Multi-View Face Recognition from Single RGBD Models of the Faces

Donghun Kim[a], Bharath Comandur[a], Henry Medeiros[b], Noha M. Elfiky[a], Avinash Kak[a,*]

[a] *School of Electrical and Computer Engineering, Purdue University, 465 Northwestern Ave, West Lafayette, IN 47907*
[b] *Department of Electrical and Computer Engineering, Marquette University, 1551 W Wisconsin Ave, Milwaukee, WI 53210*

Abstract

This work takes important steps towards solving the following problem of current interest: Assuming that each individual in a population can be modeled by a single frontal RGBD face image, is it possible to carry out face recognition for such a population using multiple 2D images captured from arbitrary viewpoints? Although the general problem as stated above is extremely challenging, it encompasses subproblems that can be addressed today. The subproblems addressed in this work relate to: (1) Generating a large set of viewpoint dependent face images from a single RGBD frontal image for each individual; (2) using hierarchical approaches based on view-partitioned subspaces to represent this training data; and (3) based on these hierarchical approaches, using a weighted voting algorithm to integrate the evidence collected from multiple images of the same face as recorded from different viewpoints. We evaluate our methods on three datasets; a dataset of 10 people that we created and two publicly available datasets which include a total of 48 people. In addition to providing important insights into the nature of this problem, our results show that we are able to successfully recognize faces with accuracies of 95% or higher, outperforming existing state-of-the-art facial recognition approaches based on deep convolutional networks.

Keywords: Face recognition, Depth cameras, Manifold representations, Multi-view face recognition, RGBD models

1. INTRODUCTION

Face recognition is now considered a reliable and nonintrusive biometric. Several algorithms that have been proposed during the last decade can now achieve accuracies that far exceed 90%. Such high levels of accuracy, however, can only be obtained for 'normalized' frontal face images. These algorithms perform less than adequately when constraints are removed on the orientation of the camera vis-à-vis the face. Although there have been many attempts at automating face image normalization to replicate such results in unconstrained scenarios, most methods that have been proposed to date are of questionable reliability. The general problem of recognizing faces under unconstrained conditions remains largely unsolved even for seemingly easy scenarios such as when there is sufficient illumination, the target motion is slow compared to the camera frame rate, and high resolution cameras are employed. A solution to this general problem would be relevant in a number of scenarios, which include face verification and identification in static imagery [1, 2, 3], video [4, 5], and camera networks [6, 7].

The problem of recognizing faces under unconstrained conditions, also known as face recognition

*Corresponding author
Email addresses: zava@purdue.edu (Donghun Kim), bcomandu@purdue.edu (Bharath Comandur), henry.medeiros@marquette.edu (Henry Medeiros), nelfiky@purdue.edu (Noha M. Elfiky), kak@purdue.edu (Avinash Kak)

in the wild, means assigning a face identity label to a set of face images collected by an assortment of cameras at random orientations with respect to a face. Imagine a human subject being tracked by the cameras at a crowded public place like an airport or a city square. This problem has become very important in recent years with the advent of camera networks. Most major cities now have surveillance cameras installed in public places. As the reader can imagine, in its most general form, it is an extremely challenging problem. When we attempt to perform face recognition in images from video and other multi-view scenarios, there is no guarantee that any of the collected images would constitute a full frontal view. In addition, we must also cope with other effects that are caused, for example, by uncontrolled illumination.

While the difficulties mentioned above can be expected to degrade the performance of any face recognition algorithm, one could raise the following question: Is it possible to compensate for some of the difficulties by leveraging the availability of multiple images recorded from different viewpoints. That is, whether multiple images from different viewpoints of the same face can compensate for lack of a single frontal image and for lack of controlled illumination. It is this question that is the focus of this paper. And, if the reader accepts the validity of the question, the problem becomes one of how one should go about pooling the visual evidence from the different viewpoints for classifying a face.

Some previous approaches attempt to solve this problem by resorting to the recent availability of large scale datasets of labeled faces in the wild [8, 9, 10, 11]. While such approaches have obtained impressive results in such scenarios, achieving accuracies as high as 99.5% and surpassing the 97.53% accuracy obtained by human observers, they suffer from two main limitations. First, they fundamentally rely on the existence of such massive datasets for training purposes. While such datasets might be readily available for celebrities and other notorious personalities, generating such datasets for a broader population might be challenging. We are more interested, therefore, in a more practical scenario in which the classifier can be trained with a single snapshot of the target. Second, even when such large datasets are available, these methods have been shown to map poorly to alternative datasets collected from a general population, which limits their practical applicability. In [11], for example, the authors have shown that when their approach based on multiple deep convolutional networks is applied to a real-world dataset of faces collected by the authors, the accuracy falls to 66% compared to the 99.5% accuracy obtained in the LFW dataset.

This paper makes a small but important step in our understanding of whether it is possible to attempt face recognition under unconstrained conditions when our training data consists of a single frontal RGBD image of each human subject. Since, as mentioned above, the general problem of unconstrained face recognition is quite broad, we focus here on this particular subproblem in order to get a better understanding of the issues involved in pooling together the visual evidence from multiple viewpoints. Within the context of our subproblem, given the RGBD images, we are faced with issues such as how to best extract viewpoint oriented 2D images from the models; how to best extract class discriminatory information from the 2D images that are likely to reside on low-dimensional manifolds in high-dimensional measurement spaces [12, 13, 14]; and, finally, how to construct a classifier that makes an identity decision based on a set of test face images collected from *random* viewpoints.

In order to solve these problems, we first create multi-view training data from single RGBD images of the human face. We then view-partition the manifolds where the data resides in order to identify the optimal subspaces in which groups of similar faces can be found. We explore two different approaches for view-partitioning the training data: pose based and appearance based.

Subsequently, we investigate how to best carry out multi-view classification by comparing the view-partitioned approaches to global approaches to multi-view classification. We study two different types of global approaches, one in which all of the training data for all human subjects is thrown into a single global subspace, and the other in which we create a person-specific global subspace separately for each human subject.

Finally, the view-partitioned approaches create the possibility of carrying out weighted voting when combining the classification labels for each of the query images into a single identity label. We do so by devising a weighting mechanism that uses the reconstruction error of a query image with respect to the different views in order to determine its weight when used in a multi-view classification mechanism.

