly different poses will not match, there would be sufficient matches if the pose variations are within a tolerable limit. Thus, sufficiently large number of face samples with arbitrary views must be included in the face constellation. By increasing the number of matched pairs among gradually varying poses and maintaining the linkages associated with such matched pairs, the constellation can cover the entire space of possible pose variations of the face. By virtue of the correspondences maintained, if two control points (i.e. the *reference points*) in a face image (the reference image) are known, then the other face images can be registered with respect to the reference image. Given an unknown image, the presence of a face can be determined by finding local patches in the given image that match with the local patches of the faces in the constellation. Thus the likelihood that the patch belongs to a valid face image is directly proportional to the number of matches found. The following subsections explain the proposed approach in more detail.

### 3.1 Finding Distinctive Correspondence Points and Reference Points

In order to find the distinctive correspondence points, we first consider a pair of images, say $I_i$ and $I_j$, and extract invariant feature points from both images using the scale invariant feature transform (SIFT) [23]. Let $F_i = \{f_{i,k}\}$, $k = 1,2,\dots,N_i$ and $F_j = \{f_{j,l}\}$, $l = 1,2,\dots,N_j$ be the collection of feature points obtained from images $I_i$ and $I_j$ respectively. $N_i$ and $N_j$ are the number of feature points extracted from images $I_i$ and $I_j$ respectively. Each feature point includes both appearance ($f^A$) and geometric ($f^G$) descriptors. For instance, the notation $f_{i,k}^A$ ($f_{i,k}^G$) represents the appearance (geometric) descriptor of the $k^{th}$ feature in image $I_i$. As indicated in section 2.1, the appearance descriptor is a gradient histogram in the local neighborhood of the feature point. The geometric descriptor can be represented as $f_{i,k}^G = (x_{i,k}, y_{i,k}, \sigma_{i,k}, \phi_{i,k})$, where $(x, y)$ is the position of the feature point $k$ in image $I_i$, $\sigma$ is the scale and $\phi$ is the gradient orientation.

When the background is highly cluttered or when the pose variation between the faces in the two images are significant, the extracted SIFT features would have a substantial number of common or non distinctive features together with the required distinctive features. Thus, it is necessary to select only the distinctive features, which lie in the most discriminative parts of a face [30]. We can consider two different hypotheses: either the feature $f_{i,k}$ is a distinctive part of a face ($H_{i,k} = 1$) or it is a non-distinctive feature ($H_{i,k} = 0$). In order to identify the distinctive features and discard common features, a likelihood ratio test is performed on all the features extracted from both the images as shown in equation (1). Those features with likelihood ratio less than one are considered as common features and are discarded.

$$\gamma(f_{i,k}) = \frac{P(H_{i,k} = 1|\Omega)}{P(H_{i,k} = 0|\Omega)} < 1, \quad (1)$$

where $\gamma(f_{i,k})$ is the likelihood ratio characterizing the distinctiveness of feature $f_{i,k}$ when some independent data ($\Omega$) have been observed over this feature. For further details on finding the distinctive features, please refer to Appendix A. Applying this constraint will reduce the number of features by approximately 10 percent. Although this number is small, these non-distinctive features would typically cause errors because of their close similarity with the distinctive features, thereby resulting in many false matches. Fig. 1 shows the results before and after removing the common features in 1(a) and 1(b) respectively, along with the resulting correspondence points found in 1(c) and 1(d) respectively. As expected, wrong correspondence points that do not belong to the face region were found (as shown in 1(c)) when the common features were not removed. These undesired points can disrupt the outcome of the subsequent registration stage, which will be discussed in section 3.2.

After identifying the distinctive feature points, we need to establish correspondences between the feature points in the two images. Many studies have used the RANSAC algorithm [28] to reject outliers and select the correspondence points between two images [27, 31]. RANSAC method performs well when majority of the feature points in one image have real correspondences in the other image. However, in our case, the number of real

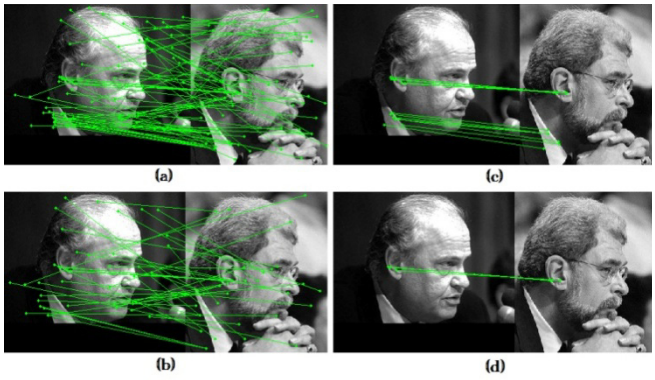



Fig. 1. Effect of not discarding (first row) and discarding (second row) the non-distinctive features in finding the correspondence points between two faces. (a) Result of initial SIFT feature matches using Lowe's method; (b) Result after discarding the common features; (c) Correspondence feature points obtained from (a); (d) Correspondence feature points obtained from (b). Although the number of correspondence feature points obtained in (d) is small compared to (c), the wrong correspondence points are all correctly removed.

correspondences is typically a very small fraction (1-2%) of all the possible matches found. Thus RANSAC method is likely to fail in this scenario [27]. To overcome this problem, we modified the registration method proposed by Goshtasby and Stockman [32] for face detection in multiple poses. Instead of using the boundary of the convex hull to determine the initial transformation parameter as the convex hull varies when the pose changes, we used the SIFT's geometric information of the reference correspondence points found to compute the transformation matrix. After discarding the common features, all potential correspondences between the two images are found by applying Lowe's matching algorithm [23]. Suppose that $M$ potential correspondences are detected. Let $f^G_{i,k(m)} \leftrightarrow f^G_{j,l(m)}$denote the $m^{th}$ potential correspondence, where $k(m)$ is the index of the $m^{th}$ corresponding feature point in $F_i$, $l(m)$ is the index of the $m^{th}$ corresponding feature point in $F_j$ and $m = 1, \dots, M$. These potential correspondences are further filtered using geometric constraint to produce only the salient correspondences. Initially, each matching pair $f^G_{i,k(m)} \leftrightarrow f^G_{j,l(m)}$ is considered a reference correspondence. The geometry of all other matched feature points in each image are normalized with respect to the position, scale and orientation of the reference corresponding point in that image as follows:

$$\begin{bmatrix} x^N_{i,k(n)} \\ y^N_{i,k(n)} \end{bmatrix} = \frac{1}{\sigma_{i,k(m)}} \begin{bmatrix} cos\,\phi_{i,k(m)} & sin\,\phi_{i,k(m)} \\ -sin\,\phi_{i,k(m)} & cos\,\phi_{i,k(m)} \end{bmatrix} \begin{bmatrix} x_{i,k(n)} \\ y_{i,k(n)} \end{bmatrix} - \frac{1}{\sigma_{i,k(m)}} \begin{bmatrix} x_{i,k(m)} \\ y_{i,k(m)} \end{bmatrix}$$

$$\sigma^N_{i,k(n)} = \frac{\sigma_{i,k(n)}}{\sigma_{i,k(m)}},$$

$$\phi^N_{i,k(n)} = \phi_{i,k(n)} - \phi_{i,k(m)}.$$

$$\begin{bmatrix} x^N_{j,l(n)} \\ y^N_{j,l(n)} \end{bmatrix} = \frac{1}{\sigma_{j,l(m)}} \begin{bmatrix} cos\,\phi_{j,l(m)} & sin\,\phi_{j,l(m)} \\ -sin\,\phi_{j,l(m)} & cos\,\phi_{j,l(m)} \end{bmatrix} \begin{bmatrix} x_{j,l(n)} \\ y_{j,l(n)} \end{bmatrix} - \frac{1}{\sigma_{j,l(m)}} \begin{bmatrix} x_{j,l(m)} \\ y_{j,l(m)} \end{bmatrix}$$

$$\sigma^N_{j,l(n)} = \frac{\sigma_{j,l(n)}}{\sigma_{j,l(m)}},$$

$$\phi^N_{j,l(n)} = \phi_{j,l(n)} - \phi_{j,l(m)}. \quad (2)$$

where $(x^N_{i,k(n)}, y^N_{i,k(n)}, \sigma^N_{i,k(n)}, \phi^N_{i,k(n)})$ and $(x^N_{j,l(n)}, y^N_{j,l(n)}, \sigma^N_{j,l(n)}, \phi^N_{j,l(n)})$, $(n = 1, \dots, M; n \neq m)$ are the normalized geometric descriptors of the other matched feature points with respect to the potential reference point. The matched pairs that are geometrically consistent with respect to the reference correspondence $f^G_{i,k(m)} \leftrightarrow f^G_{j,l(m)}$ are found by applying the threshold individually to each geometric attribute as follows.

