Int J Comput Vis
DOI 10.1007/s11263-007-0119-z

A Viewpoint Invariant, Sparsely Registered, Patch Based, Face Verifier

Simon Lucey · Tsuhan Chen

Received: 4 March 2007 / Accepted: 17 December 2007
© Springer Science+Business Media, LLC 2008

Abstract Sparsely registering a face (i.e., locating 2–3 fiducial points) is considered a much easier task than densely registering one; especially with varying viewpoints. Unfortunately, the converse tends to be true for the task of viewpoint-invariant face verification; the more registration points one has the better the performance. In this paper we present a novel approach to viewpoint invariant face verification which we refer to as the "patch-whole" algorithm. The algorithm is able to obtain good verification performance with sparsely registered faces. Good performance is achieved by not assuming any alignment between gallery and probe view faces, but instead trying to learn the joint likelihood functions for faces of similar and dissimilar identities. Generalization is encouraged by factorizing the joint gallery and probe appearance likelihood, for each class, into an ensemble of "patch-whole" likelihoods. We make an additional contribution in this paper by reviewing existing approaches to viewpoint-invariant face verification and demonstrating how most of them fall into one of two categories; namely *viewpoint-generative* or *viewpoint-discriminative*. This categorization is instructive as it enables us to compare our "patch-whole" algorithm to other paradigms in viewpoint-invariant face verification and also gives deeper insights into why the algorithm performs so well.

Keywords Face verification · Patch-whole modeling · Viewpoint invariance

S. Lucey (✉) · T. Chen
The Robotics Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
e-mail: slucey@ieee.org

T. Chen
e-mail: tsuhan@cmu.edu

1 Introduction

Ideally, one would like to solve the problem of pose-invariant face recognition by representing faces in 3D as pose variation is inherently linear rather than non-linear in 2D. Unfortunately, there are many practical reasons why representing a face in 3D is often untenable. For example, one may be attempting to recognize a face(s) from a image/video medium that is inherently 2D (e.g., Internet, television, etc.). As a result there is an inherent need for accurate and robust view-point invariant face recognition algorithms that can perform well with a single 2D image.

All face recognition algorithms require some degree of registration so as to normalized for unwanted shape variation. In recent work Gross et al. (2004) demonstrated that improved face recognition performance can be attained using dense registration (39–54 fiducial points depending on the pose) rather than sparse registration (3 fiducial points located on the eyes and nose tip) for the task of pose-invariant face recognition (see Fig. 1). Similarly, Blanz and Vetter (2003) demonstrated good performance using extremely dense offline registration (75,972 vertex points on laser-scan 3D images) and medium density registration (at least 7–8 fiducial points depending on pose) with the online 2D images. A problem with both these approaches, however, is that automatic dense registration of the face across viewpoints remains a very difficult task; making most these algorithms still very reliant on manual registration.

Sparse registration (i.e., 2–3 fiducial points such as the eyes and nose) of the face is generally considered an easier[1] problem than dense registration. This can mainly be

[1] We should note, however, that all experiments conducted in this paper were performed with manually found fiducial points centered around the eyes.



Dense Registration Sparse Registration

Fig. 1 Examples of densely (*left image*) and sparsely (*right image*) registered face images. This paper will be concerning itself with the challenging problem of viewpoint invariant face verification with sparse registration

attributed to the nature of the sparse points being located (i.e., eyes, nose, etc.) as they typically contain strong edges and have a similar appearance across subjects. Techniques for sparse registration are more mature than their denser cousins, and can now perform very well on frontal faces (see Everingham and Zisserman 2006 for a review). Some sparse registration algorithms can now perform well across viewpoints (see Lucey and Matthews 2006 for details). In this paper we present an algorithm that is able to achieve good view-invariant face verification performance with sparse registration, making the construction of an accurate automated pose-invariant face recognition system far more feasible.

1.1 Categorizing Viewpoint Invariant Methods

Given that we are restricted to 2D appearance one can describe the task of face verification learning in terms of estimating the likelihood functions,

$$p(\mathbf{x}^g, \mathbf{x}^p|\omega), \quad \omega \in \{\mathcal{C}, \mathcal{I}\} \tag{1}$$

where ω refers to the classes where the gallery view ($\mathbf{x}^g$) and probe view ($\mathbf{x}^p$) images are similar ($\mathcal{C}$) and dissimilar ($\mathcal{I}$) in terms of subject identity. We shall refer to $\mathcal{C}$ and $\mathcal{I}$ as the client and imposter classes respectively. There is no need in this formulation for subject labels, as we assume there is only a single gallery and probe image per subject. The likelihoods in (1) are learnt offline from a finite *world set*. The world set contains a large number of subject faces representative of the population of subject faces expected during verification, but are usually independent, in terms of identity, to the subjects involved in the online verification process.

Given that the likelihood functions in (1) have been learnt, we shall now discuss how they are evaluated. Herein, we shall refer to $p(\mathbf{x}^g, \mathbf{x}^p|\omega)$ as a density function, rather than a likelihood function, as we now assume the parametric form of ω is fixed.

In this paper we shall categorize these approaches in two ways namely, *viewpoint-generative* and *viewpoint-discriminative*. This categorization shares many similarities with the *generative* and *discriminative* categorization employed in machine learning and pattern recognition (Jebara 2004; Bishop 2006) for solving classification problems. Generative classification methods attempt to model the inputs and outputs to a classification system jointly. This is often inefficient if we only need conditional distributions of output given input. Discriminant classification approaches are motivated by a minimalist approach of just learning the input-output mapping required for the task at hand. This minimalist approach can often lead to sizeable improvements in classification accuracy given a finite training set. The viewpoint-generative and viewpoint-discriminative approaches are defined as follows:

Viewpoint-Generative: The most common approach in literature (Beymer and Poggio 1995; Zhao and Chellappa 2000; Gross et al. 2004; Blanz and Vetter 2003; Blanz et al. 2005) for viewpoint invariant face verification is to find a regression function between the gallery and probe viewpoints in terms of their 2D appearance. One can then apply this regression function to *generate* what the probe image's appearance is from the gallery viewpoint. This concept can be written formally in terms of the density functions in (1) where we take the conditional expectation of $\mathbf{x}^g$ with respect to the client density function in (1),

$$\tilde{\mathbf{x}^g} = \int_{\mathbf{x}^g} \mathbf{x}^g \, p(\mathbf{x}^g|\mathbf{x}^p, \mathcal{C}) d\mathbf{x}^g \tag{2}$$

so as to gain an estimate $\tilde{\mathbf{x}^g}$ of what the claimant's probe image $\mathbf{x}^p$ looks like from the gallery view. A simple nearest neighbor classifier is then used to gain a match-score ms of how similar the client gallery image $\mathbf{x}^g$ and the claimant's estimate $\tilde{\mathbf{x}^g}$ are in terms of some distance metric. Given that we assume $p(\mathbf{x}^g, \mathbf{x}^p|\mathcal{C})$ is Gaussian, the solution to (2) can be explicitly found. As we shall discuss in Sect. 4 we can equivalently frame this problem as a least-squares regression problem (Bishop 2006).

Viewpoint-Discriminative: Recently, another approach has become popular in literature (Kanade and Yamada 2003; Liu and Chen 2005; Kim and Kittler 2005). This approach attempts to model what is discriminative between the client ($\mathcal{C}$) and imposter ($\mathcal{I}$) classes. This approach has some inherent advantages over the viewpoint-generative

approach as more emphasis is given to *discrimination*, rather than the *generation* of a gallery view image from the probe view appearance. We can express this approach formally as attempting to estimate the match-score directly as a log-likelihood ratio based on (1),

$$ms = \log P(\mathcal{C})\, p(\mathbf{x}^g, \mathbf{x}^p|\mathcal{C}) - \log P(\mathcal{I})\, p(\mathbf{x}^g, \mathbf{x}^p|\mathcal{I}) \qquad (3)$$

where $P(\mathcal{C})$ and $P(\mathcal{I})$ are the priors for the client and imposter distributions respectively. In an ideal world, this would be the optimal approach for performing viewpoint face verification as it would realize the optimal Bayesian decision boundary between clients and imposters. In practice, unfortunately, such a strategy is too naive as one typically has no idea of the true likelihood functions $p(\mathbf{x}^g, \mathbf{x}^p|\omega)$ or even their parametric form. As we shall discuss in Sect. 5, viewpoint-discriminative methods vary based on their simplifying assumptions to realize a reasonable approximation to the decision boundary seen in (3).