This paper makes four main contributions. First, it presents a novel hierarchical approach for multi-view facial recognition. Second, it proposes a weighted voting scheme to improve face recognition by combining face images from different viewpoints. Third, a new dataset of RGBD face images for the evaluation of multi-view face recognition is introduced. Finally, the paper includes an extensive evaluation and analysis of several mechanisms to perform data clustering and classification for the purpose of facial recognition.

The remainder of this paper is organized as follows. Section 2 discusses some of the most relevant works related to the topic of face recognition in relatively unconstrained scenarios such as in videos and camera networks. Section 3 proposes several approaches to construct facial recognition algorithms that can be trained from a single RGBD image of each target subject, and Section 4 discusses the method we employ to combine the classification results obtained from several query images. An extensive experimental evaluation is then presented in Section 5, which is then followed by our concluding remarks in Section 6.

2. PRIOR WORK

Dealing with face recognition using non-frontal imagery at test time generally requires constructing partial or full 3D models of the human head and morphing the models in order to best describe the test images. For the case of static imagery, there are two different classes of algorithms that come under this category. In the first class, the training protocol includes generating off-normal images of the face by directly applying a pose-transform to the frontal image [15, 16, 17]. At test time, the recognizer first locates prominent facial features and then uses these locations to geometrically register the input with multiple example views. Subsequently, a correlation based operation is used to find the best match from the database. In the second approach, the goal is to use some sort of a range sensor to create a generic 3D point cloud model of either the whole head or of a set of salient points on the head [18, 19, 20, 21, 22]. Subsequently, this model, along with the accompanying texture information, can be manipulated to create off-normal training images for a human subject. At test time, a query image is generally manually annotated for the salient features of a face and the 3D model morphed to fit to the query image through these salient points.

2.1. Recognizing Faces in Videos

Recognizing a face in a video involves the following processes that may need to run simultaneously: a tracking/detection mechanism, a crucial alignment step, and a recognition algorithm, which generally attempts to exploit the availability of multiple image frames. All of these are complex and are active subjects of ongoing research [23, 24, 25, 26, 27, 28, 29, 30]. Regarding face tracking, a comprehensive survey of existing approaches is given in [31]. In a particularly relevant example, Marras et al. [32] proposed a particle filtering method that uses the reconstruction error from learned subspaces to determine facial orientation. As for face detection, although it is still a largely unsolved problem, during the past decade, much progress has been made in this area [33, 34, 35, 36]. While face detection is generally regarded as the starting point for all facial analysis tasks [37], face alignment can be regarded as an important and essential intermediate step for many subsequent higher level tasks that range from biometric recognition to mental state understanding. We discuss the issue of facial alignment in more detail below

2.1.1. Facial Alignment

The problem of face alignment is a well-studied area of computer vision that has created a wealth of scientific research. Countless applications [26, 27, 29, 38, 39], and a large number of approaches have been proposed to tackle it. Face alignment is widely used by face recognition algorithms to improve their robustness against pose variations. For example, in the

stage of face registration, the first step is usually to locate prominent facial points and use them as anchor points for affine warping, while other face recognition algorithms, such as feature-based (structural) matching [2, 40], rely on accurate face alignment to build the correspondence among local features (e.g. eyes, nose, mouth, etc.) to be matched.

Over the last two decades, numerous techniques have been developed for face alignment with varying degrees of success. Çeliktutan et al. [41] surveyed many traditional methods for face alignments of both 2D and 3D faces. More recently, Yang et al. [42] provided an empirical study of recent face alignment methods, aiming to draw some empirical yet useful conclusions and make insightful suggestions for practical applications. In general terms, face alignment can be formulated as a problem of searching over a face image for pre-defined feature points (also called face shape), which typically starts from a coarse initial shape, and proceeds by refining the shape estimate step by step until convergence. During the search process, two different sources of information are typically used: facial appearance and shape information. In particular, typically, faces are modeled as deformable objects which can vary in terms of shape and appearance. Much of the early work revolved around Active Shapes Models (ASMs) and Active Appearance Models (AAMs) [30, 43, 44]. In ASMs, facial shape is expressed as a linear combination of shape bases learned via Principal Component Analysis (PCA), while appearance is modeled locally using (most commonly) discriminatively learned templates.

AAMs, first proposed by Cootes et al. [43], are linear statistical models of both the shape and the appearance of the deformable object. They are able to generate a variety of instances by a small number of model parameters, and therefore have been widely used in many computer vision tasks, such as face recognition [45], object tracking [46], and medical image analysis [47]. Despite their popularity and success, AAMs have been traditionally criticized for the limited representational power of their holistic representation especially when used in unconstrained conditions. One possible way to overcome these drawbacks is to use part-based representations, due to the observation that local features are generally not as sensitive as global features to lighting and occlusion. ASMs are a notable example is of part-based models [48, 44], which combines the generative appearance model for each facial part and the Point Distribution Model for global shapes. More recently, the focus has shifted to a family of methods known as Constrained Local Models (CLMs) [49, 50, 51] that build upon ASM to model individual facial parts using discriminatively trained local detectors [51, 52, 53, 54]. In the training phase, a CLM learns an independent local detector for each facial point, and also a prior shape model to characterize the deformation of face shapes. For testing, face alignment is typically formulated as an optimization problem to find the best fit of the shape model to the test image.

Research in multi-view face recognition has been significantly influenced by the availability of large annotated datasets consisting of face images recorded under unconstrained conditions [55, 56, 57, 58, 59, 60]. For example, these datasets have been used to develop a variety of cascaded regression-based techniques [61, 62, 63, 64, 38, 65, 66, 67, 68] which have proved very successful in solving the face pose alignment problem. The motivation behind cascaded regression is that, since performing regression from image features to face shape in one step is extremely challenging, we can divide the regression process into stages, by learning a cascade of vectorial regressors. As in related computer vision tasks such as human pose estimation [69, 70], such methods are particularly successful when associated with generative deformable part models [61]. Despite the substantial progress made in facial alignment in recent years, it is still unclear whether the ability to determine the orientation of a face may eventually translate into more accurate facial recognition approaches for unconstrained scenarios in which face orientations may vary dramatically and frontal reconstructions are bound to be heavily distorted.