$$\begin{aligned} C^x_m &= \{n: |x^N_{i,k(n)} - x^N_{j,l(n)}| < T^X_H\}, \\ C^y_m &= \{n: |y^N_{i,k(n)} - y^N_{j,l(n)}| < T^Y_H\}, \\ C^\sigma_m &= \{n: |log\,\sigma^N_{i,k(n)} - log\,\sigma^N_{j,l(n)}| < T^\sigma_H\}, \\ C^\phi_m &= \{n: |\phi^N_{i,k(n)} - \phi^N_{j,l(n)}| < T^\phi_H\}. \end{aligned} \quad (3)$$

where $C^x_m$, $C^y_m$, $C^\sigma_m$ and $C^\phi_m$ are the four groups of indexes to those potential correspondences which satisfy the geometric constraints in the normalized space created by the reference correspondence. The thresholds $(T^X_H, T^Y_H, T^\sigma_H, T^\phi_H)$ control the amount of tolerance to variations between the two images. The lower the tolerance, the higher the degree of match required between the two feature points. This reduces the amount of false matches but also reduces the amount of allowed variations, consequently affecting the ability to support larger pose changes. Normalizing the features to unit length of the reference feature $m$, we allow the position of the match points to vary with half the unit length. Thus we used 0.5 for $T^X_H$ and $T^Y_H$. Similarly, to provide the same tolerance level, we set $T^\phi_H$ as 25 degrees and $T^\sigma_H$ as $log(1.5)$.

The intersection of the four sets of indices in equation (3), $C_m = \{C^X_m \cap C^Y_m \cap C^\sigma_m \cap C^\phi_m\}$, forms the match cluster for the reference correspondence $f^G_{i,k(m)} \leftrightarrow f^G_{j,l(m)}$. Let the size of a match cluster, i.e., the number of elements in $C_m$, be denoted as $s_m$. The procedure is then repeated for all the matched pairs, $f^G_{i,k(m)} \leftrightarrow f^G_{j,l(m)}$, $(m = 1, \dots, M)$, to obtain the cluster that has the maximum value of $s_m$. Let $m^* = argmax_m\,(s_m)$ denote the set of values of $m$ that maximize $s_m$. If $m^*$ has more than one reference correspondence with the same maximum match cluster size, the reference correspondence which results in the lowest Euclidean distance among its underlying matched pairs in the normalized space is taken as the real correspondence as indicated in equation (4).

$$C_R = \begin{cases} C_{m^*}, & if\ |m^*| = 1 \\ C_{m'}, m' = \underset{m \in m^*}{argmax} \left( \sum_{n \in C_m} R^N_{i \leftrightarrow j,n} \right), & if\ |m^*| > 1 \end{cases} \quad (4)$$

where $C_R$ is the cluster of indexes to the real correspondence points between the two images and $R^N_{i\leftrightarrow j,n} = \left(x^N_{i,k(n)} - x^N_{j,l(n)}\right)^2 + \left(y^N_{i,k(n)} - y^N_{j,l(n)}\right)^2$ is the square of the Euclidean distance between the normalized pairs in the underlying cluster $C_m$. For a valid match, we require three points from one plane to be matched to three points from another plane. Since each pair contains information about the location, scale and orientation, a minimum of two pairs of the matching feature points are sufficient to produce a valid match (i.e., the size of the cluster $C_R$ should be greater than or equal to 2). However, a higher number of matched pairs increase the confidence level of the matching operation. This procedure is summarized in Algorithm 1.

**Algorithm 1:**Finding distinctive correspondence points and Reference points.

**Input:** Two face images $i$ and $j$

**Output**: Real correspondence points between two faces

**1**. Extract SIFT features from two images.

**2**. Remove common features from them using appearance descriptor of SIFT.

**3**. Consider the $K$ potential correspondence points $f^G_{i,m} \leftrightarrow f^G_{j,m}$ , $(m = 1,2,\dots,K)$between two images ($i$ and $j$) obtained from the Lowe's matching algorithm,
where $f^G_{i,m} = (x_{i,m}, y_{i,m}, \sigma_{i,m}, \phi_{i,m})$ contains the locations, scale and orientation of the feature $f^G_{i,m}$calculated by the SIFT algorithm.

**4**. **For** $m = 1,2,\dots,K$ **Do**

(a).Assume the feature $f^G_{i,m}$ is the reference point.

(b). Normalize all the geometric descriptors of potential points $f^G_{i,k} = (x_{i,k}, y_{i,k}, \sigma_{i,k}, \phi_{i,k}), (k = 1,\dots,K \,\&\, k \neq m)$in the first image $i$ to the geometric descriptor of the reference point $f^G_{i,m} = (x_{i,m}, y_{i,m}, \sigma_{i,m}, \phi_{i,m})$.

(c). Name the normalized geometric descriptors as $\{f^N_{i,k}\} = \{(x^N_{i,k}, y^N_{i,k}, \sigma^N_{i,k}, \phi^N_{i,k})\}, (k = 1,\dots,K \,\&\, k \neq m)$.

(d). Repeat (a), (b) and (c) for the second image $j$.

(e). Find the cluster of consensus points ($C_m$) with respect to the reference point $m$, which are the points that their normalized geometric descriptors are in vicinity$\left|\{f^N_{i,k}\} - \{f^N_{j,k}\}\right| < T_H^{x,y,\sigma,\phi}$.

**End**

**5**. Consider the pair with the maximum number of inliers $f^G_{i,R} \leftrightarrow f^G_{j,R}$ as the reference pair, where $R = argmax(C_m),\ (m = 1,2,\dots,K))$.

**5.1.** If more than one pair was found as the reference pair, select the one that its inliers in the two images have closer distance to each other $(C_R = argmin_{C_m}\left(\sum_{n\,\in\, C_m} R^N_{i\leftrightarrow j,n}\right))$.

**6**. Return all the pairs $f^G_{i,r} \leftrightarrow f^G_{j,r}$ , $(r\epsilon\ C_R)$ as the real correspondence points and the pair $f^G_{i,R} \leftrightarrow f^G_{j,R}$ as the *reference points* between two images.

## 3.2 Registering Faces from Multi-views

After finding the correspondence points, the next task is to align the various face images differing in position, scale and pose to a reference face. This is needed in order to compute the similarity measure in the overlap region and to determine whether the detected region is indeed a valid face.

Given any two images, the correspondence points obtained using the procedure described in section 3.1 serve as links between these images. Each image has links to its neighborhood images. If the number of images is large enough, it is possible that all the images will have connection to one another through one or more paths among the links. Thus, one or more images will have links to the reference image registered to a *reference map*. The connection among these face images forms a constellation connection as illustrated in Fig. 2. The constellation allows a relationship to be established among the training face images and the reference image.

The *reference map* is constructed from one reference image using two control points (also called *reference points*) as shown in Fig. 3(a). It is recommended that an un-occluded frontal face image with sufficient resolution and minimal background clutter be used as the reference image. The *reference map* encodes the position, scale and orientation of the reference face in the form of unit vectors (represented by the green and white arrows in Fig. 3(a)). The position of the two control points are known as they are selected manually. The control points should be representative and able to withstand pose variations. In all our experiments, the nose base and upper part of the nose in the middle of the two eyes are selected as the control points (see Fig. 3(a)). The reference image is pinned to the *reference map* using these two control points. The next image that contains some correspondence points with the reference image is registered to the *reference map* using the geometric transformation given by the correspondence points. With two images pinned to the *reference map*, a third image that has correspondence points to either the reference image or the second image is then registered. This process is then iterated until all the images having correspondence to at least one image being registered. As more images are registered to the *reference map*, the possibility that another image can be registered increases. Thus, face images of different poses can eventually be processed, with each major pose forming a node in the constellation graph as shown in Fig. 2. As such, the ability of the face detector to handle pose, expression and illumination variations only depends on the availability of training face images in the database. Therefore, there is no limit on the type of facial variations that the system can cope with, as long as the features used to establish the correspondence are robust to local variations encountered in both images to be matched.