Although similar, the viewpoint-generative and viewpoint-discriminative approaches differ to machine learning's generative and discriminative approaches. Specifically, the viewpoint-generative approach attempts to model jointly the gallery and probe view appearance for the same subject. With this model one can generate a missing appearance viewpoint from a single gallery or probe image. This modeling approach, however, is not concerned with generating a match-score. Instead, a nearest neighbor classifier is used to obtain the match-score. Both the generative and discriminative approaches to machine learning are concerned with generating a match-score from the modeling procedure. From this perspective both generative and discriminative machine learning approaches can be employed within the viewpoint-discriminative approach.

1.2 Contributions

In this paper we review and analyze *viewpoint-generative* and *viewpoint-discriminative* approaches to face verification. Our paper is broken down as follows. In Sect. 4 we review common approaches for viewpoint-generative face verification. We make a contribution by demonstrating how these various approaches are really variations on the same technique and demonstrate empirically which variation performs best. In Sect. 5 we describe in detail viewpoint-discriminative methods, reviewing and analyzing two approaches for face verification, namely the well-known viewpoint-differential (Moghaddam and Pentland 1997; Kanade and Yamada 2003; Kim and Kittler 2005) method as well as a naive approach we refer to as the viewpoint-joint method. We discuss their advantages and disadvantages, hypothesizing how particular elements of each approach could be combined to make a more effective algorithm. In Sect. 6 we then review and analyze existing patch-based variants of these approaches. Based on this analysis we propose a novel approach we refer to as the "patch-whole" method (see Fig. 2). Our method exhibits superior performance in evaluations when compared to existing approaches in literature.

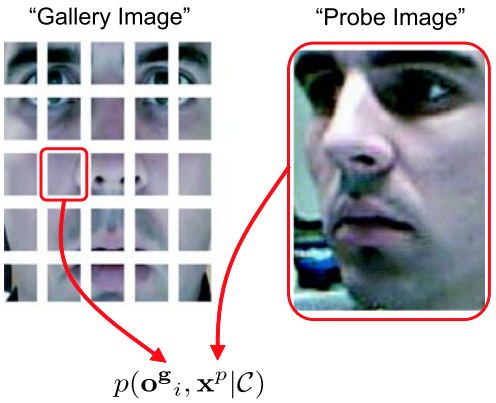


Fig. 2 In this paper we demonstrate that good performance, which is robust to pose mismatch, can be obtained by modeling the marginal distribution of gallery patch appearance $\mathbf{o}^g$ at position i with the whole appearance of the probe image $\mathbf{x}^p$ (note we employ the notation $\mathbf{x}$ for representing the whole facial appearance, and $\mathbf{o}$ for representing patch appearance). We refer to this approach as our "patch-whole" method

Compared to our previous work (Lucey and Chen 2006), this paper performs a more exhaustive evaluation and gives additional insights into what component of our algorithm is leading to improved performance. The patch-whole algorithm that we present in this paper is quite different to the one presented in (Lucey and Chen 2006). Specifically, we introduce the use of a regularization term, within our patch-whole framework, to encourage generalization and abandon the use of heuristically chosen feature compaction techniques (like the discrete cosine transform (DCT)) used previously. We also propose and analyze extensions to our patch-whole approach such as: (i) balancing the energy between patch-whole pairs, (ii) employing a symmetrical match-score and (iii) combining multiple patch-size models.

2 Related Work

Blanz et al. (2005) categorized viewpoint-invariant face recognition algorithms into two alternate paradigms; namely *viewpoint-transformed* and *coefficient-based*. Viewpoint-transformed approaches essentially act in a pre-processing manner to transform/warp the probe image, based on estimated pose parameters, to match the gallery image in pose. Coefficient-based recognition attempts to estimate the light-field (Gross et al. 2004) of the face (i.e. the face under all viewpoints, or at least the face under the gallery and probe

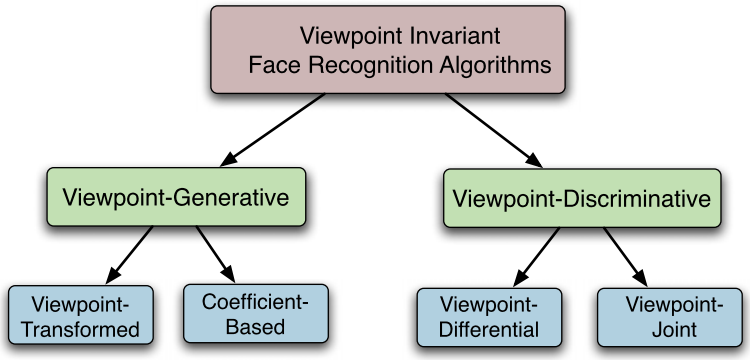


Fig. 3 This figure depicts our proposed taxonomy of paradigms within viewpoint-invariant face recognition. One can see that the viewpoint-transformed and coefficient-based paradigms of (Blanz et al. 2005) are subsumed under the viewpoint-generative paradigm; this shall be discussed in more detail in Sect. 4. Viewpoint-differential techniques like those of Moghaddam and Pentland (1997), Kanade and Yamada (2003) or (Kim and Kittler 2005) are categorized under the viewpoint-discriminative paradigm. We shall also propose a naive method in Sect. 5 which we refer to as the viewpoint-joint approach. This approach shall form the central framework of our proposed "patch-whole" method

viewpoints) based on a single image; this is done for both the gallery and probe image. Notable examples of viewpoint-transformed recognition can be seen in the work of Beymer and Poggio (1995) as well as Zhao and Chellappa (2000). The work of Sanderson et al. (2006) can also be loosely categorized as a viewpoint-transformed approach, with the transform/warp being applied to the histogram of texture extracted from the face image rather than the face image itself. Examples of coefficient-based recognition can be see in the work of Gross et al. (2004) and Blanz et al. (2005).[2] Figure 3 depicts a graphical relation between viewpoint-transformed and coefficient-based paradigms. As we shall see in Sect. 4, both the viewpoint-transformed and coefficient-based paradigms can be thought to be variations of the viewpoint-generative paradigm we proposed in Sect. 1.1.

Although useful, the initial categorization of Blanz et al. does not satisfactorily describe all paradigms in viewpoint-invariant face recognition literature. The *viewpoint-differential* paradigm, as we refer to it, attempts to model the difference of gallery and probe images for clients and imposters. This paradigm places more emphasis on learning what is important for good recognition across viewpoints, rather than good reconstruction of the face/lightfield. The work of Kanade and Yamada (2003), and Kim and Kittler (2005) are notable examples of the viewpoint-differential paradigm. Figure 3 depicts how this paradigm relates to the viewpoint-transformed and coefficient-based paradigms proposed by Blanz et al., and the viewpoint-generative and viewpoint-discriminative paradigms we proposed in Sect. 1.1.

[2]We should note that in the work of both Gross et al. (2004) and Blanz et al. (2005) that the coefficient-based approach was applied to both the texture *and* 3D shape components of the face. Since we are working with sparsely registered faces in this paper, we shall be comparing techniques that solely transform the texture component of the face.

3 Evaluation and Database

Verification is performed by accepting a claimant when his/her match-score is greater than or equal to *Th* and rejecting him/her when the match-score is less than *Th*, where *Th* is a given threshold. Verification performance is evaluated using two measures; being false rejection rate (FRR), where a true client is rejected against their own claim, and false acceptance rate (FAR), where an impostor is accepted as the falsely claimed client. The FAR and FRR measures increase or decrease in contrast to each other based on the threshold *Th*. The overall verification performance of a system is typically visualized in terms of a receiver operating characteristic (ROC) or detection error tradeoff (DET) curve. A simple measure for overall performance of a verification system is found by determining the equal error rate (EER) for the system, where FAR = FRR.