2.1.2. Exploiting Multiple Image Frames for Video-based Facial Recognition

Rather than attempting to carry out face frontalization on each frame, most video-based facial recognition approaches try to leverage the availability of

multiple images of the face, generally at a variety of poses and under different illuminations to improve the detection performance.

A well-known approach to overcome the problem of pose and illumination changes consists of recording training videos of the human subjects and, subsequently, using the frames of training videos as the gallery of images for each subject in the database. At test time, a query video recording is compared with all of the images in the database. In fact, the test consists of comparing each frame of the query video with all the images in the database and accumulating the matching scores for the query video [71, 72, 73, 74].

Instead of performing face recognition based on a frame-by-frame analysis of the training and test data, it is also possible to treat a video as a temporal stream in the three dimensional space formed by two spatial and one temporal coordinates. One can analyze this 3D space holistically to extract information that characterizes the dynamic properties of a face. Zhou et al. [75] pioneered that kind of work by tracking the subjects in the videos and extracting their faces to construct priors for the different views of the different faces. Lee et al. [5] focused on automatically learning the transition probabilities between the different possible appearances of a face in a video. Along the same lines, Liu et al. [76] used a Hidden Markov Model (HMM) for modeling the face appearance along with the head pose changes in the training videos for each human subject.

Another class of methods, the so-called ensemble approaches, focuses on the fact that a query video, when treated as a stream of temporal information, may not correspond to any of the gallery videos recorded previously for all of the human subjects [77, 78, 79, 80, 81, 82, 73]. So at test time, a frame-by-frame comparison between the query video and the gallery videos is carried out to create virtual gallery videos for each human subject using the gallery video frames that are most similar to the query video. Subsequently, face recognition is based on comparing the query video with the virtual gallery videos for the different human subjects.

2.2. Multi-Camera and Multi-View Face Recognition — Recognizing Faces in the Wild

Superficially it may seem that there should be no difference between multi-camera (or multi-view) face recognition and video face recognition. As it turns out, the two are very different because, with video, the variations in the viewpoints are bound to be localized to where the camera happens to be. On the other hand, when you have multiple cameras viewing the same subject, they could be mounted at spatially dispersed locations. One typical example are the cameras mounted at an airport that are tracking the same human subject with the goal of identifying the individual from the snippets of images recorded by the cameras. The multi-camera face recognition research can be divided into the following two categories: when a face to be recognized is in the intersection of the fields of view of all the cameras, and when that is not the case. We briefly discuss both cases below.

For the case of multi-camera face recognition when a face is in the intersection of the fields of view of all the cameras, most works focus on choosing the view that provides the most reliable evidence for recognizing a face and subsequently using a traditional approach for carrying out the recognition task [83, 84, 85]. In [85], for example, the reliability of each camera depends on how well both face detection and recognition can be carried out. Subspace learning methods provide one tool to determine the orientation of a face. Li et al. [86] proposed one of the first approaches for clustering faces into subspaces according to their poses. Their method was based on a supervised version of Independent Subspace Analysis (s-ISA). Although their experiments indicate that s-ICA provides better face pose classification than Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Topographic Independent Component Analysis (TICA), the results are largely qualitative. Kan et al. [87] proposed a method that finds optimal linear transformations to map images from different views (or different sensing modalities) into a common subspace. Their approach shows performance improvement over previous linear subspace learning approaches such as [88], but their multi-view classification evaluation is

restricted to the range [-45^o,$+45^o$]. In addition, manifold learning approaches have been shown to be more robust and have better generalization capabilities than linear methods such as ISA [89, 90, 91]. Furthermore, none of the methods mentioned above is concerned with the problem of incorporating multiple query images from different perspectives into the classification problem.

While not directly applicable to the multipose face recognition problem, another related non-linear subspace learning approach was proposed by Goudelis et al. in [92]. In that work, the authors proposed a face verification (i.e., binary recognition) method which employs a kernelized discriminant that maximizes impostor distance measures while minimizing the client (i.e., non-impostor) distance measure. Their method showed impressive single-digit equal error rates (ERR) in several challenging datasets with varying face poses.

When there is no overlap between the view-fields of the cameras involved, person re-identification becomes a fundamental issue in multi-camera face recognition [93, 94, 95, 96, 97, 98, 99, 100]. The notion of re-identification addresses the following issue related to face recognition in the wild: If a group of people is being tracked by a network of non-overlapping cameras, how can we ensure that the face fragments extracted from two different cameras belong to the same individual? Person re-identification is a complex problem on its own and is currently a subject of active research. It is, however, beyond the scope of the work reported in this paper. Obviously, after collecting the face fragments for each individual, there remains the problem of aggregating the evidence and attempting final face recognition. This aggregation problem is one of the issues this paper attempts to solve.

A few recent contributions focus on aggregating the evidence provided by the individual cameras in a multi-camera approach to face detection and recognition. In [7], An et al. aggregate the information from different cameras with the help of a dynamic Bayesian network, which contains a node for each camera and a node for each human subject the system is expected to recognize. At training time, the structure of the network and its parameters are learned with person-specific dynamics from the gallery videos. At test time, faces are recognized by maximizing the posterior probabilities derived from the camera and the human subject nodes. Du et al. [6] aggregate evidence in multi-camera scenarios by tracking a human head from camera to camera. The head model used in this work is a texture-mapped sphere represented by spherical harmonics. Recognition is carried out by comparing the head model coefficients of the training images with those that apply to a test subject. Another approach that tracks a head and then associates a pose with it was proposed by Harguess et al. [101]. At training time, all frontal images obtained from multiple cameras are used for building a generic cylinder head model and a lower dimensional subspace. At test time, the pose of the head is estimated through the cylinder head model that is constructed during the training process. This pose is used to weight the reliability of a partial view of a query face assuming that the reliability goes down as the query viewpoint moves away from the frontal view.