### *3.2.1 Reducing Propagation Error*

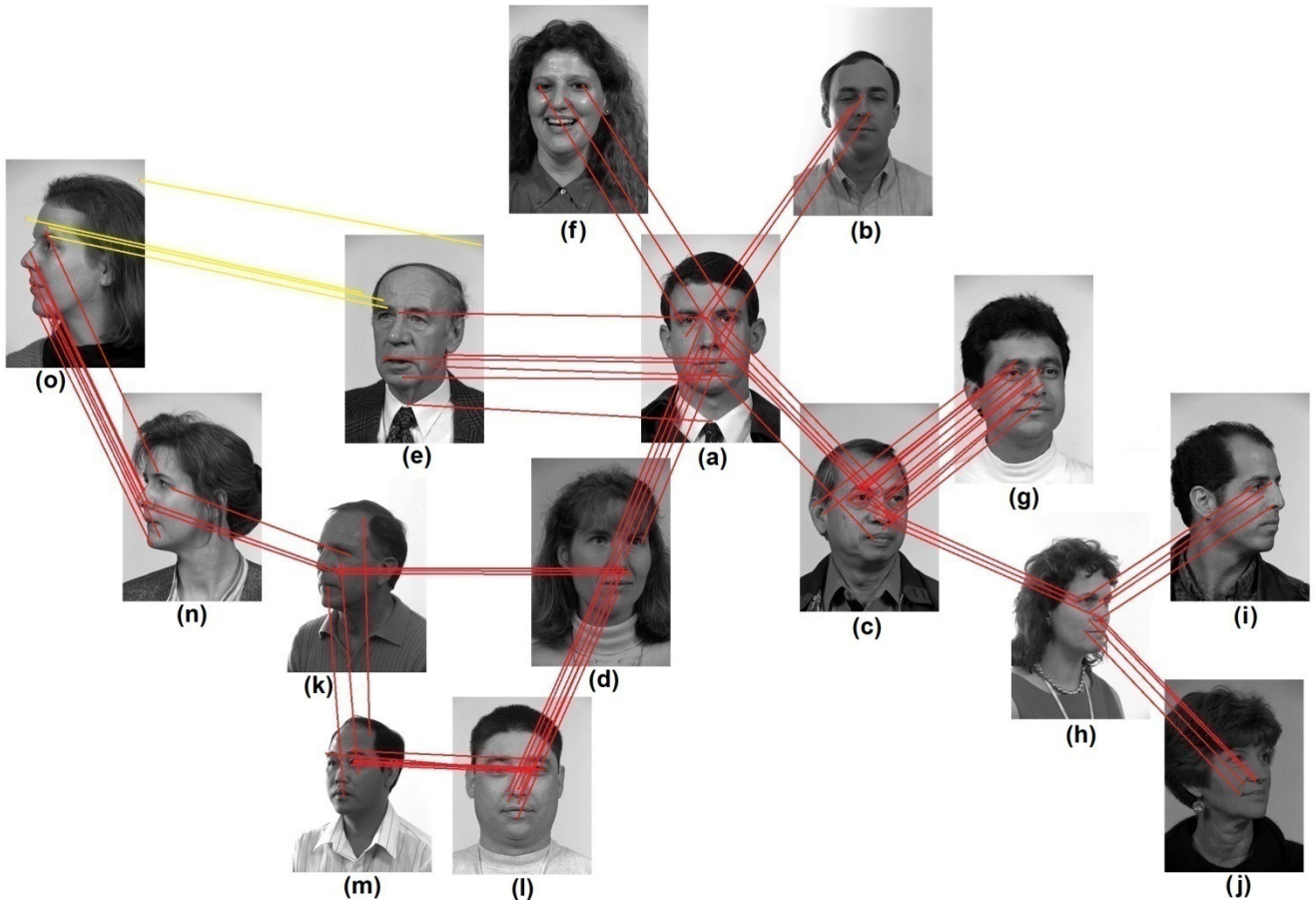



Fig. 2. Constellation formed from links between faces of different poses and scales having reliable correspondences. (a) reference image selected which has frontal pose and has all the facial characteristics (i.e. no occlusion, defect etc.). All other images differing in views, from the left profile pose to oblique and right profile poses, are then mapped to the *reference map* by finding a path that links it to the reference image. Image (l) is registered to the *reference map* via its link with image (d) and image (a). Such links form the path for registration. As all the correspondences are correct, the link is strong. Image (o) is registered to the *reference map* through image (e) and image (a). However, there are wrong correspondences between (o) and (e), resulting in a weak link (yellow lines). Such an error will cause wrong registration for image (o) and all images connected to it, such as image (n),the lady's face image in left profile view, if (n) is linked to (a) via the path (o) and (e). Thus, if the weak link can be identified and removed, then (o) can be registered through the path {(a),(d),(l),(m),(k),(n),(o)}, resulting in correct registration.

As a face image is registered to the *reference map* based on the correspondence points found with another face image that is already registered, any error in the correspondence will result in the error being propagated as more faces are being registered to this link. To avoid this problem, we propose a mechanism to remove the weak links from the constellation, since such links have a higher chance of error. We noticed that weak links usually arise from correspondences found on the non-face region. Thus we define a likelihood test to determine the degree of links as follows: Let $F = \{f_k\},\ k = 1,2,\dots,K$ be the collection of feature points in a given image, $\beta$, which have connection between the image and the other images in the constellation. To detect the weak links, the likelihood ratio $\vartheta(f_k)$ is computed for all features $(f_k)$ in $F$ as shown in equation (5). Let $H_k$ denote a binary random variable indicating whether the feature $f_k$ originates from a part of face $(H_k = 1)$ or not $(H_k = 0)$. This likelihood ratio should be greater than one for valid face features as shown in equation (5).

$$\vartheta(f_k) = \frac{p(H_k = 1|\psi)}{p(H_k = 0|\psi)} > 1. \tag{5}$$

The above likelihood ratio is computed from a sequence of $M$ independent observations of data $\psi$ when forming the constellation. Thus $H_k$ follows a Bernoulli distribution [33]. Note that the number of correspondence between two images in the constellation is not high (typically less than 50 in our experiment). Thus the number of valid correspondences is higher than the noise since most of the wrong correspondences have been removed earlier and that the reference image has minimal background clutter. Therefore, to evaluate the likelihood ratio in (5), we use Bayesian estimation. We consider uncertainty on the distribution parameter $\theta$ and use a *Beta* probability distribution as a prior with parameters $\alpha_1$ and $\alpha_0$ as indicated in equation (6). For further details, please refer to Appendix B.

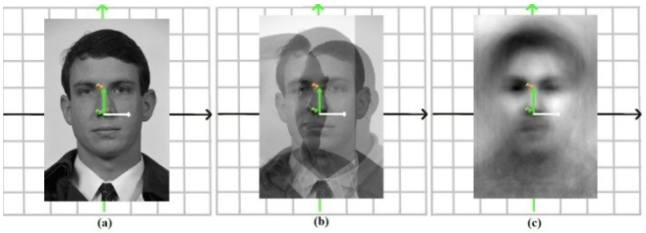


Fig. 3. (a) *Reference map* shaped from one frontal view as reference image and two manually selected control points that pin the reference image to the *reference map*. Once the reference image is pinned to the *reference map*, the transform function mapping the other images of different poses and scales to the reference image can be obtained based on their links to the reference image; (b) Composition of left oblique and right profile images registered to the *reference map*; (c) Composition of all images with paths linking them to the reference image during the training stage and registered to the *reference map* to form the final model.

$$P(\theta|\alpha_1,\alpha_0) = Beta(\theta|\alpha_1,\alpha_0) = \frac{\Gamma(\alpha_1+\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}, \quad (6)$$

where $\Gamma(x)$ is the gamma function:

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du, \quad (7)$$

Thus, the posterior prediction for random variable $H_k$, which indicates if the feature $f_k$ comes from a part of the face, is given by:

$$P(H_k = 1|\psi) = \int_0^1 \theta\, P(\theta|\psi)\, d\theta = \int_0^1 \theta Beta(\theta|L_k+\alpha_1, L-L_k+\alpha_0)\, d\theta = \frac{L_k+\alpha_1}{L+\alpha_1+\alpha_0}, \quad (8)$$

where $L$ is the total number of images used to form the constellation (considered as the training stage) and $L_k$ is the number of images in the constellation where feature $f_k$ can establish a connection to image $\beta$. For cases where feature $f_k$ comes from the non-face region, we consider it as a white noise with average value $\bar{w}$. This implies that the average number of times it can make a connection with different images is $\bar{w}$. This value cannot be large since noise features have random appearance and thus does not possess strong connections between many images under the appearance and geometric constraints. Thus, its likelihood probability can be computed in the same way as (8), which is shown in (9).

$$P(H_k = 0|\psi) = \frac{\bar{w}+\alpha_0}{L+\alpha_1+\alpha_0}, \quad (9)$$

Substituting (8) and (9) into (5) indicates that the number of times the face feature $f_k$ has participated in establishing connection with the images in the training set should be greater than a threshold ($L_k > (\bar{w} + \alpha_0 - \alpha_1) = Thr_w$) for it to be considered as a valid face feature. Otherwise, it is a weak link and should be discarded. In setting the threshold $Thr_w$, a small value will increase the noise while a large value will unnecessarily remove valid features. We empirically observe that a value of 5 is appropriate for $Thr_w$. Alternatively, if the likelihood ratio in (5) is ranked, then it is not necessary to set the threshold $Thr_w$. After finding the distinctive correspondence points among all the images in the training set and computing the likelihood ratio in equation (5), rank all the links between image pairs in descending order. We can then start to establish the connection between images that have higher degree of match. Consequently the constellation center will comprise images having strong links while the edges of the constellation will be filled with weak links. Hence, any error will usually not propagate through to the center of the constellation but instead is limited to some inferior links at the edges of the constellation and hence does not affect the results significantly.