Experiments were performed on a subset of the FERET database (Phillips et al. 2000), specifically images stemming from the *ba*, *bb*, *bc*, *bd*, *be*, *bf*, *bg*, *bh*, and *bi* subsets; which approximately refer to rotation's about the vertical axis of 0°, +60°, +40°, +25°, +15°, −15°, −25°, −40°, −60° respectively. In all experiments, gallery images stem from the frontal pose *ba* with probe images stemming from all other view-points. The database contains 200 subjects which were randomly divided into sets $g1$ and $g2$ both containing 100 subjects. The world set is used to learn any non-client data-dependent aspects of the verification system. The evaluation set is used to obtain performance rates for the verification system. The $g1$ and $g2$ sets were used interchangeably as the world and evaluation sets. All images were geometrically normalized, using a similarity warp based on the eye coordinates, to give a cropped face image of 74×64 pixels. It was found during our experiments that employing an

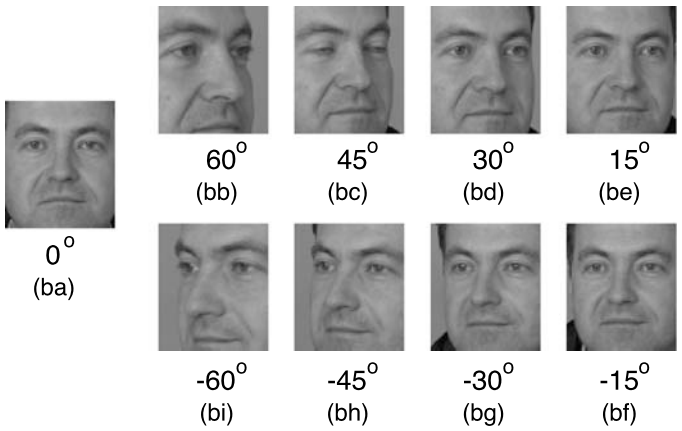


Fig. 4 Example of FERET images used in experiments with poses stemming from the *ba*, *bb*, *bc*, *bd*, *be*, *bf*, *bg*, *bh*, and *bi* subsets; which approximately refer to rotation's about the vertical axis of 0°, +60°, +40°, +25°, +15°, −15°, −25°, −40°, −60° respectively

affine warp, based on eye and nose coordinates, led to poorer performance. All images employed were in grayscale.

4 Viewpoint Generative Methods

4.1 Viewpoint-Transformed Methods

Given that we have a sparse registration of the face, a common approach (Blanz et al. 2005; Blanz 2006) in literature has been to find the regression/transformation matrix $\mathbf{W}$ between offline probe $\mathbf{X}^p$ and gallery $\mathbf{X}^g$ view examples. One can solve for $\mathbf{W}$ typically by minimizing,

$$\text{tr}[(\mathbf{X}^g - \mathbf{W}\mathbf{X}^p)^T(\mathbf{X}^g - \mathbf{W}\mathbf{X}^p)] + \alpha \cdot \text{tr}[\mathbf{W}^T\mathbf{W}] \tag{4}$$

where α is a regularization factor that is employed so as to avoid over fitting. In this form we assume that a unit bias has been appended to the probe example matrix $\mathbf{X}^p = [[\mathbf{x}^p{}_1^T, 1]^T, \ldots, [\mathbf{x}^p{}_N^T, 1]^T]$ containing N examples; no such bias is applied to the gallery example matrix $\mathbf{X}^g = [\mathbf{x}^g{}_1, \ldots, \mathbf{x}^g{}_N]$. Note that each column vector in $\mathbf{X}^g$ and $\mathbf{X}^p$ corresponds to each other in terms of subject identity. The solution to $\mathbf{W}$ is simply,

$$\mathbf{W} = \mathbf{X}^g\mathbf{X}^{pT}(\mathbf{X}^p\mathbf{X}^{pT} + \alpha\mathbf{I})^{-1}. \tag{5}$$

Typically the form of $\mathbf{x}^g$ and $\mathbf{x}^p$, which are the column vectors making up $\mathbf{X}^g$ and $\mathbf{X}^p$ respectively, have been vectorized images of the face. The regularization factor α is estimated through a cross-validation procedure.[3]

When one wants to match online a client's gallery image $\mathbf{x}^g$ with a claimant probe image $\mathbf{x}^p$, one performs two steps:

1. Synthesize the gallery view from the probe image $\mathbf{x}^p$:

$$\tilde{\mathbf{x}^g} = \mathbf{W}[\mathbf{x}^{pT}, 1]^T. \tag{6}$$

2. Then measure the distance between the synthesized gallery image $\tilde{\mathbf{x}^g}$ and the true gallery image $\mathbf{x}^g$:

$$ms = d(\mathbf{x}^g, \tilde{\mathbf{x}^g}) \tag{7}$$

where ms is the match-scored used for verification. For the purposes of this paper we shall be using a Euclidean distance.

We should note that the regression matrix can be expressed in terms of the offline probe images in the world set $\mathbf{W} = \mathbf{A}\mathbf{X}^{pT}$. One can then replace all dot products in (4) with kernel operations $k(\mathbf{x}, \mathbf{y})$ and then attempt to solve for $\mathbf{A}$ (see Bishop 2006) in a non-linear space. For the purposes of this paper, however, we shall restrict ourselves to only the linear case.

4.2 Coefficient-Based Methods

Given a sparse registration of the face, another common approach (Blanz et al. 2005; Gross et al. 2004; Blanz 2006) is to perform coefficient based pose-invariant face recognition. Typically, this approach performs PCA[4] on the offline probe view examples $\mathbf{X}^p$ and gallery view examples $\mathbf{X}^g$ where the column vectors in each matrix correspond to the same subject and there is only a single sample of each subject per view. We then obtain a compressed coefficient representation of both views,

$$\mathbf{C} = \mathbf{V}[\mathbf{X}^{gT}, \mathbf{X}^{pT}]^T \tag{8}$$

where $\mathbf{V}$ is the ensemble of eigenvectors from the PCA process and $\mathbf{C}$ is the ensemble of compact coefficients cor-

[3] Whenever cross-validation is mentioned in our work concerning a regularization factor we employed a five-fold cross-validation procedure where we randomly divide our face dataset 5 times into world, evaluation and imposter sets. The optimal regularization factor is then exhaustively found for each fold, with the final value being chosen as the average value across the five folds.

[4] Throughout this paper whenever principal component analysis (PCA) is employed the first n eigenvectors are preserved corresponding to 99% of the energy in that dataset.

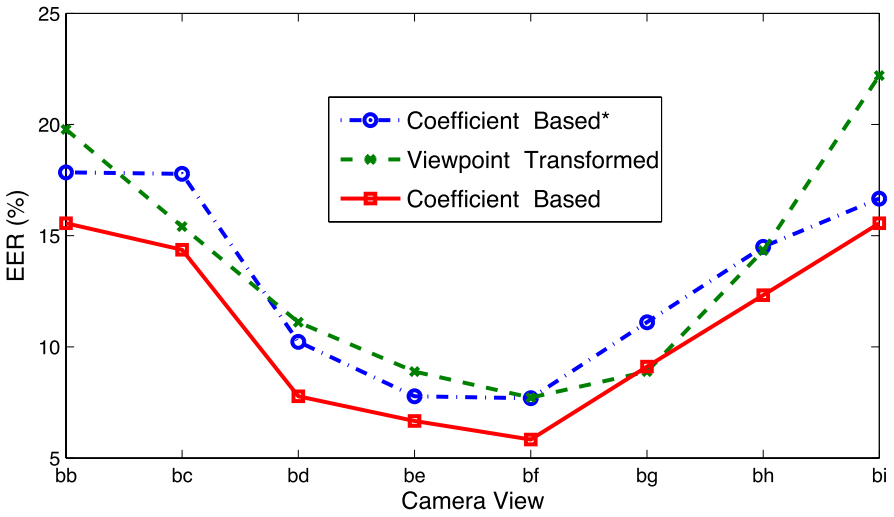


Fig. 5 Results demonstrate that the coefficient-based method, given suitably chosen regularization factors, outperforms the viewpoint-transform method for the task of face verification. Note a coefficient-based method with *no* regularization is included (denoted by the *) for completeness. Results were evaluated on set $g1$, using set $g2$ as the world set

responding to the subject identities in the columns of $\mathbf{X}^g$ and $\mathbf{X}^p$. Note, the offline gallery and probe means have been subtracted from the columns of $\mathbf{X}^g$ and $\mathbf{X}^p$.