3. VIEW PARTITIONED SUBSPACES FOR MULTI-SUBJECT FACE DATA

As stated above, the main problem that this paper investigates is that of face recognition from a set of partial views as recorded by a spatially distributed network of cameras. In order to discriminate between the faces of different individuals, we use a vector representation for the images so that each image is regarded as a point in a high D-dimensional space. In order to cope with the curse of dimensionality, we want to create lower dimensional representations of the faces, but do so in such a way that the discriminatory information between the faces is not lost. As already alluded to in the previous sections, face data when collected from different viewpoints is likely to reside on manifolds and any dimensionality reduction approach must take into account the structure of the manifold — in both the original measurement space and in the target low-dimensional space. So the first research issue faced is how to represent the training data for the different individuals in a manifold-based

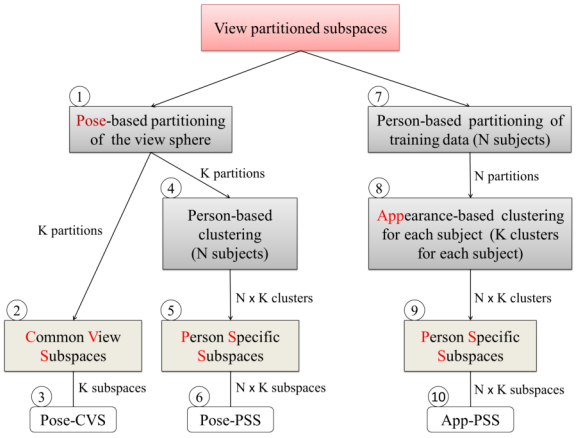

Figure 1: Variations of the view partitioned subspace model for multi-subject face images.

low-dimensional representation. We address this issue by creating multiple *view partitioned* subspaces. By view partitioning we mean simply dividing the view sphere according to some criterion.

The goal of the present section is to introduce two criteria for partitioning the training data for subspace construction. The first is based on the pose parameters associated with the training images; the second uses the appearance similarity of the images. We have previously used both of these approaches for solving the simpler problem of head pose estimation [102]. Our conclusion in that study was that, for the purpose of pose estimation, the appearance based partitioning method produced better results than the pose based partitioning method. For the purpose of face recognition, we must now also factor in the person-to-person image variations. In this context, for each partition of the training dataset, we can either construct a single subspace for all the individuals in the database, or we can create person-specific subspaces. Figure 1 illustrates these variations on the top-level pose-based and appearance-based subspace construction techniques. We will discuss each of the boxes in Figure 1 in detail later in this section.

Before focusing on the issue of how best to construct the subspaces, we are faced with the serious challenge of accumulating a large number of viewpoint variant images of the faces of different individuals for training purposes. In this paper, we have solved this problem by recording a single RGBD image for each individual in the database and then synthetically generating all the needed viewpoint variant images from the recorded RGBD image. In the next section we briefly discuss this process.

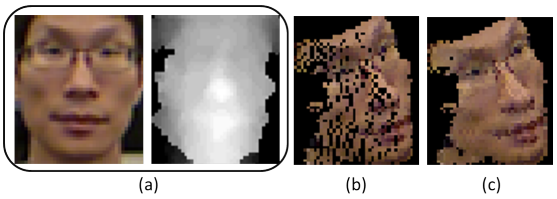

Figure 2: The sequence of steps for generating a pose-transformed version of a frontal RGBD image: (a) the original RGBD image for the frontal pose (the RGB data is shown on the left side and the depth is shown on the right); (b) pose transformed and projected result from the data in (a); (c) 2.5D interpolated result.

3.1. Creating Viewpoint Based Face Images from a Single Frontal RGBD Scan of a Human Subject

As we have previously described in [102], the 3D position (X, Y, Z) associated with an RGBD "pixel" at the raster coordinates (x, y) is given by:

$$X = \frac{Z_D}{f_c}(x - u_x),\ Y = -\frac{Z_D}{f_c}(y - u_y),\ Z = Z_D, \quad (1)$$

where Z_D is the depth value recorded by the sensor, f_c is the focal length, and u_x and u_y are the center coordinates of the image plane. Given the 3D points obtained in this manner, we first remove the background by thresholding the point cloud according to its depth histogram using Otsu's algorithm [103]. The foreground, i.e., the set of points with Z coordinate lower than Otsu's threshold, corresponds to the 3D points on the surface of a face (Figure 2 (a)).

The resulting 3D point cloud model is simply a collection of 9-dimensional vectors of the form $\mathcal{M} = [\mathbf{x}_{2D}, \mathbf{Z}_D, \mathbf{X}_{3D}, \mathbf{V}_{RGB}]^T$, where $\mathbf{x}_{2D}$ and $\mathbf{Z}_D$ are a matrix of (x, y) pixel positions and their corresponding depth values, $\mathbf{X}_{3D}$ contains the three spatial coordinates and $\mathbf{V}_{RGB}$ contains the three color values.

This cloud model includes 2D-3D correspondence information with texture data. Given a single RGBD image of the frontal pose, we generate T training images by first applying T pose transformations to its point cloud, and then projecting the resulting point clouds back on the camera image plane. The process to generate a virtual view image I_t is described by

$$I_t = \mathcal{T}\left(\mathbf{K}\,[\mathbf{I}\,|\,\mathbf{0}^T]G(\mathbf{p})\,\mathbf{X}_{3D}\right), \tag{2}$$

where $\mathbf{K}$ is the intrinsic camera calibration matrix, $\mathcal{T}(\cdot)$ stands for the conversion from the vectorized image with RGB values to the 2D image on the camera image plane, $G(\cdot)$ is the 3D transformation including the translation parameters $\mathbf{t} = [\ t_x \quad t_y \quad t_z\ ]^T$ and the Euler rotation matrix $\mathbf{R}$ computed from the rotation parameters $\boldsymbol{\theta} = [\ \theta_{rx} \quad \theta_{ry} \quad \theta_{rz}\ ]^T$ as

$$G(\mathbf{p}) = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix}, \tag{3}$$

where $\mathbf{p} = [\ \theta_{rx} \quad \theta_{ry} \quad \theta_{rz} \quad t_x \quad t_y \quad t_z\ ]$ is the pose parameter vector. In short, given a model $\mathcal{M}$, RGB values with respect to 3D points are matched to a new 2D image plane by transformation and projection.