Given a test image, we can compute the number of matches of the features obtained with that of the constellation. If there is sufficient match, then the face is detected. However, this is limited to only one face per test image and is slow. The next section will address these issues.

## 4 Detecting Multiple Faces of Arbitrary View

In this section, we generalize the proposed face detection algorithm to handle multiple faces of arbitrary views. In our proposed approach, it is possible to approximately locate the features belonging to specific facial traits once a given image is registered. Therefore, to detect multiple faces, we need to cluster all the features that approximately form a face region.

The posterior probability function $P(\rho|F)$ gives the probability that the set of features $F = \{f_1, f_2, \dots, f_K\}$ completely register to a *reference map*, ($\rho$). Applying the Bayes theorem gives the prior probability $P(\rho)$ and likelihood $P(\{f_1 f_2 \dots f_K\}|\rho)$ as:

$$P(\rho|F) = \frac{P(\rho)P(F|\rho)}{P(F)}, \quad (10)$$

where $P(\rho)$is the prior probability and $P(F|\rho)$ is the likelihood of observing the features $F = \{f_1, f_2, \dots, f_K\}$ given a *reference map* ($\rho$). To simplify the evaluation of the likelihood term $P(F|\rho)$, we assume that the discriminative features of face images are independent. Hence, the equation (10) can be re-written as:

$$P(\rho|F) = P(\rho) \times \prod_{k=1}^{K} \frac{P(f_k|\rho)}{P(f_k)}, \quad (11)$$

where $P(f_k|\rho)$ is the likelihood term indicating the probability that the feature $f_k$ is a part of a face (i.e., $P(H_k = 1)$). This term is similar to the equation (8). $P(\rho)$ is the prior probability of the *reference map* and is a constant scaling factor that can be factored in the formulation and hence not evaluated. The term $P(f_k)$ is the prior probability of face features. This term represents the spatial distribution of the face features around the *reference map* and can be modeled using a Gaussian 2D distribution given by:

$$P(f_k) = \frac{1}{2\pi\delta_x\delta_y} e^{-\left(\frac{(x_k-x_0)^2}{2\delta_x^2}+\frac{(y_k-y_0)^2}{2\delta_y^2}\right)} \quad (12)$$

where $(x_k, y_k)$ is the position of feature $f_k$ and $(x_0, y_0)$ is the position of the center point between the two *reference points*. $\delta_x$ and $\delta_y$ indicate the spread of the face features and is dependent on the size of the face in the given image. In all our experiments, we approximate $\delta_x = \delta_y =$ (1.5 × *reference map* scale).

During face detection, all features of the test image $\beta$ are extracted using the SIFT method. Each feature in the test image is compared against the features of all images in the face constellation. An image feature $f_{\beta,m}$ is considered to be similar in appearance to the feature $f_{i,k}$ from the $i^{th}$ image in the constellation if the Euclidean distance between their appearance descriptors $f^A_{\beta,m}$ and $f^A_{i,k}$ is less than a threshold $T^{\partial}_{i,k}$. This threshold is not a global value and has to be determined individually for every feature in the constellation. This is done at the training stage. Given a feature $f_{i,k}$ from image $i$ is matched to the feature $f_{j,l}$ from image $j$, the appearance threshold $T^{\partial}_{i,k}$ for feature $f_{i,k}$ is the Euclidean distance between $f^A_{i,k}$ and $f^A_{j,l}$. If $f_{i,k}$ from image $i$ is involved in many matched pairs, the appearance threshold $T^{\partial}_{i,k}$ is given by the maximum Euclidean distance between $f^A_{i,k}$ and the appearance descriptors of the other features that match with $f_{i,k}$. Let $F_m$ be the set of features in the constellation that have high appearance similarity to the image feature $f_{\beta,m}$. This process is repeated for all the features in test image $\beta$ and the collection of all matching features in the constellation $F = \cup_{m=1}^{M} F_m = \{f_1, f_2, \dots, f_K\}$ is obtained. Each matching feature $f_k$ is used to estimate a *reference map* in the test image. The estimated *reference map*s are clustered using mean-shift algorithm [34]. Other clustering methods that do not require a priori knowledge about the number of clusters can also be used. In addition, clustering reduces the processing time as comparison is only required against one feature per cluster instead of all features present. Let cluster $\chi$ has $K_\chi$ features. Each cluster center can be considered as a potential *reference map* ($\rho_\chi$). Among these potential *reference map*s, the valid *reference map*s can be obtained as follows:

$$P(\rho_\chi|F) > Thr_F, \quad (13)$$

which is reformulated using (11) as:

$$\prod_{k\in\chi} \frac{P(f_k|\rho_\chi)}{P(f_k)} > \frac{Thr_F}{P(\rho_\chi)}, \quad (14)$$

The right hand side of equation (14) can be replaced by another constant threshold. While evaluating $P(f_k|\rho_\chi)$, we consider a uniform prior for *Beta* distribution ($\alpha_1$=$\alpha_0$=1) in equation (8) and substituted $L_k$ value for features in the constellation that have at least one match in cluster $\chi$. For those features in the constellation that do not have any match in cluster $\chi$, $L_k$ is set to zero. The computed value for each cluster is then compared to a threshold to classify it as a face or non-face, thereby yielding multiple detected faces. This threshold is determined experimentally such that it minimizes the classification error.

The probabilistic framework also gives the ability to handle partially occluded faces and faces with highly cluttered background because it does not require finding features from all parts of a face. Occluded faces can still be detected if distinctive features from some parts of the face satisfying equation (14) can be found. Since the distribution of the features from the background clutter is usually more dispersed compared to the face, the likelihood that sufficient features in a cluster from the background but similar to a face is rather low. Therefore, it is possible to detect faces even in highly cluttered background.

## 5 EXPERIMENTS

We evaluated our proposed algorithm using the standard color FERET dataset [35, 36], the CMU face image dataset [38] and the Face Detection Dataset and Benchmark (FDDB) [37]. FERET dataset contained images of 1,209 persons of various age, gender and ethnicity. Each person had an average of 10 images, with different illumination conditions and pose variations. The number of images in each viewpoint was not uniform. For instance, a person in the dataset may have more than three frontal faces but only one full profile face. We resized all the images to 256x384 pixels. The CMU dataset contained images acquired under natural settings. The image sizes and resolutions were not uniform. Some face images were very small. Faces in this dataset had high background clutter and each image contained more than one face from different viewpoints. In the FDDB dataset, 2,845 images were taken from the faces in the wild dataset [39] and separated into 10 folds randomly. All the 2,845 images were captured under natural settings and contain 5,171 faces annotated manually as a benchmark. Faces in this dataset were very challenging with cases of occlusion, varying poses, low resolution, and out of focus. For all the above datasets, we converted the color images into gray scale.

Fig. 4. Correspondence points found between the image pairs of the same view (Frontal view); (a) Potential correspondence points obtained using Lowe's method [23]; (b) Ground truth found manually; (c), (d) and (e) RANSAC method with different models and (f) Our method.

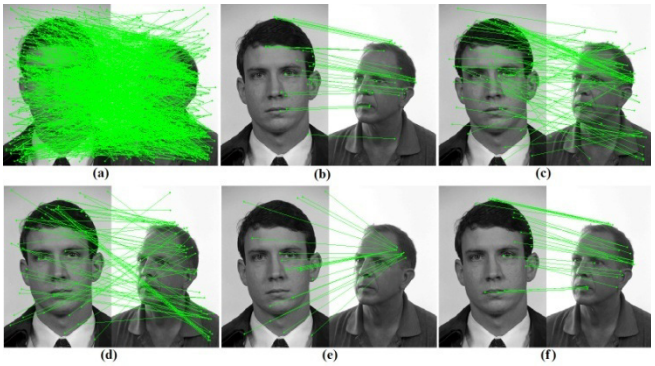


Fig. 5. Correspondence points found between the image pairs of different views; (a) Potential correspondence points obtained using Lowe's method [23]; (b) Ground truth found manually; (c), (d) and (e) RANSAC method with different models and (f) Our method.