When one wants to match online a client's gallery image $\mathbf{x}^g$ with a claimant probe image $\mathbf{x}^p$, one typically performs three steps:

1. Estimate the joint compressed coefficient $\mathbf{c}^p$ for the claimant's probe image,

$$\mathbf{c}^p = \mathbf{D}(\mathbf{D} + \beta\mathbf{I})^{-1}\mathbf{V}[\tilde{\mathbf{x}^g}^T, \mathbf{x}^{pT}]^T \tag{9}$$

   where $\mathbf{D}$ is the diagonal eigenvalue matrix corresponding to the eigenvectors in $\mathbf{V}$, and β is a regularization factor. Note that $\tilde{\mathbf{x}^g}$ is estimated from $\mathbf{x}^p$ through the view-point transformed method in (6). Both $\tilde{\mathbf{x}^g}$ and $\mathbf{x}^p$ have had their offline means removed before applying (9). The regularization factor β is estimated through a cross-validation procedure.
2. Estimate the joint compressed coefficient $\mathbf{c}^g$ for the client's gallery image by applying (9) again, except using the real $\mathbf{x}^g$ and the estimated $\tilde{\mathbf{x}^p}$ via (6).
3. Then measure the distance between the client's gallery coefficient $\mathbf{c}^g$ and the claimant's probe coefficient $\mathbf{c}^p$:

$$ms = d(\mathbf{c}^g, \mathbf{c}^p) \tag{10}$$

   where ms is the match-scored used for verification. For the purposes of this paper we shall be using a Euclidean distance.

The inclusion of the regularization factor β in (9) can be understood if we assume that all appearance vectors $\mathbf{x}$ contain Gaussian[5] noise such that,

$$\mathbf{x} \sim \mathcal{N}(\mathbf{Vc}, \beta\mathcal{I}) \tag{11}$$

where $\mathbf{c}$ is the compact appearance vector and $\mathbf{V}$ is the mixing matrix (typically a matrix of eigenvectors estimated through a PCA process). The inclusion of this Gaussian noise is useful for generalization as it is tantamount to synthetically generating hundreds of training examples. The maximum *a posteriori* (MAP) solution to $\mathbf{c}$ given $\mathbf{x}$, $\mathbf{V}$ and β is given in (9). For a full derivation of (9) please refer to (Bishop 2006; Blanz 2006).

Interestingly, there is a link between the coefficient-based method just described and the viewpoint-transformed method described in Sect. 4.1. Specifically that the match-score realized in (10) can be expressed as the summation of the two asymmetric match-scores,

$$ms = \underbrace{d(\mathbf{x}^g, \tilde{\mathbf{x}^g})}_{\text{profile}\Rightarrow\text{gallery}} + \underbrace{d(\tilde{\mathbf{x}^p}, \mathbf{x}^p)}_{\text{gallery}\Rightarrow\text{probe}} \tag{12}$$

based on (6) and (7) assuming $\mathbf{V} = \mathcal{I}$ and $\beta = 0$. In these circumstances the final match-score is just the sum of the match-score realized from applying the viewpoint-transformed method in both directions (i.e. $\mathbf{x}^g \Rightarrow \tilde{\mathbf{x}^p}$ and $\mathbf{x}^p \Rightarrow \tilde{\mathbf{x}^g}$). If the viewpoint-transformed approach was perfect there would be no need to obtain match-scores in both directions. However, since the process is lossy and introduces noise there may be an advantage in taking the average across both directions. The application of an appropriately chosen $\mathbf{V}$ and β discourages overfitting.

4.3 Experiments

To emphasize the importance of regularization in viewpoint-generative methods, and the advantage of coefficient-based methods over viewpoint-transformed we present verification results in Fig. 5. One can see in these results that employing suitably chosen regularization factors, and employing a coefficient-based method over viewpoint-transformed

[5]Note, we will be using $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote a multi-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

leads to the best results. It is interesting to note that these results are in direct contradiction to the results seen by Blanz et al. (2005) in their comparison of viewpoint-transformed and coefficient based methods. One reason for this difference could stem from how Blanz et al. formulated their viewpoint-transformed and coefficient-based methods. Specifically, Blanz et al. used pre-existing state of the art frontal face recognizers with their viewpoint-transformed method, and only a simple nearest neighbor classifier with the coefficient-based method. One could argue that this introduced a major bias towards the viewpoint-transformed method. Additionally, the dataset they conducted their evaluation on contained less viewpoint variation than our dataset ($\pm 45^\circ$ compared with $\pm 60^\circ$), which could also be a factor as the major difference in performance between methods can be seen at the more extreme viewpoints. Finally, Blanz's method involved mappings of shape and texture components of the face whereas our approach is only dealing with texture.

5 Viewpoint-Discriminative Methods

Viewpoint-discriminative methods can be described as methods that attempt to model the decision boundary between client and imposter classes. In this section we will look at a variety of strategies for encouraging this generalization.

5.1 Viewpoint-Joint Methods

As discussed in the introduction, in theory it would be optimal to use the raw holistic appearance vectors $\mathbf{x}^g$ and $\mathbf{x}^p$ to estimate the actual joint likelihood functions in (1) and then obtain match-scores through the application of (3). In practice, however, this approach leads to very poor performance due to: a) the unknown parametric form of the joint density functions, and b) the finite nature of the offline world set. Fortunately, the bias seen in the density functions towards the offline world set can be alleviated somewhat through the employment of a regularization factor during estimation. This results in the following approximation,[6]

$$p(\mathbf{x}^g, \mathbf{x}^p|\omega) \rightsquigarrow p(\mathbf{c}^g, \mathbf{c}^p|\omega) \tag{13}$$

where,

$$\mathbf{c}^q = \mathbf{D}^q(\mathbf{D}^q + \beta\mathbf{I})^{-1}\mathbf{V}^q\mathbf{x}^q \tag{14}$$

[6]It should be emphasized that we are attempting to approximate the output of the density function for the purposes of classification, not the generative distribution itself. To make this difference clear, we use the $\rightsquigarrow$ to denote our approximation.

given that $q \in \{g, p\}$, $\mathbf{V}^q$ is the matrix of eigenvectors and $\mathbf{D}^q$ is the diagonal matrix of corresponding eigenvalues for offline world set examples stemming from view q. In (14) we assume the offline mean for view q has been subtracted from $\mathbf{x}_q$. One can then apply (3) to obtain a match-score for verification. This approach while giving reasonable results is very sensitive to the correct selection of β. As per our previous approaches, β is selected through a cross-validation procedure.

5.2 Viewpoint-Differential Methods

A number of approaches have been employed in literature in order to estimate the 2D appearance likelihoods in (1). One of the most well known has been the intra-personal (i.e. client) and extra-personal (i.e. imposter) approach of Moghaddam and Pentland (1997). In this approach the authors attempt to model the *differential* appearance between gallery and probe images $\mathbf{x}^g$ and $\mathbf{x}^p$, in order to make the approximation,

$$p(\mathbf{x}^g, \mathbf{x}^p|\omega) \rightsquigarrow p(\mathbf{x}^g - \mathbf{x}^p|\omega) \tag{15}$$

from the offline examples present in the world set. These likelihoods are attempting to model the holistic face appearance for both the client ($\omega = \mathcal{C}$) and imposter classes ($\omega = \mathcal{I}$). As pointed out by Moghaddam and Pentland, there is an inherent advantage in modeling the differential appearance, rather than joint appearance, of the client and imposter classes as the differencing step reduces the variation of the pattern being modeled.