Generating 2D images projected from rotated 3D points has two general problems to be considered. First, some 3D points in the model can overlap in a projected image plane. To handle this, only the closest sample to the camera is projected to the image plane. Second, when a pose-transformed point cloud is projected back onto the camera image plane, one often ends up with "holes" on account of the variable depth resolution of an RGBD sensor. An example of this effect is shown in Figure 2 (b), which illustrates a pose-transformed version of the frontal RGBD image. We eliminate such holes by applying bilinear interpolation to neighboring points using the constraint that the points used for bilinear interpolation possess roughly the same depth values. Figure 2 (c) shows a projection when such interpolation is a part of the projection operator. Figure 3 shows additional examples of the training images generated according to this process.

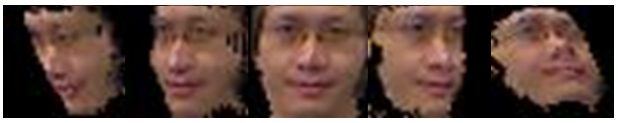

Figure 3: Examples of generated training images for one subject.

3.2. Applying ISOMAP for Clustering Multi-subject Face Images

When face images are viewed from different directions, the image data falls on a low-dimensional manifold in a high-dimensional measurement space [12, 13, 14]. It is this fact that is responsible for much interest in topics such as manifold-based learning and data clustering [104, 105, 106, 107, 89, 108, 109]. Much of this work is based on the intuition that if we could first create an appropriate low-dimensional representation for the underlying manifold, that would simplify the logic needed for establishing the decision boundaries required for the classification of the data.

We have previously investigated three of the main methods existing today for understanding data on manifolds, namely: 1) Locally Linear Embedding (LLE) [107]; 2) ISOMAP [89]; and 3) Representations that can be obtained by the Kambhatla and Leen algorithm [106]. Our study concluded that ISOMAP gives us the best partitioning of the data in terms of giving us the least average reconstruction error in the subspaces in each of the view partitions of the data [110]. The goal in this section is to demonstrate the clustering that is achieved when ISOMAP is applied to the multi-subject face images.

As described in the previous section, we record a single frontal RGBD scan for each human subject and then create viewpoint dependent training images from the scan by applying an appropriate projection transform to the scan. The clustering results we show in this section were obtained on the image data collected in this manner. These results involve the training images collected from the RGBD scans for the three subjects shown in Figure 4 (a).

Note the manifold structure for each of the three subjects in Figure 4 (b) in the space spanned by the three leading eigenvectors when all of the data for all three subjects is subject to a PCA based dimen-

sionality reduction. Each subject-specific manifold in this figure is illustrated with a different color that matches the color of the border for the corresponding human subject in (a). As the reader can see, all three manifolds look similar globally. However, when the manifolds are examined more carefully by focusing on the local curvatures, one can see the differences between the three that are caused by the different facial features, eyewear, etc. Shown in (c) is the mean manifold for the three subjects. The mean manifold is obtained by averaging the three principal coordinates in the 3D PCA space on the basis of the identity of the pose labels associated with the images. Note that Figures 4 (a), (b), and (c) are just for human visualization of the structure of the image data for the three human subjects.

With regard to the dimensionality reduction of this face data with ISOMAP, the extent to which the algorithm can capture both the global shape variations in the manifolds shown in Figure 4 (b) and, at the same time, retain the local shape characteristics, depends on the parameter γ, which controls the size of the immediate neighborhood of a data point that ISOMAP uses for calculating point-to-point geodesic distances. Figures 5 (a), (b), and (c) show how the ISOMAP representation calculated from the original data changes as we vary γ. What the ISOMAP algorithm accomplishes can be thought of as the unfolding of the manifold. Since small values of γ will cause geodesic distances to become more sensitive to local shape variations in the manifold, it is not surprising that the "unfolded manifolds" returned by ISOMAP for $\gamma = 6$ look like what is shown in Figure 5 (a). As this parameter becomes larger and larger, the sensitivity to small shape variations disappears and what emerges is the overall global shape as seen in Figure 5 (c). When we apply the KMeans algorithm to the ISOMAP representation with $K = 9$ pose classes, the corresponding clustering results we get are as shown in Figure 5 for the three different values of γ .

We now show that clustering of the multi-subject data represented by the results in Figure 5 does NOT yield a usable partitioning of the view sphere. Shown in Figure 6 is a random sampling of the images in each of the clusters in Figure 5. What is even more important with regard to the results shown in Figure 6 are the triple of data entries, with each entry of the form `SI : X` where `SI` is one of $\{\texttt{S1}, \texttt{S2}, \texttt{S3}\}$ and where `X` is an integer. The three entries $\{\texttt{S1}, \texttt{S2}, \texttt{S3}\}$ stand for "Subject 1," "Subject 2," and "Subject 3," respectively, these being the three subjects arranged left-to-right in Figure 4 (a). The integer `X` in `SI : X` stands for the number of images for the subject `SI` in the cluster. Given this notation, out of 9 clusters, we have 5 clusters that consist exclusively of images from the same subject. Additionally, in the remaining 4 clusters, we have exactly 2 subjects represented. There does not exist a single cluster that contains images from all three subjects. It is therefore evident that the sort of viewspace we achieve automatically with such partitioning does not correspond to an even distribution of the different face poses of the three subjects. As we have shown in [110], however, this algorithm does typically give us a good view sphere partitioning of the images as long as they belong to a single human subject.

As it turns out what works for the case when multi-subject images are considered together is a pose based partitioning of the viewspace. As far as pose-based criteria are concerned, that is accomplished trivially since the images generated from the RGBD data are tagged with the face poses. Figure 7 (a) illustrates nine partitions that are manually determined in the pitch and yaw space. Shown in (b) of the figure is a visualization of all the images in the space spanned by the three leading eigenvectors extracted from all of the images. The (c) of the figure shows the partitioning applied to the mean of the images from the three human subjects, the means being computed for the same pose parameters.

In the following section, we will discuss locally optimum subspace construction from the given clusters of images. Before launching into the next section, we must point out that there has been much research in the past in fitting locally linear subspaces to data that resides on nonlinear manifolds [111, 112, 5, 72].