### 5.1 Setup

We used SIFT features [23] and applied the proposed method outlined in section 3 on all the image-pairs. The SIFT parameters such as the number of octaves of the Gaussian scale space and the number of scale levels within each octave were determined such that at least 1,000 features were extracted from each image. Potential correspondence points between the image-pairs were found using Lowe's method [23]. Based on Matlab implementation and using a 2.66GHz C2Q PC with 3 GB RAM running Windows Vista 32-bit OS, the entire process to locate the correspondence points took less than a second for an image pair of size 256x384 pixels.

### 5.2 Finding Distinctive Correspondence Points

In order to evaluate the ability of our proposed method to find the distinctive correspondence points among the many false matches, we arbitrarily chose two images of the same view (Fig. 4) and another of different view (Fig. 5). Figures 4(a) and 5(a) show the results after applying the Lowe's method [23]. We obtained the ground truth data shown in Fig. 4(b) and 5(b), by manually checking all the correspondences in (a). For comparison, we also implemented the RANSAC 8-points with Fundamental constraint (Fig. 4(c), 5(c)), RANSAC with Affine constraint (Fig. 4(d), 5(d)) and RANSAC with Euclidean constraint (Fig. 4(e), 5(e)). Results obtained using our proposed method are shown in Fig. 4(f) and 5(f). The results indicate that for image pairs with the same view, our method is able to find the maximum number of correspondence points with respect to the ground truth data. For image pairs with different views, our proposed method yields the least number of false correspondence points.

To measure the quantitative accuracy, we use the mutual informative measure described in [40, 41]. Mutual information between the predicted label and the ground truth label is given by:

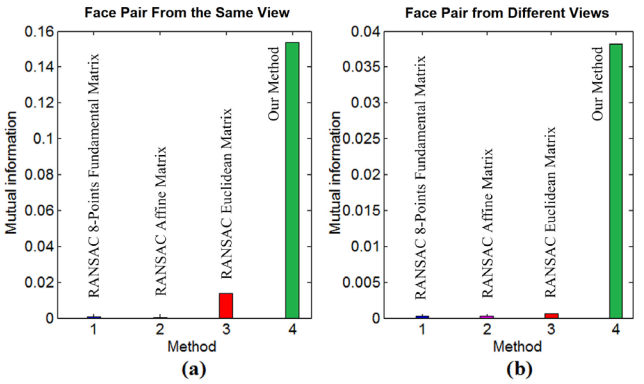



Fig. 6. The overall performance of finding correspondence points measured using mutual information criteria [40, 41]. (a) image pair with the same view, (b) image pair with a different view.

$$I(y,t) = \sum_{y=0}^{1}\sum_{t=0}^{1} p(y,t)\, log \frac{p(y,t)}{p(y)p(t)}, \qquad (15)$$

where $p(y=1,t=1)$ represents the true positive normalized with the number of test points, $p(y=1,t=0)$ represents the normalized false positives and so on. The two individual terms can be calculated by applying summation in one variable, for instance $p(y) = \sum_{t=0}^{1} p(y,t)$.

Fig. 6 shows the results obtained on image pairs with the same view (Fig. 6(a)) and different views (Fig. 6(b)). Our method outperforms the other approaches in both cases. This implies that our proposed method is capable of finding distinctive correspondence points even in the presence of a large number of false matches.

### 5.3 Face Registration

This experiment was carried out on the FERET dataset. We randomly chose 500 images of different people, with one image per person of a random view, to form the initial training set. We then applied our proposed method to register these images. We took a frontal view image as the reference image and manually selected the nose base and upper part of nose in the middle of two eyes as the

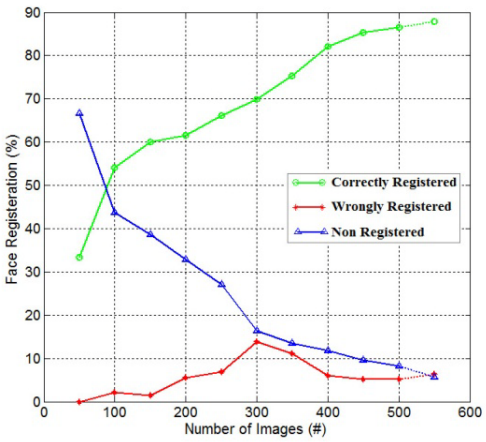



Fig. 7. Percentage of face registration versus the number of images used. The percentage of correctly registered images increases when the number of images increases while the percentage of non-registered images decreases. The rate of wrongly registered images is kept low by the propagation error prevention algorithm. It is rather constant at about 6% after 400 images are used for training. This low error percentage does not affect the model construction significantly.

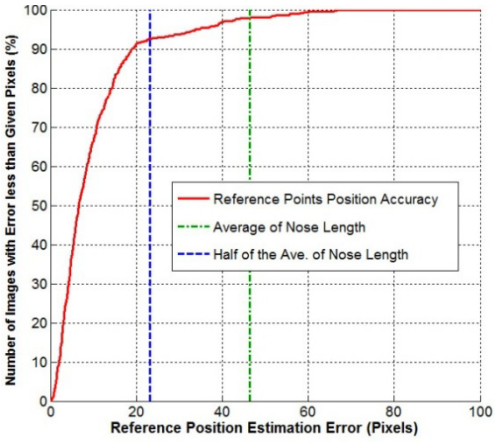



Fig. 8. Cumulative plot of the *reference point*'s location estimation accuracy. Analysis is done for all the images that are registered to the *reference map*, but excludes those that cannot be registeredby our method.

*reference points* for registration. Marking these two points is the only manual intervention needed. Starting with 50 training images to form the initial constellation of links, we evaluated the percentage of images which were successfully registered, wrongly registered and could not be registered over the total number of images available in the training dataset. An image was considered as correctly registered when the estimated reference vector falls within a circle centered at the base of the nose, with a radius of half of the nose length (distance between the nose base and upper part of the nose in the middle of the two eyes). We then repeated the procedure by increasing the number of training images until all the available images were used. The registration result is shown in Fig. 7.

We manually checked all the images using the same criteria as above to ensure correct registration and found that only 6% of the images were wrongly registered after all 500 images were added to the initial training set. Those images without any link to the reference image were classified as non-registered faces. The results confirm that as the number of training images increased, the number of correctly registered images with different viewpoints also increased. Thus, the probability of finding a path to the reference image increased, resulting in successful registration.

Fig. 8 shows the cumulative plot of the accuracy of estimating the *reference points* in the 500 training images. The horizontal axis shows the error in estimating the position of the *reference points*, which is obtained by computing the Euclidean distance between the estimated *reference points* and the ground truth data (marked manually in all the images). The vertical axis shows the cumulative percentage of images having estimation error less than the given number of pixels in the horizontal axis. The number of images is normalized by the total number of registered images (440 images out of 500 as shown by the solid lines in Fig.7). The plot shows that our algorithm is able to estimate the *reference points* to within half of the average nose length in more than 90% of the images. This is achieved despite using only one manually registered image and that the dataset contains faces with large pose variations. In our implementation, the nose length is used to normalize the images.

### 5.4 Detecting Multiple Faces with Arbitrary Views

In this experiment, we compared our approach with the method by Toews and Arbel [18, 42]. Similar to [18], 500 images of different people in random viewpoint were randomly selected from the FERET database as the training set. The significant difference between our method and [18] is that their method requires manually selecting all the control points in all the 500 training images. In our case, we only manually selected two control points on only one image (the reference image). Thus unlike [18], the amount of manual intervention needed in our proposed approach is independent on the number of training images used. For testing, we selected ten different random folds, each with 500 images of different persons not in the training set. Again, one image per person from a random viewpoint was used. The average Receiver Operating Characteristic (ROC) curve obtained is shown in Fig. 9. The ROC curve was obtained by changing the threshold value of the probabilistic classifiers in equation (14) from zero to infinity. Reference vectors, which fall within a circle centered at the reference point with a radius of half the distance of the two control points, and scale and orientation within 20% of the reference vector, were considered as correct detection. The results obtained show that our method outperforms Toews and Arbel [18] by an average of 12% lower equal error rate. Based on this curve, we computed the threshold corresponding to the equal error rate (EER) point and use this threshold in the equation (14) for all the subsequent experiments.