It has been reported (Moghaddam and Pentland 1997) that techniques centered around linear discriminant analysis (LDA), like those seen in the Fisherface (Belhumeur et al. 1997) algorithm, can obtain similar performance to Moghaddam and Pentland's approach. LDA based approaches employ a similar paradigm to the approach of Moghaddam and Pentland, in terms of differential appearance, although they are not framed within a strict probabilistic framework. Approaches centered around variants of LDA, have recently reported good performance on the problem of pose mismatched face recognition (Kim and Kittler 2005).

For the experiments in this paper, we will be assuming the client and imposter classes of the differential appearance likelihood function in (15) are modeled through a normal distribution. The employment of other distributions is possible. However, with the number of training observations being far less than the dimensionality of the training data, a normal distribution becomes a natural choice as it can be completely described in terms of its first and second order moments. Further, these normal distributions are estimated

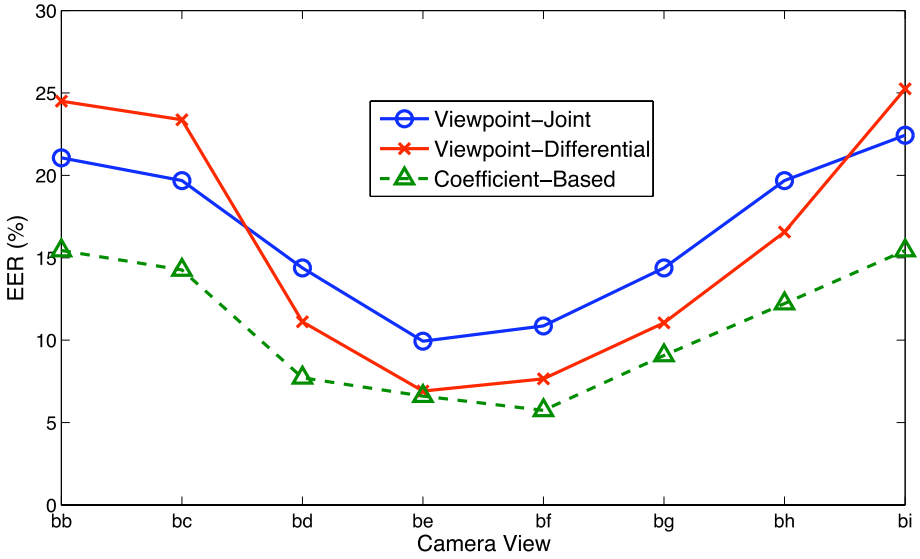


Fig. 6 Comparison of results between the viewpoint-joint, viewpoint-differential and coefficient-based methods. Results demonstrate that there is benefit in modeling the whole joint appearance (i.e. the viewpoint-joint and coefficient-based methods), especially in the presence of a large viewpoint mismatch (i.e. greater than 40°). Differential approaches still receive good performance in the presence of small viewpoint mismatch. Results were evaluated on set $g1$, using set $g2$ as the world set

within a subspace, found using PCA, that preserves all major modes of extra-personal variation. Constraining the distribution to lie within this subspace ensures that the covariance matrix, describing the client and imposter classes, is not rank deficient. A match-score is then obtained through a log-likelihood ratio as described in (3).

5.3 Experiments

In Fig. 6 one can see a performance breakdown of algorithms representing the three methods discussed thus far, namely (i) coefficient-based, (ii) viewpoint-joint and (iii) viewpoint-differential. The viewpoint-transformed method was omitted from this analysis as the approach is just a variant of the coefficient-based method. The coefficient-based method obtains the best performance overall, in comparison to the viewpoint-joint and viewpoint-differential algorithms.

In Fig. 7, we conducted an additional experiment where we tested the performance of all three algorithms for the situation where the gallery image is "badly" misaligned with the probe image. We synthetically created this misalignment by performing a 180° circular shift on the gallery image in the x and y directions. An example of this synthetic misalignment can be seen at the bottom of Fig. 7. It must be emphasized, for these experiments, that the circular shift operation was applied to both the offline and online gallery images; requiring the likelihood functions for all three algorithms to be re-estimated. Interestingly, there was no noticeable degradation in performance for the coefficient-based and viewpoint-joint algorithms, whereas the viewpoint-differential algorithm suffered catastrophic degradation in comparison to the original results seen in Fig. 6.

An immediate observation one can make about the experimental results in Figs. 6 and 7 is that any algorithm that relies on modeling differential appearance intrinsically relies on "some level" of alignment between the gallery and probe images. As pose mismatch increases, the alignment of gallery and probe images tends to degrade; resulting in poorer verification performance. We should also point out that the coefficient-based approach significantly outperforms the viewpoint-joint approach in Fig. 6. This poor performance demonstrates some of the intrinsic problems in attempting to model the raw joint likelihood functions for the client and imposter classes in the viewpoint-joint approach. Conversely, the coefficient based method employs a naive but well performing nearest neighbor classifier requiring no learning. The only learning associated with the coefficient based method stems from the regression matrices required for mapping between gallery and probe views and vice-versa.

6 Patch Variants

6.1 Differential-Patch

Recently, Kanade and Yamada (2003) proposed an effective extension to the holistic viewpoint-differential approach of Moghaddam and Pentland. This extension is centered around the decomposition of a face image into an ensemble of sub-image patches $\mathbf{x} = \{\mathbf{o}\}_{r=1}^{R}$. An example of this decomposition can be seen in Fig. 8. This decomposition was motivated by hypothesized deficiencies in holistic appearance-based template matching. In holistic template matching, if we use the whole face region for comparison, it is not easy to take into account changes in appearance due to pose differences, because the appearance in a different part of a face changes in a different manner due to its complicated three-dimensional shape (e.g. the nose). By treat-

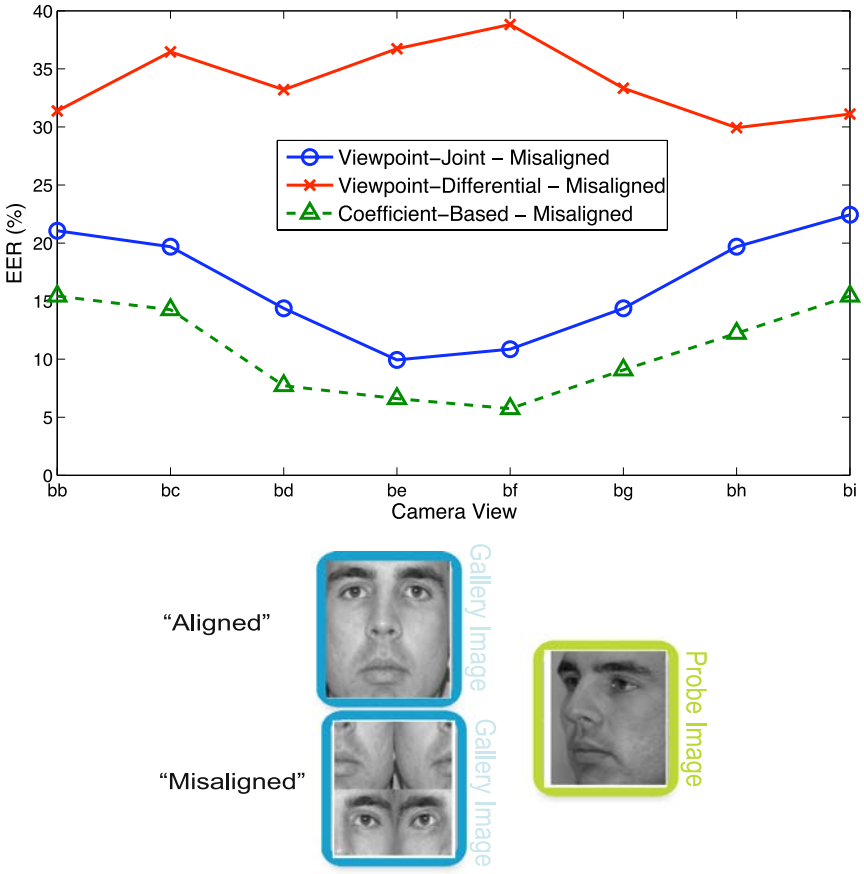


Fig. 7 Demonstration of how algorithms that model joint appearance, such as the coefficient-based and viewpoint-joint methods, are less prone to the effects of bad alignment between gallery and probe images. All results in this figure were carried out on misaligned gallery images. Refer to Fig. 6 for the aligned performance of the algorithms. Note that there is minimal difference in performance between the aligned and misaligned experiments for the algorithms that model joint appearance. However, there is a catastrophic drop in performance for the viewpoint-differential algorithm for the misaligned experiment