3.3. Constructing Subspaces from View-Partioned Clusters

Whether we cluster all of the training data using a pose-based approach or separately cluster the data corresponding to each individual using an

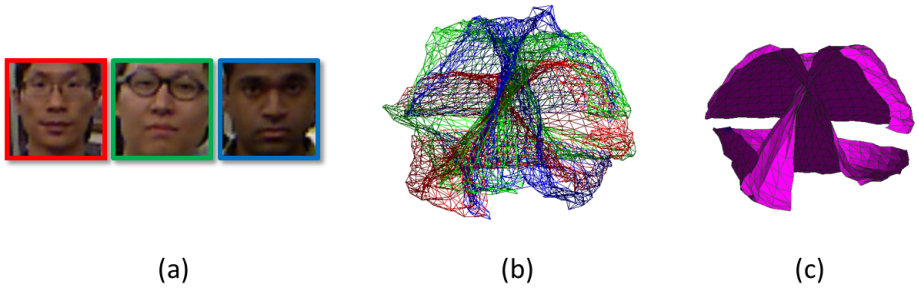

Figure 4: Visualization of the manifolds corresponding to three subjects as obtained by ISOMAP: (a) Three subjects, (b) Visualization of person-specific manifold structure in the PCA space, (c) Mean manifold for the person-specific manifolds in (b).

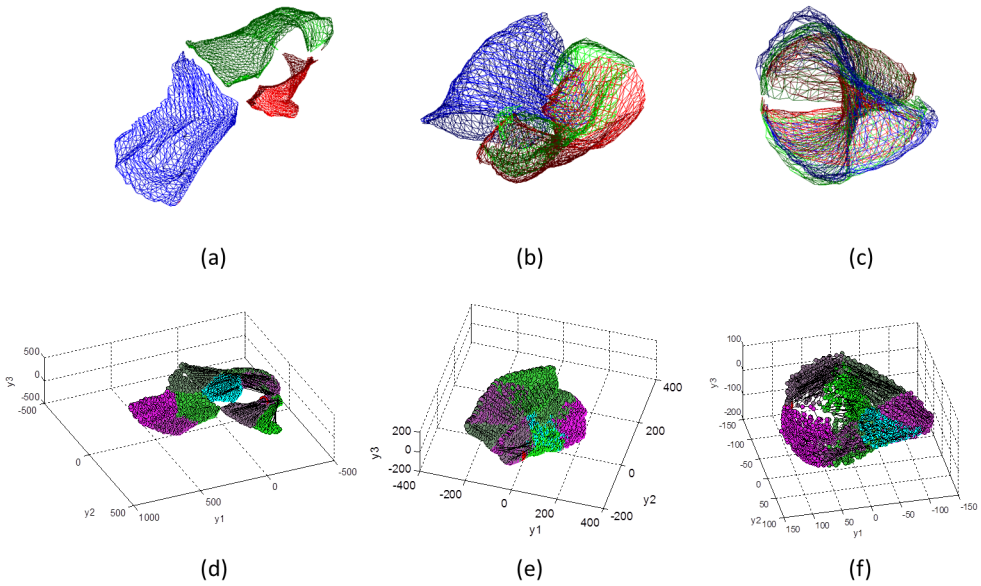

Figure 5: Top row: ISOMAP-based representation of multi-subject face images with (a) $\gamma = 6$, (b) $\gamma = 10$, (c) $\gamma = 27$. Bottom row: Clustering results in ISOMAP representation with (d) $\gamma = 6$, (e) $\gamma = 10$, and (f) $\gamma = 27$.

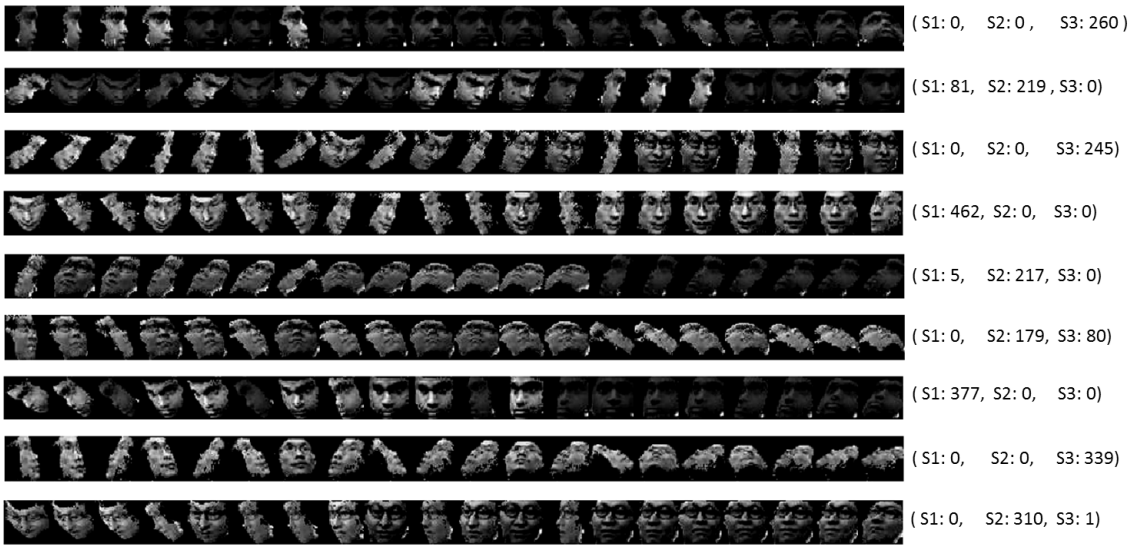

Figure 6: Clustered image samples on the result of Figure 5 (d) with $K = 9$ and $\gamma = 6$ for three subjects.