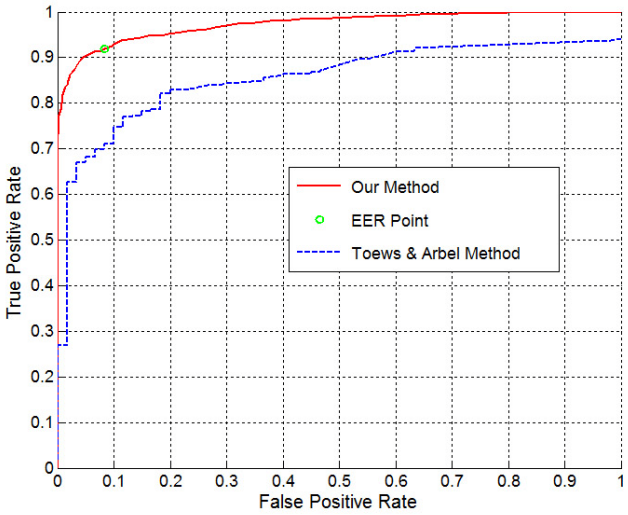



Fig. 9. ROC curve comparing our method and Toews and Arbel method [18] on 500 randomly selected test images of different views from the FERET dataset. Equal Error Rate (EER) point shown in this graph has been used to determine the optimum threshold in (14).

We further evaluate our method trained using the FERET dataset, on the CMU face dataset [38]. Sample results are shown in Fig. 10. Most of the undetected faces are due to low resolution (or small faces). For such faces, the size is close to or below the lower bound of the scale of SIFT operator. Therefore, the features extracted are not distinctive and this affected the accuracy of our face detector. Another reason for the failure is due to an insufficient number of training images, causing failure to find the appropriate link to the reference face. We found that in the FERET dataset used for training, there is an insufficient number of left profile or left oblique faces, causing the model built to be incomplete. As a result, such faces cannot be detected as seen in Fig. 10(a) and 10(d). However, after adding 50 randomly selected images of left profile and left oblique faces from the FERET dataset to the training stage, such images can be successfully detected as seen in Fig. 11. Adding this amount of training images also increased the number of correctly registered images, while decreasing the number of non-registered faces as illustrated by the added dash lines at the tails of the curves in Fig. 7.

The third dataset used to evaluate our method is FDDB [37]. We used the same data as Jain and Learned-Miller [43]. We compared the faces detected with the ground truth data and the degree of match is obtained by calculating the ratio of an intersecting area using:

$$S(D,G) = \frac{Area(D) \cap Area(G)}{Area(D) \cup Area(G)}, \qquad (16)$$

where $D$ is the detected face and $G$ is the ground truth face. $Area(\,)$ indicates the area of face which typically is defined by an ellipse or a square window. Those detected faces with match values, $S(D,G)$, greater than 0.5 are considered as a correctly detected face in the discrete score. In the continuous score, the direct value of $S(D,G)$ is used. For further details on the evaluation protocol, please refer to [37]. Fig. 12 shows the comparison of our

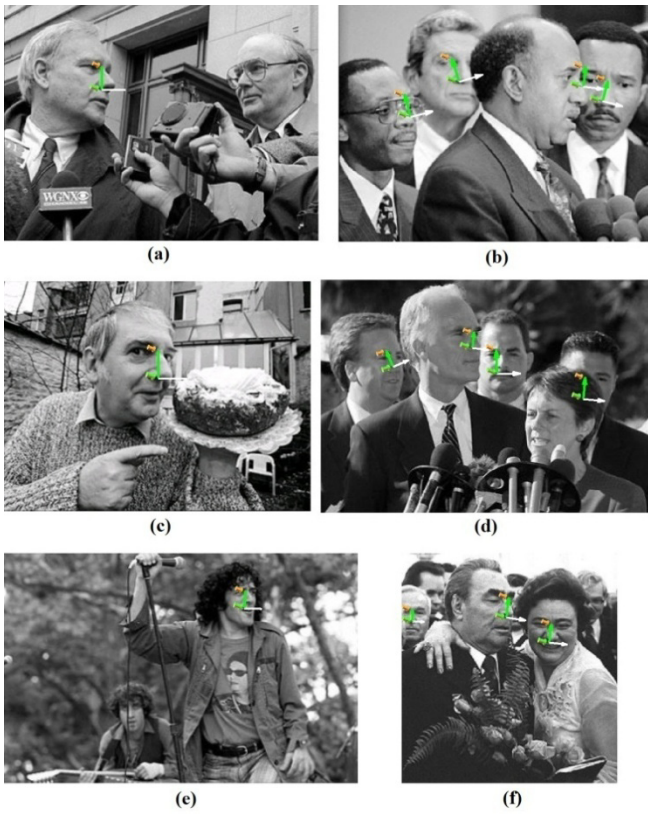



Fig. 10. Sample results of our proposed method applied to the CMU database. The results show that this framework is invariant to viewpoints, occlusion and scale changes. Some faces in the images are not detected. The small faces are not detected due to poor resolution of these faces, causing SIFT feature extraction to fail. For those left profile faces not detected, the number of training images available in the dataset was not sufficient to cover these views.

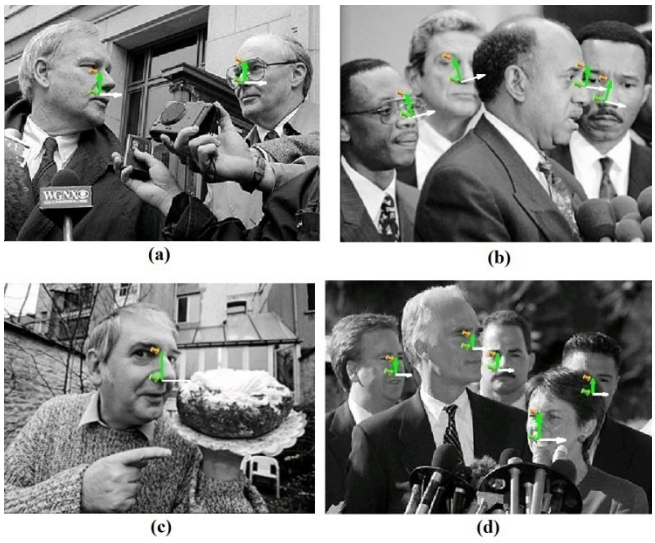
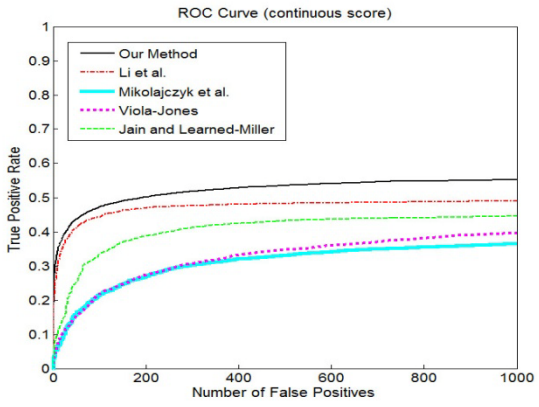
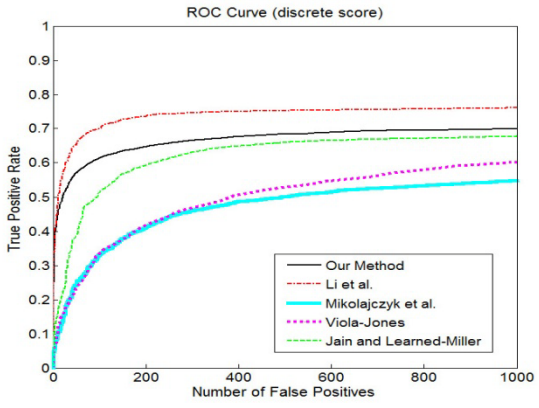



Fig. 11. Sample results of our proposed method after adding 50 random images of left profile and left oblique faces in the FERET dataset to the training stage. It shows that almost all images can be successfully detected.




(a) Continuous Score




(b) Discrete Score

Fig. 12. ROC curves obtained using the evaluation method proposed in [37] on the FDDB dataset, showing both (a) Continuous score and (b) Discrete score. The proposed method is compared against Li. et al. [44], Jain and Learned-Miller [43], Viola and Jones [5] and Mikolajczyk et al. [45].