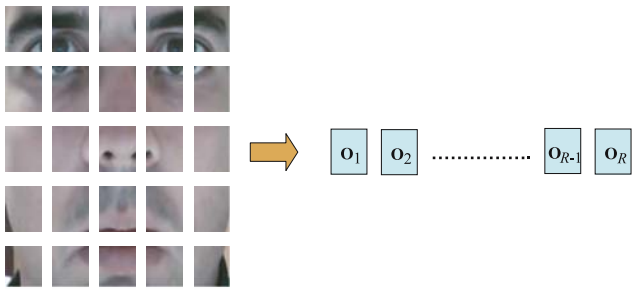
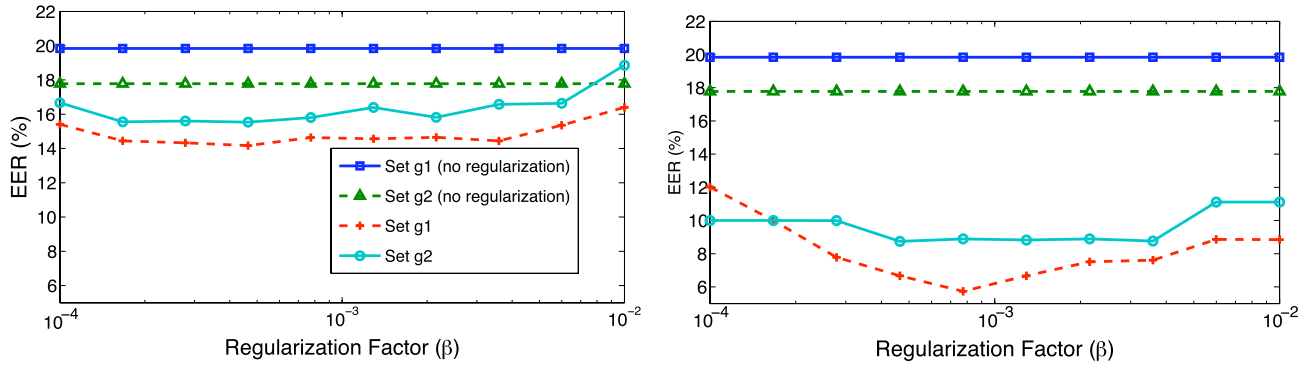


Fig. 8 In this paper we will be using a patch-based representation of the face such that $\mathbf{x} = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_{R-1}, \mathbf{o}_R]$

ing the face as an ensemble of independent patches we can, to some extent, circumvent this problem by learning how the discrimination of each local region of the face varies as a function of pose. Kanade and Yamada (2003) proposed gaining distributions based on the "sum of squared differences" (SSD), however in a recent evaluation Lucey and Chen (2006) we demonstrated that better performance can be obtained by making the following approximation based on the actual patch values,

$$p(\mathbf{x}^g, \mathbf{x}^p|\omega) \rightsquigarrow \prod_{i=1}^{R} p(\mathbf{o}^g{}_i - \mathbf{o}^p{}_i|\boldsymbol{\lambda}_{\omega_i}). \tag{16}$$

The parametric form of $\boldsymbol{\lambda}$ is assumed to be a multidimensional Gaussian distribution. A 2D discrete cosine transform was used to preserve the 32 most energy preserving dimensions in each patch. This dimensionality reduction was performed so as to ensure the covariance matrices are well ranked. We should note that an extension to Kanade and Yamada's work was proposed by Liu and Chen (2005) for the case where an ellipsoid can be additionally registered to the gallery and probe faces. Since we are restricting our work to sparsely registered faces this extension is outside the scope of this paper.



(a) Viewpoint *bb* (+60°) (b) Viewpoint *be* (+15°)

Fig. 9 This figure depicts verification performance as a function of a regularization factor β (see (4)) for two pose mismatches, specifically: **a** viewpoint *bb* (+60°) and, **b** viewpoint *be* (+15°). For completeness we have also included verification performance for when $\beta = 0$ (i.e. no regularization). One can see that employing a non-zero β increases performance for both pose mismatches and across both evaluation sets $g1$ and $g2$. Interestingly, the advantage of the regularization factor β seems to be much greater for smaller (i.e. viewpoint *be*) rather than larger (i.e. viewpoint *bb*) pose mismatches

6.2 Patch-Whole Methods

Although giving good performance, it was demonstrated in Sect. 5 that viewpoint-differential methods suffer an inherent drawback. Specifically, any algorithm that relies on differential appearance, whether at the holistic or patch level, intrinsically relies on "some level" of alignment between the gallery and probe images. As pose mismatch increases, the alignment of gallery and probe images tends to degrade, resulting in poorer verification performance.

To overcome this limitation we propose to make an alternate approximation that is not reliant on differential appearance,

$$p(\mathbf{x}^g, \mathbf{x}^p|\omega) \rightsquigarrow \prod_{i=1}^{R} p(\mathbf{o}^g{}_i, \mathbf{x}^p|\boldsymbol{\lambda}_{\omega_i}) \tag{17}$$

where $\mathbf{o}^g{}_i$ refers to an image patch, at position i, within the gallery image, $\mathbf{x}^p$ refers to the whole appearance of the probe image and $\boldsymbol{\lambda}$ refers to the parametric form of the distribution (Gaussian). We refer to this approach as the *patch-whole* method. An immediate questions arises however when inspecting (17): why is there any benefit in estimating these likelihoods in a piece-wise patch fashion? We propose there are two main benefits to our patch-whole method. First, our approach enables us to employ the advantages of a patch-based representation for recognition. In a similar manner to the patch-based differential method developed by Kanade and Yamada our approach allows one to learn how the discrimination between each local region of the gallery image and the whole probe image varies as a function of pose. Second, unlike Kanade and Yamada's approach our method does *not* assume any corresponding alignment between the probe and gallery image; allowing for improved performance in the presence of large pose mismatch.

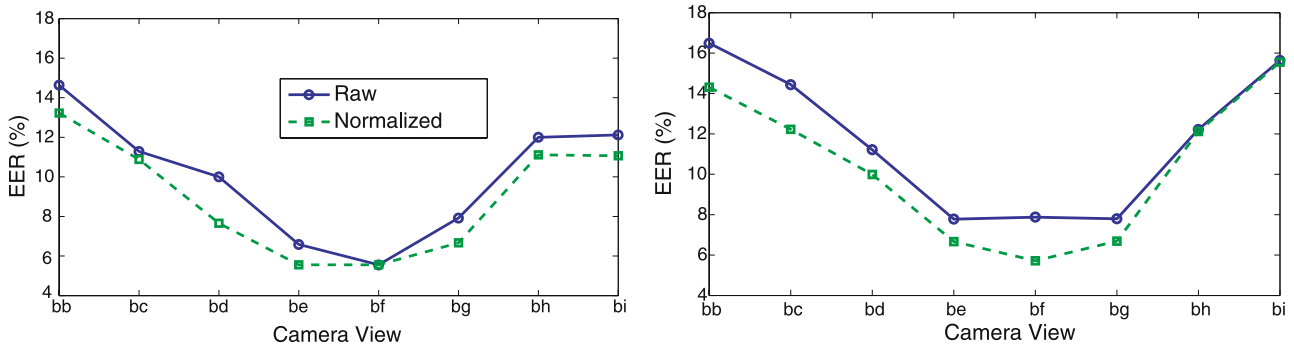
Although useful, the raw application of (17), in comparison to existing methods, still obtains poor verification performance. However, a number of steps can be taken to additionally boost the performance of our patch-whole method.