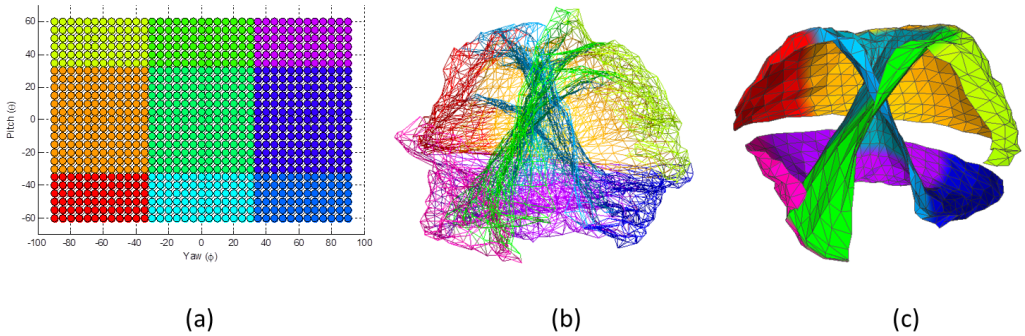

Figure 7: Visualization of the pose-based clustering with $K = 9$: (a) Manual pose partition, (b) Partitioned subject-specific manifolds in the PCA space, (c) A partitioned mean manifold in the PCA space.

appearance-based method, we need to decide how to handle the person-to-person variations in the training data. That is, we need to choose whether the multi-subject data should be represented through common view subspaces as at node 2 in Figure 1, or through a finer person-specific decomposition as at nodes 5 and 9. In the common-view subspace we place the training images of all the subjects in a single pose-partitioned subspace. For the person-specific subspaces, images of each subject are placed in either a single pose-partitioned subspace or in a single appearance-based cluster.

Recent literature in face recognition suggests that we are likely to achieve higher recognition accuracies if we construct person-specific subspaces [113, 5, 114, 115, 116, 117]. The reason has to do with the fact that, as mentioned above, each individual's face data resides on a separate manifold, but the structure of these manifolds gets lost in a low-dimensional subspace that integrates over all of the data for all the training subjects. One can argue that if an attempt was made to retain the manifold structure in the low-dimensional space constructed using PCA — as would be the case in person-specific subspaces — one would get better results no matter what classification rule is used for face recognition.

In light of the merits of the person-specific subspaces as stated in the literature, but keeping in mind that not enough is known about what strategies might work the best for face recognition in the wild, we keep both options open. That is, this work evaluates both the Common View Subspace (CVS) construction and what we refer to as Person Specific Subspaces (PSS).

3.3.1. Common View Subspace and Person Specific Subspace Models

The CVS model is only available through the pose-based partitioning criterion as shown at node 3 of Figure 1 (as mentioned in Section 3.2 above, partitioning the CVS model based on subject appearance does not provide a useful representation). We call this model Pose-CVS; it is created by first partitioning the view sphere and then placing all of the training images for all the subjects in a common subspace for each partition. As a result, the CVS model consists of multiple PCA subspaces, one for each pose partition, and the principal components of training samples in each subspace. Here, each training sample is labeled with the index of a human subject. Accordingly, for a given number of views K, the CVS model is represented by

$$Model_{CVS} = \left\{ \left\{ S^{(k)}, \mathbf{Y}_h^{(k)} \right\}_{k=1}^{K} \right\}_{h=1}^{H}, \quad (4)$$

$$= \{L_{cvs,h}\}_{h=1}^{H}, \quad (5)$$

where $L_{cvs,h} = \left\{ S^{(k)}, \mathbf{Y}_h^{(k)} \right\}_{k=1}^{K}$, the k^{th} cluster-based subspace is $S^{(k)} =< \mathbf{r}^{(k)}, \mathbf{U}^{(k)}, \mathbf{\Lambda}^{(k)} >$ with center $\mathbf{r}^{(k)}$, eigenvector matrix $\mathbf{U}^{(k)}$, and eigenvalues matrix $\mathbf{\Lambda}^{(k)}$. Additionally, $\mathbf{Y}_h^{(k)}$ denotes a set of projected training samples with respect to the h^{th} subject in the k^{th} cluster. Here, we can also interpret $\mathbf{Y}_h^{(k)}$ as the set of points of the h^{th} subject on the hyperplane represented by $S^{(k)}$.

Again, as shown in Figure 1, the person-specific subspaces can be constructed for either pose-based partitioning of the view sphere or appearance-based partitioning (nodes 6 and 10, respectively). When the view sphere is partitioned directly in the pose space, as at node 1, the person-specific subspaces are constructed by fitting a PCA model to all the training images for each human subject separately. We call this approach Pose-PSS. On the other hand, in nodes 7 and 8, we first partition all the training images on the basis of their human identity and carry out appearance-based clustering of the images for each human subject using ISOMAP followed by K-Means clustering (see [110] for more details). This approach is called App-PSS. Consequently, both PSS models can be expressed in the following form:

$$Model_{PSS} = \left\{ \left\{ S_h^{(k)}, \mathbf{Y}_h^{(k)} \right\}_{k=1}^{K} \right\}_{h=1}^{H}, \quad (6)$$

$$= \{L_{pss,h}\}_{h=1}^{H}, \quad (7)$$

where K is the number of clusters formed for each human (based on pose or appearance), and H denotes the number of human subjects. The k^{th} cluster-based subspace for the h^{th} subject is represented as $S_h^{(k)} =< \mathbf{r}_h^{(k)}, \mathbf{U}_h^{(k)}, \mathbf{\Lambda}_h^{(k)} >$ with $\mathbf{r}_h^{(k)}$ the center of the

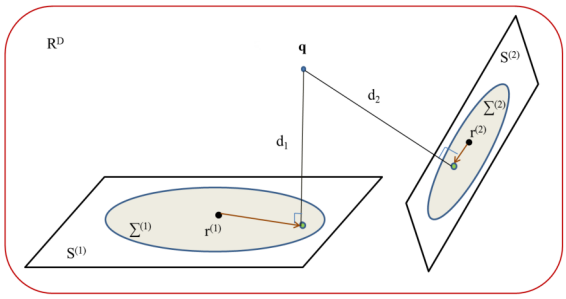

Figure 8: Geometric interpretation of the reconstruction error distance for two subspaces $S^{(1)}$and $S^{(2)}$ in R^D.

k^{th} cluster, eigenvector matrix $\mathbf{U}_h^{(k)}$, and eigenvalues matrix $\mathbf{\Lambda}_h^{(k)}$. Also, $\mathbf{Y}_h^{(k)}$ is the set of points on the hyperplane represented by $S_h^{(k)}$.