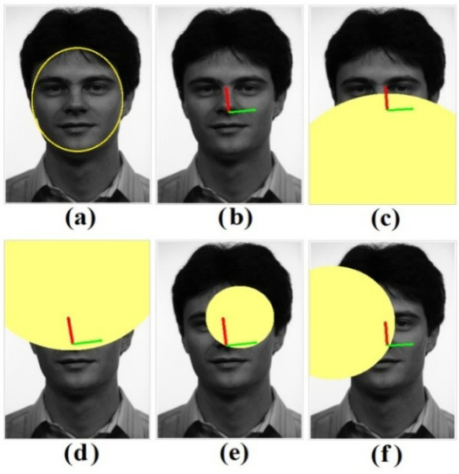



Fig. 13. Detection under occlusion. (a) Original face image.. (b) Detection and localization result without any occlusion. (c) – (f) The largest occlusion before the detection fails when the size of the occluding disk is increased from bottom, top, center and left/right (symmetrical for frontal face) respectively.

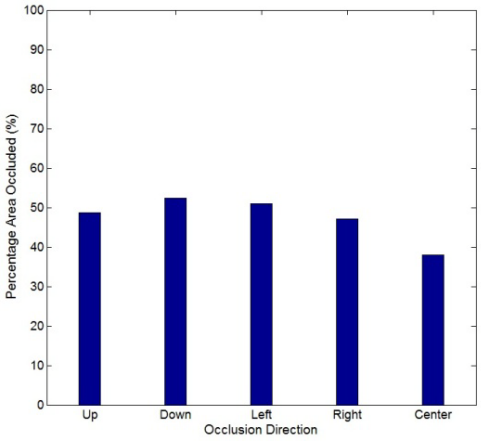



Fig. 14. Percentage of occlusion in the 500 images taken from FERET dataset just before the face detection fails.

method with the other reported methods that have conducted evaluation on the FDDB database. Our approach performs better than the other reported methods in the continuous score category because it is able to locate faces more accurately. Our approach automatically determine the *reference points* in the faces, while the other methods only return a window that included a part of the detected face. However, in discrete score, our method is not as good as Li et al. [44]. This is partly because we did not perform any post-processing to validate the face region, but merely draw an ellipse with parameters similar to the reference image and centered it at the midpoint of the two *reference points* to meet the requirement of the evaluation. Thus, for non-frontal faces, the estimated ellipse will cover some background regions. In the extreme case of a profile face, the ellipse drawn will only cover about half the face. Thus achieving greater than 50% overlap is not possible which lowers our performance. Nevertheless, the performance is still comparable as the other reported methods could not detect non-frontal or occluded faces. It should be emphasized that our method is trained with only 500 FERET images. Even then, we only manually label one image with two points, whereas the other methods require significantly more positive and negative samples to train, thus requiring a large number of manual labeling.

### 5.5 Occlusion Analysis

In order to evaluate the robustness of our proposed approach against occlusion, we used a circular disk to occlude the face and then grow the disk gradually. A random face in frontal view different from the training set (Fig. 13(a)) is selected from the FERET dataset. The disk is then placed at four locations: top, bottom and left or right mid-points and center of the image, to minimize any ambiguity due to features chosen and the exact position of the face. Fig. 13(b) shows the result without any occlusion. The maximum level of occlusion which our method is able to tolerate is shown in Fig. 13(c), (d), (e), and (f) respectively for different disk's location. The ratio of the overlapping area between the disk and the face region relative to the total face area is computed as a percentage. Using this protocol, we conduct the experiment on 500 randomly selected FERET dataset images. We plot the percentage for all the four locations

just before the detection fails (see Fig. 14). This shows that the proposed approach can tolerate about 40% to 50% occlusion, as long as sufficient distinctive features (typically eyes and mouth regions that have high similarity for most faces) can be detected to establish the correspondence.

### 5.6 Computation Time

We implemented the face detection and localization portion using C++. The average time to detect and localize a face in an image of 320×240 pixels was about 540ms using an ordinary PC (2.8GHz C2D PC with 3 GB RAM and Linux OS using a single core). As a comparison, the execution time for the window-based method of Huang et al. [8] and Li and Zhenqiu [7] is 250~330ms (using a Pentium 4-3.0GHz), and 200ms (using a Pentium 3-700MHz) respectively. However, these methods yield only coarse detection, requiring face alignment to localize the face. On the contrary, our approach performs both detection and localization, providing the position, scale and amount of in-plane rotation of a detected face simultaneously. Thus, the computational time of our method is not significantly slower. By contrast, our method is up to 7 times faster than that of Toews and Arbel [18, 42] on the same machine. This is because our method is able to select more informative and distinctive features, resulting in a more compact and efficient model construction.

The majority of the required time is spent on SIFT feature extraction (480ms) [46]. The time needed for comparing the model and the extracted features is not significant. Based on our experiments, with 500 training images, we have about 8,000 features in the constellation. After mean-shift clustering, the number is reduced to about 800 features. The comparison time between the 800 features in the constellation and the extracted features is only about 60ms. In addition, our method has a huge time advantage in the training or model construction stage. The window-based methods and Toews and Arbel [18, 42] require all the training images to be manually labeled, which is a very time consuming task. For example, Huang et al. [8] reported labeling 30,000 frontal, 25,000 half-profile and 20,000 full-profile faces manually. However, our approach requires only manually labeling two control points in only one image, the reference image. There is no further manual intervention needed.

## 6 Conclusion

In this paper, we have presented an approach to simultaneously detect and localize multiple faces from arbitrary views and different scale images. We have shown that the proposed method is able to handle arbitrary views, varying illumination conditions and complex backgrounds. Advantageously, the probabilistic framework is able to detect multiple and occluded faces. To prevent any error propagation due to registering wrongly matched features, weak links are detected and removed. Unlike other approaches, the proposed approach does not require manually labeled faces apart from two points in the reference face image for training. Despite its simplicity, experimental results show that the performance is better than the other state-of-the-art approaches for multi-view face detection. Note that the proposed method does not have any prior assumptions about the training images used. More diverse training images enrich the face constellation, allowing it to detect more complex variations of the face. Thus, the capability of the proposed approach will only grow with more diverse images, so long as the intra-class variation is not too large that sufficient correspondence points cannot be found.

Our future work is to extend the proposed approach to detect other object classes and to solve the problem of establishing correspondences when the intra-class variations are large.

## Acknowledgment

This research was supported by A*STAR and the Institute for Infocomm Research, Singapore. The authors would also like to thank Karthik Nandakumar, Jessica Ng, Mark David Rice and Chuohao Yeo for their assistance in improving the paper.

## References

[1] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image Vis. Comput.*, vol. 21, no. 11, pp. 977-1000, 2003, DOI: 10.1016/s0262-8856(03)00137-9.

[2] J. Wright, A.Y. Yang, A. Ganesh, et al., "Robust Face Recognition via Sparse Representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210-227, 2009, DOI: 10.1109/TPAMI.2008.79.

[3] Z. Baochang, G. Yongsheng, Z. Sanqiang and L. Jianzhuang, "Local Derivative Pattern Versus Local Binary Pattern: Face Recognition With High-Order Local Pattern Descriptor," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 533-544, 2010, DOI: 10.1109/TIP.2009.2035882.

[4] M. Meytlis and L. Sirovich, "On the Dimensionality of Face Space," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 7, pp. 1262-1267, 2007, DOI: 10.1109/TPAMI.2007.1033.

[5] P. Viola and M.J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137-154, 2004, DOI: 10.1023/B:VISI.0000013087.49260.fb.

[6] H. Schneiderman and T. Kanade, "Object detection using the statistics of parts," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 151-177, 2004, DOI: 10.1023/B:VISI.0000011202.85607.00.

[7] S.Z. Li and Z. Zhenqiu, "FloatBoost learning and statistical face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1112-1123, 2004, DOI: 10.1109/TPAMI.2004.68

[8] C. Huang, H.Z. Ai, Y. Li and S.H. Lao, "High-performance rotation invariant multiview face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 671-686, 2007, DOI: 10.1109/TPAMI.2007.1011.

[9] G. Dedeoglu, T. Kanade and S. Baker, "The Asymmetry of Image Registration and Its Application to Face Tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 5, pp. 807-823, 2007, DOI: 10.1109/TPAMI.2007.1054

[10] A. Gee and R. Cipolla, "Determining the Gaze of Faces in

Images," *Image Vis. Comput.*, vol. 12, no. 10, pp. 639-647, 1994, DOI: 10.1016/0262-8856(94)90039-6.

[11] L. Xiaoming, "Generic Face Alignment using Boosted Appearance Model," *Proc. Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1-8, 2007.