6.2.1 *Regularization*

A major problem with the raw approach in (17) stems from the finite nature of the offline world set used to estimate the likelihood functions. Specifically, the likelihood functions in (17) are too biased towards the offline world set, rather than the online evaluation set. This problem is similar to the regularization problem seen in Sect. 5.1 for estimating the joint-holistic likelihood functions. We can lessen this bias by assuming that both $\mathbf{o}^g{}_i$ and $\mathbf{x}^p$ are affected by some Gaussian noise with isotropic variance β. We can then obtain MAP estimates of compact appearance vectors $\mathbf{c}_i^g$ and $\mathbf{c}^p$ such that,

$$p(\mathbf{o}^g{}_i, \mathbf{x}^p|\boldsymbol{\lambda}_{\omega_i}) \approx p(\mathbf{c}_i^g, \mathbf{c}^p|\boldsymbol{\lambda}_{\omega_i}) \tag{18}$$

where $\mathbf{c}_i^g$ and $\mathbf{c}^p$ are estimated from $\mathbf{o}^g{}_i$ and $\mathbf{x}^p$ respectively through the application of (14) (see Sect. 5.1) by letting $\mathbf{x}_q \in \{\mathbf{o}^g{}_i, \mathbf{x}^p\}$. Separate eigenvector $\mathbf{V}^q$ and eigenvalue $\mathbf{D}^q$ matrices are estimated for each representation. Admittedly, different regularizing factors can be used for $\mathbf{o}^g{}_i$ and $\mathbf{x}^p$, but for simplicity we chose to use the same factor β for all representations. Results for varying β can be seen in Fig. 9 for the specific viewpoints of *bb* (+60°) and *be* (+15°). For both sets $g1$ and $g2$, and both viewpoints, one can see that there is an inherent benefit in choosing a non-zero regularization factor. There is an especially large jump in per-



(a) Viewpoint *bb* (+60°)] (b) Viewpoint *be* (+15°)]

Fig. 10 This figure depicts a comparison, in terms of Equal Error Rate (EER), between *Normalized* and *Raw* gallery-patch and probe-whole compact features. One can see that their is an inherent benefit in power normalizing these compact features as demonstrated by the results at: **a** viewpoint *bb* (+60°) and, **b** viewpoint *be* (+15°). Experiments were carried out using a 16 × 16 size patch. This normalization procedure aids verification performance by balancing the energy contained in the probe-whole appearance vector with the smaller gallery-patch. Results were evaluated on set $g1$, using set $g2$ as the world set

formance for the *be* pose mismatch, giving a good indication of how especially biased the non-regularized distributions of smaller pose mismatches were to the offline world set.

6.2.2 Energy Normalization

When learning the dependencies between $\mathbf{c}_i^g$ and $\mathbf{c}^p$ there may be problems stemming from there being less energy in the compact patch $\mathbf{c}_i^g$ than the compact holistic vector $\mathbf{c}^p$ due to their differing sizes. To alleviate this problem we employed a normalization procedure. Specifically, we ensured that both compact appearance vectors $\mathbf{c}_i^g$ and $\mathbf{c}^p$ have unit norm before gathering statistics. The advantage of this strategy can be seen in Fig. 10 where we can see verification performance for *Normalized* and *Raw* compact features. One can see across both sets $g1$ and $g2$ there is an inherent benefit in power normalizing the gallery-patch and probe-whole compact features.

6.2.3 Symmetrical Match-Score

An additional refinement to our patch-whole method can be made by employing a symmetrical match-score. In (17) we denote a likelihood function where the gallery image is decomposed into patches and the probe image is treated as a whole. Let us denote the match-score from this evaluation as $\overrightarrow{ms}$. In principle, there is no reason why the operation in (17) cannot be reversed, that is the probe image is decomposed into patches and the gallery image is treated as a whole. Let us denote the match-score obtained from these reverse-likelihood functions as $\overleftarrow{ms}$.

As discussed in Sect. 4.2 for the viewpoint coefficient-based method, where there was a definite advantage in averaging match-scores stemming from warps/transformations in two directions (i.e., from gallery view to probe view and vice-versa). Borrowing upon this concept we propose that such an approach can be applied to our patch-whole method such that,

$$ms = \overrightarrow{ms} + \overleftarrow{ms} \tag{19}$$

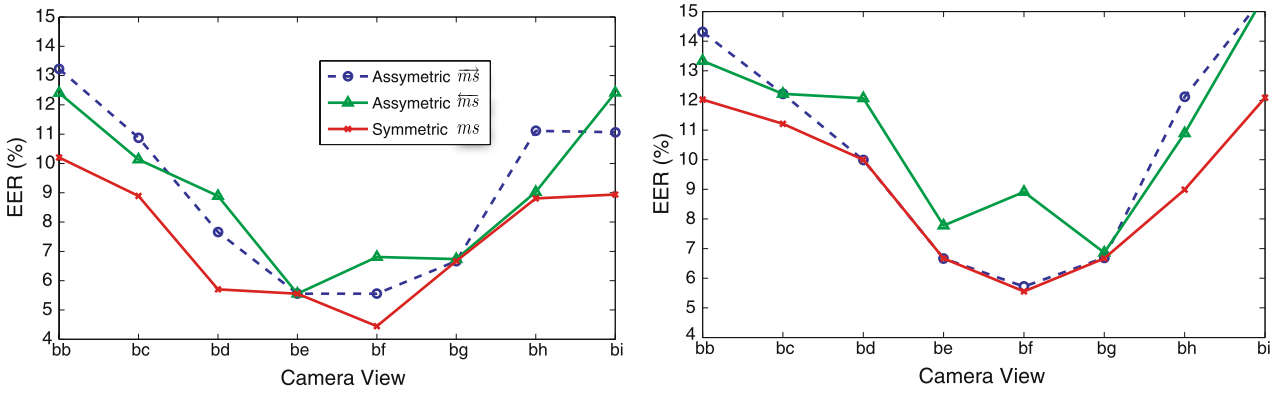
where we refer to ms as our symmetrical match-score. Results for this approach can be seen in Fig. 11 in comparison to the asymmetric match-scores $\overrightarrow{ms}$ and $\overleftarrow{ms}$. One can see there is a some advantage in employing the symmetrical over asymmetrical match-scores.

6.2.4 Patch-Size

An obvious question to ask when employing any patch-based computer vision technique is: what size patch is optimal? In Fig. 12 we give an empirical answer to this question by evaluating our patch-whole method over a variety of patch sizes. From this figure one can see there is no one optimal patch-size, although in our experiments patch-sizes of 16–20 pixels seemed to give the best overall results. In Fig. 12 we also obtained performance for when we combine the match-scores from a variety of patch sizes such that,

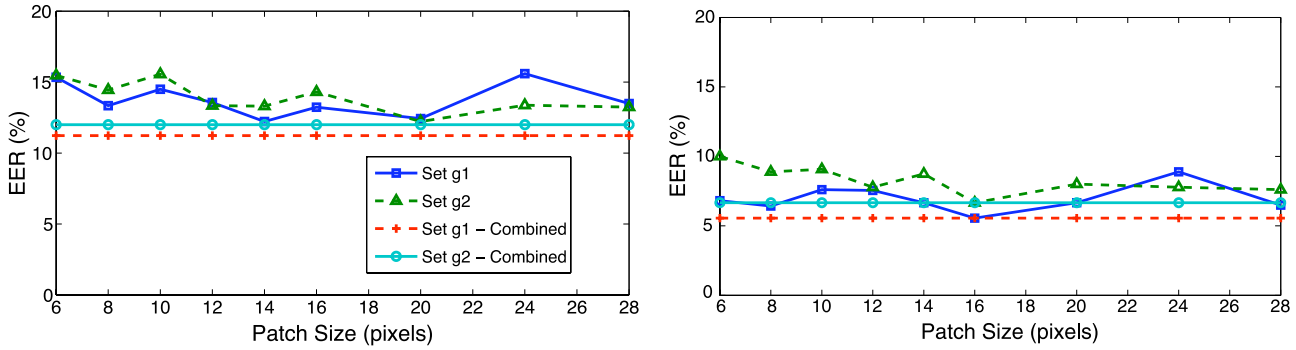
$$ms = \sum_{sz} ms^{(sz)} \tag{20}$$

where $ms^{(sz)}$ is the match-score obtained for patch size sz, resulting in the final match-score ms. This combination strategy is similar to the product rule mentioned in the classifier combination work of Kittler et al. (1998). We tested other strategies for combination such as the sum, min and max rules but found empirically the product rule to perform



(a) Set $g1$ (b) Set $g2$

Fig. 11 This figure depicts a comparison between two asymmetric match-scores and the symmetrical match-score for the patch-whole algorithm. Results across both evaluation sets indicates an advantage in employing the symmetrical match-score. All experiments were carried out using a patch size of 16×16. Results in (**a**) were derived by evaluating on set $g1$ and employing $g2$ as the world set. Results in (**b**) were obtained by swapping evaluation and world sets



(a) Viewpoint *bb* (+60°) (b) Viewpoint *be* (+15°)

Fig. 12 This figure depicts verification performance as a function of patch-size across two viewpoints, specifically: **a** *bb* (+60°) and **b** *bb* (+15°). One can see that there is an advantage in combining multiple patch-representations within the patch-whole framework

best. One can see in Fig. 12 that the combined method obtains performance equal to, and in one case superior to, individual patch size match-scores.