[12] L. Gu and T. Kanade, "A Generative Shape Regularization Model for Robust Face Alignment," *Proc, European Conf. Computer Vision, Pt I,* pp. 413-426, 2008.

[13] Z. Jianke, G. Luc Van and S.C.H. Hoi, "Unsupervised face alignment by robust nonrigid mapping," *Proc. Int. Conf. Computer Vision,* pp. 1265-1272, 2009.

[14] L. Xiaoming, "Discriminative Face Alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1941-1954, 2009, DOI: 10.1109/TPAMI.2008.238

[15] A.U. Batur and M.H. Hayes, "Adaptive active appearance models," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1707-1721, 2005, DOI: 10.1109/TIP.2005.854473.

[16] A.B. Ashraf, S. Lucey and C. Tsuhan, "Learning patch correspondences for improved viewpoint invariant face recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1-8, 2008.

[17] L. Liang, R. Xiao, F. Wen and E. Sun, "Face Alignment Via Component-Based Discriminative Search," *Proc. European Conf. Computer Vision, Pt I,* pp. 72-85, 2008.

[18] M. Toews and T. Arbel, "Detection, Localization, and Sex Classification of Faces from Arbitrary Viewpoints and under Occlusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1567-1581, 2009, DOI: 10.1109/TPAMI.2008.233.

[19] G.B. Huang, V. Jain and E. Learned-Miller, "Unsupervised Joint Alignment of Complex Images," *Proc. Int. Conf. Computer Vision,* pp. 1-8, 2007.

[20] C.P. Papageorgiou, M. Oren and T. Poggio, "A general framework for object detection," *Proc. Int. Conf. Computer Vision,* pp. 555-562, 1998.

[21] T. Ojala, M. Pietikainen and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971-987, 2002.

[22] K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points," *Proc. Int. Conf. Computer Vision,* pp. 525-531, 2001.

[23] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91-110, 2004, DOI: 10.1023/B:VISI.0000029664.99615.94.

[24] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615-1630, 2005, DOI: 10.1109/TPAMI.2005.188

[25] H. Bay, T. Tuytelaars and L. Van Gool, "SURF: Speeded up robust features," *Proc. European Conf. Computer Vision, Pt 1,* i pp. 404-417, 2006.

[26] S.J. Thomas, B.A. MacDonald and K.A. Stol, "Real-Time Robust Image Feature Description and Matching," *Proc. Asian Conf. Computer Vision*, pp. 334-345, 2011.

[27] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63-86, 2004, DOI: 10.1023/B:VISI.0000027790.02288.f2.

[28] M.A. Fischler and R.C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381-395, 1981, DOI: 10.1145/358669.358692.

[29] J.C. Chen and J.J.J. Lien, "A view-based statistical system for multi-view face detection and pose estimation," *Image Vis. Comput.*, vol. 27, no. 9, pp. 1252-1271, 2009, DOI: 10.1016/j.imavis.2008.11.004.

[30] G. Dorko and C. Schmid, *Selection of scale-invariant parts for object class recognition*, IEEE Computer Soc, p. 634-640, 2003.

[31] O. Chum and J. Matas, "Optimal Randomized RANSAC," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1472-1482, 2008, DOI: 10.1109/TPAMI.2007.70787

[32] A. Goshtasby and G.C. Stockman, "Point Pattern-Matching Using Convex-Hull Edges," *IEEE Trans. Syst. Man Cybern.*, vol. 15, no. 5, pp. 631-637, 1985.

[33] K. Murphy, *Machine Learning: a Probabilistic Perspective*, MIT Press, in preparation, August 2012.

[34] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603-619, 2002, DOI: 10.1109/34.1000236

[35] P.J. Phillips, H. Moon, S.A. Rizvi and P.J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090-1104, 2000, DOI: 10.1109/34.879790.

[36] P.J. Phillips, H. Wechsler, J. Huang and P.J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image Vis. Comput.*, vol. 16, no. 5, pp. 295-306, 1998, DOI: 10.1016/s0262-8856(97)00070-x.

[37] V. Jain and E. Learned-Miller, "FDDB: A Benchmark for Face Detection in Unconstrained Settings," Technical Report UM-CS-2010-009, Dept. of Computer Science, University of Massachusetts, Amherst, online available at http://vis-www.cs.umass.edu/fddb/, 2010.

[38] H. Schneiderman and T. Kanade, "CMU Profile Face Images " available at http://vasc.ri.cmu.edu//idb/html/face/profile_images/index.html, 2011.

[39] G.B. Huang, M. Ramesh , M. Berg and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," Technical Report 07-49, University of Massachusetts, Amherst, online available at http://vis-www.cs.umass.edu/lfw/, Oct. 2007.

[40] D.J.C. Mackay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.

[41] H.M. Wallach, "Evaluation Metrics for Hard Classifiers," Technical Report, Cavendish Lab., Univ. Cambridge, Cambridge, online available at http://www.inference.phy.cam.ac.uk/hmw26/, Sep. 2006.

[42] M. Toews and T. Arbel, "A statistical parts-based model of anatomical variability (vol 26, pg 497, 2007)," *IEEE Trans. Med. Imaging*, vol. 26, no. 5, pp. 757-757, 2007, DOI: 10.1109/tmi.2006.895907.

[43] V. Jain and E. Learned-Miller, "Online domain adaptation of a pre-trained cascade of classifiers," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* pp. 577-584, 2011.

[44] J. Li, T. Wang and Y. Zhang, "Face detection using SURF cascade," *Proc. Int. Conf. Computer Vision Workshops*, pp. 2183-2190, 2011.

[45] K. Mikolajczyk, C. Schmid and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," *Proc. European Conf. Computer Vision*, pp. 69-82, 2004.

[46] A. Vedaldi, "SIFT Implemenation in C++ (SIFT++)," available at http://www.vlfeat.org/~vedaldi/code/siftpp.html, 2011.

**Seyed Mohammad Hassan ANVAR** received the BS degree in electronic engineering and Communication systems from Shiraz University in 2002 and the MS degree in electronic engineering and Communication systems from Yazd University in 2005. He has conducted many experimental researches in the field of electronic, robotic and artificial intelligence. In August 2009, he awarded SINGA scholarship to pursue his PhD study at the Nanyang Technological University and the Institute for Infocomm Research, Singapore. His current research focuses on object recognition and face detection from different viewpoints. He is a student member of the IEEE.

**Wei-Yun YAU** (S'90–M'98–SM'05) received his BEng (Electrical) from the National University of Singapore (1992), MEng degree (1995) and PhD degree (1999) from the Nanyang Technological University. From 1997 to 2002, he was with the Centre for Signal Processing, Singapore leading the research and development effort in biometrics signal processing. His team won the top 3 positions in both speed and accuracy at the international Fingerprint Verification Competition 2000. Currently, he is a Programme Manager with the Institute for Infocomm Research, leading the research in Interactive Social Tele-Experience programme. Wei-Yun also serves as a member of the IAPR's Technical Committee on Biometrics (TC4) and an Exco member of the Asian Biometric Consortium. He is also currently the Chair of the IPTV Working Group, Singapore and the Chair of Biometrics Technical Committee, Singapore. Wei-Yun is the recipient of the TEC Innovator Award 2002, the Tan Kah Kee Young Inventors' Award 2003 (Merit), IES Prestigious Engineering Achievement Awards 2006 and the Standards Council Distinguished Award 2007. His research interest includes biometrics, active vision system, personalized media and interactive visual interface. A paper he co-published in 2008 received the Pattern Recognition Journal Honorable Mention 2010.

**Eam Khwang TEOH** received his BE and ME degrees in Electrical Engineering from the University of Auckland, New Zealand in 1980 and 1982 respectively and the PhD degree in Electrical & Computer Engineering from the University of Newcastle, New South Wales, Australia in 1986. Since June 1985, he has been with the School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore as a Lecturer, Senior Lecture and currently, an Associate Professor in the Division of Control & Instrumentation. Dr Teoh was a Vice Dean in the School of EEE, NTU from June 1996 till May 2005. In 1994 and 2012, he was voted by the students as NTU Teacher of the Year Award/Teaching Excellence Award for the School of Electrical & Electronic Engineering. Dr Teoh has published more than 240 journal and conference papers. In 2007, one of his papers titled "Lane Detection and Tracking Using B-Snake" by Wang Y., Teoh E.K. and Shen D.G. in Image and Vision Computing, Vol 22, No 4, April 2004, pp 269-280 won the Most Cited Paper Award for the journal Image and Vision Computing. His research interests include computer vision, pattern recognition, biometric recognition systems and medical image processing. He is a member of the IEEE.