6.3 Comparison

For completeness we have conducted a comparison between the leading techniques mentioned in this paper and our own patch-whole method with extensions. One can see in Fig. 13 that our algorithm outperforms leading viewpoint-generative (i.e., coefficient-based method) and viewpoint-discriminative (i.e., patch-based differential method) by a substantial margin across all poses and both evaluation sets. An important thing to note from this result is that the viewpoint-discriminative paradigm is now substantially outperforming the viewpoint-generative paradigm. This result is consistent with our philosophy to viewpoint-invariant face verification, in that both the client and imposter statistics should be used to gain optimal performance. Viewpoint-generative methods suffer from an inherent drawback as they only rely on the client statistics.

7 Conclusions

In this paper we have proposed a novel approach, which we refer to as the "patch-whole" algorithm. This approach is able to deliver good face verification performance for faces that have only sparse registration. This approach has two advantages. First, it makes no assumption about the alignment between the gallery and probe image pairs; allowing it to deal with large pose mismatch. Secondly, it allows for a richer modeling of the joint appearance by decomposing the gallery image into an ensemble of statistically independent patches. Our approach outperformed all other approaches tested in our experiments. The performance of our algorithm in large pose mismatch was especially encouraging.

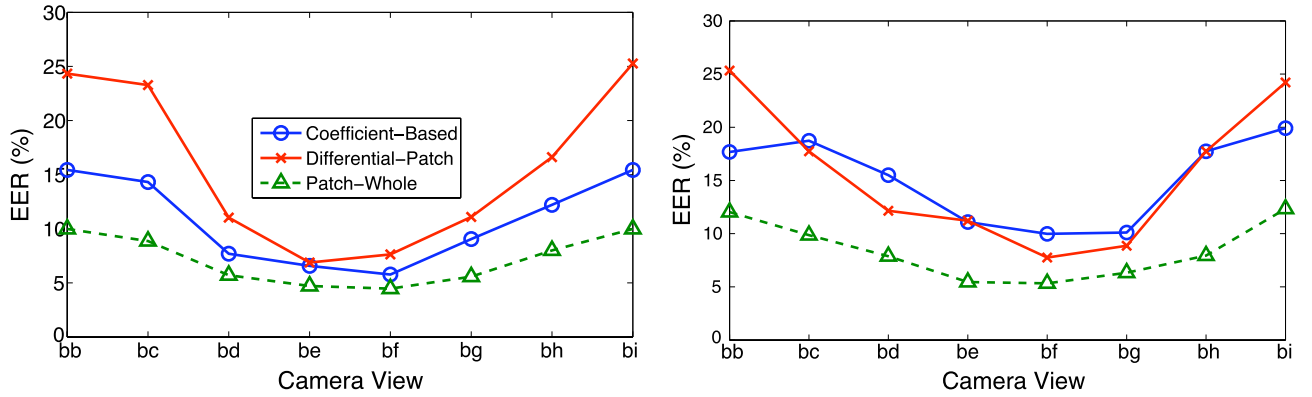


Fig. 13 This figure depicts a comparison between leading viewpoint-generative and viewpoint-discriminative methods and our own patch-whole method. Across all poses and both evaluation sets one can see that our approach substantially outperforms other leading methods. Results in (**a**) were derived by evaluating on set $g1$ and employing $g2$ as the world set. Results in (**b**) were obtained by swapping evaluation and world sets

To fairly compare our approach to what exists in literature we have also devised a taxonomy for categorizing viewpoint-invariant face recognition algorithms. Broadly, we can categorize an algorithm as being *viewpoint-generative* or *viewpoint-discriminative*. Through this categorization we make a number of additional contributions to viewpoint-invariant face recognition, namely:

- Demonstrating that the viewpoint-transformed and coefficient-based approaches of Blanz et al. (2005) are really just variants on the same approach. Empirically we demonstrated that the coefficient-based approach is slightly superior to the viewpoint-transformed approach.
- Differential methods (i.e., techniques that rely on taking the difference between gallery and probe images) have a distinct disadvantage when being employed for viewpoint-invariant face recognition. This disadvantage stems from the assumed alignment between the gallery and probe images during the differencing procedure. As a result we demonstrate empirically that methods that do not assume such a strict alignment outperform differential methods significantly.

In future work we shall be attempting to extend our "patch-whole" algorithm through the employment of discriminative rather than generative classifiers which will be able to realize more complex decision boundaries and make better use of training examples. Additionally, we shall attempt to make inroads into gaining a more sophisiticated formalism for learning the alignment between patch-whole pairs.

References

Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transaction Pattern Analysis and Machine Intelligence*, *19*(7), 711–720.

Beymer, D., & Poggio, T. (1995). Face recognition from one model view. In *International conference on computer vision*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Blanz, V. (2006). Face recognition based on a 3D morphable model. In *Automatic face and gesture recognition (FGR)* (pp. 617–624).

Blanz, V., & Vetter, T. (2003). Face recognition based on fitting a 3D morphable model. *IEEE Transaction Pattern Analysis and Machine Intelligence, 25*(9).

Blanz, V., Grother, P., Phillips, P. J., & Vetter, T. (2005). Face recognition based on frontal views generated from non-frontal images. In *International conference on computer vision and pattern recognition (CVPR)* (Vol. 2, pp. 454–461).

Everingham, M., & Zisserman, A. (2006). Regression and classification approaches to eye localization in face images. In *International conference on automatic face and gesture recognition (AFGR)* (pp. 441–446).

Gross, R., Matthews, I., & Baker, S. (2004). Appearance-based face recognition and light-fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(4), 449–465.

Jebara, T. (2004). *Machine learning discriminative and generative*. Dordrecht: Kluwer.

Kanade, T., & Yamada, A. (2003). Multi-subregion based probabilistic approach towards pose-Invariant face recognition. In *IEEE international symposium on computational intelligence in robotics automation* (Vol. 2, pp. 954–959).

Kim, T., & Kittler, J. (2005). Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(3), 318–327.

Kittler, J., Hatef, M., Duin, R., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(3), 226–239.

Liu, X., & Chen, T. (2005). Pose-robust face recognition using geometry assisted probabilistic modeling. In *International conference on computer vision and pattern recognition (CVPR)* (Vol. 1, pp. 502–509).

Lucey, S., & Chen, T. (2006). Learning patch dependencies for improved pose mismatched face verification. In *International conference on computer vision and pattern recognition (CVPR)* (Vol. 1, pp. 909–915).

Lucey, S., & Matthews, I. (2006). Face refinement through a gradient descent alignment approach. In *HCSNet workshop on the use of vision in HCI*, Canberra, Australia, 1–3 November 2006.

Moghaddam, B., & Pentland, A. (1997). Probabilistic visual learning for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(7), 696–710.

Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *10*(22), 1090–1104.

Sanderson, C., Bengio, S., & Gao, Y. (2006). On transforming statistical models for non-frontal face verification. *Pattern Recognition*, *39*(2), 288–302.

Zhao, W., & Chellappa, R. (2000). SFS based view synthesis for robust face recognition. In *International conference on automatic face and gesture recognition (AFGR)* (pp. 285–292).