3.3.2. Overall Classification Logic for a Test Image

When we use the above subspace models to classify a query image, we employ a nearest-subspace (NS) classifier that chooses a subspace in terms of the smallest reconstruction error. The reconstruction error distance calculates the orthogonal distance from a query to the hyperplane obtained by PCA. Figure 8 illustrates the reconstruction error distance from a query image point $\mathbf{q}$ to two hyperplanes $S^{(1)}$ and $S^{(2)}$ in the underlying R^D space. The reconstruction error distance is given by:

$$\begin{aligned} d(\mathbf{q}, S^{(k)}) &= \left\| \bar{\mathbf{F}}^{(k)T}(\mathbf{q} - \mathbf{r}^{(k)}) \right\|^2, \\ &= \left\| \mathbf{q} - \mathbf{F}^{(k)}\mathbf{F}^{(k)T}(\mathbf{q} - \mathbf{r}^{(k)}) \right\|^2, \quad (8) \end{aligned}$$

where $\mathbf{U}^{(k)} = \left[\mathbf{F}^{(k)}\ \bar{\mathbf{F}}^{(k)}\right]$ is the matrix whose columns are eigenvectors of the covariance matrix obtained from samples in the k^{th} subspace. $\mathbf{F}^{(k)}$ consists of the d leading eigenvectors and $\bar{\mathbf{F}}^{(k)}$ has the $(n-d)$ trailing eigenvectors of $\mathbf{U}^{(k)}$, which has rank$(\mathbf{U}^{(k)}) = n$.

App-PSS and Pose-PSS return the human identity using only the NS classifier. On the other hand, Pose-CVS requires two layered classifiers to decide the human label. First we need to select a subspace and then classify the human label with respect to that subspace. We consider two different classifiers in the second layer of Pose-CVS, a nearest neighbor (NN) classifier and an SVM classifier [118, 119].

Figure 9 illustrates the classification logic for each of the models we consider. Figure 9 (a) shows the classification logic used for both PSS approaches. From the $N \times K$ subspaces available, the test image is assigned to that subspace for which the reconstruction error distance is the smallest. This directly yields the person ID for the test images since each subspace is person specific. In other words, for the PSS model $Model_{PSS}$, given a query $\mathbf{q}$, recognizing a face is simply achieved by the nearest subspace classifier as

$$h^* = \arg\min_h d(\mathbf{q}, S_h^{(k)}), \quad (9)$$

where $d(\mathbf{q}, S_h^{(k)})$ denotes the reconstruction distance from a point $\mathbf{q}$ to the k^{th} hyperplane of the h^{th} subject.

Figures 9 (b) and (c) show how to work with two layered classifiers for the Pose-CVS model. The first layer classifier of this model is similar to the PSS models. For a query image $\mathbf{q}$, we first find the best subspace to use by minimizing the reconstruction distance as

$$j = \arg\min_k d(\mathbf{q}, S^{(k)}), \text{ for k=1,}\cdots\text{,K.} \quad (10)$$

As for the second layer classifier, Figure 9 (b) shows the NN classifier and (c) depicts the SVM classifier where LSVM and RKSVM stand for linear SVM and RBF kernel SVM, respectively.

For the NN classifier, let the training samples $\mathbf{x}_i^{(j)}$ in the j^{th} subspace thus ascertained have their local-subspace representations given by the vectors $\mathbf{y}_i^{(j)} = \mathbf{U}^{(j)}(\mathbf{x}_i^{(j)} - \mathbf{r}^{(j)})$ for $\mathbf{y}_i^{(j)} \in \mathbf{Y}_h^{(j)}$ and $i = 1, \cdots, T_j$ where T_j is the number of samples in the j^{th} subspace. Subsequently, we search in the local subspace for that training image which is closest to the query image $\mathbf{q}$. That is, we find

$$l^* = \arg\min_i \left\| \mathbf{y}_i^{(j)} - \mathbf{U}^{(j)}(\mathbf{q} - \mathbf{r}^{(j)}) \right\|^2, \text{ for } i = 1, \cdots, N_j. \quad (11)$$

The person label returned by the query image $\mathbf{q}$ is the label h associated with the nearest training sample

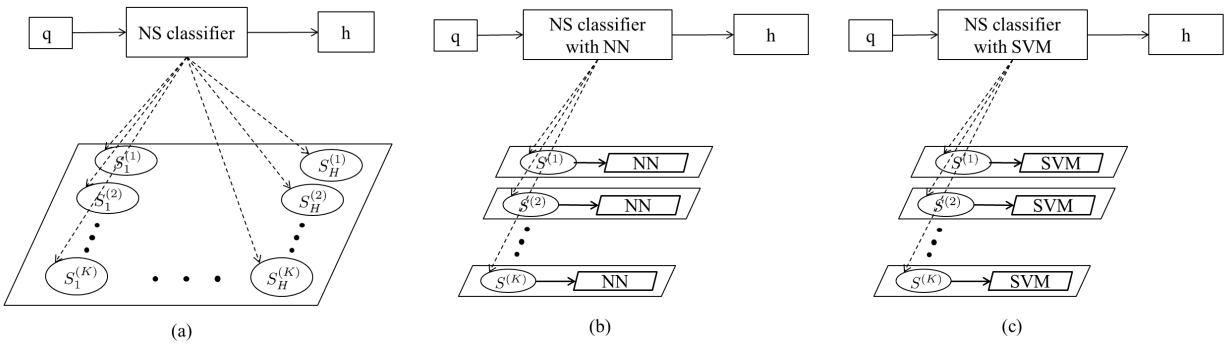

Figure 9: Classification logic for: (a) App-PSS and Pose-PSS, (b) Pose-CVS-NN, (c) Pose-CVS-LSVM and Pose-CVS-RKSVM.

image represented in the local subspace by the vector $\mathbf{y}_{l^*}^{(j)}$.

For the SVM classifier, the person label is returned by the SVM classifier trained from the local-subspace representation of the training samples $\mathbf{x}_i^{(j)}$ in the j^{th} subspace. During the training procedure, the SVM classifiers associated with common-view subspaces are learned from the training samples projected in each subspace. Here, we consider two popular kernels: a linear kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$ and a non-linear kernel with radial basis function, $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$. In this paper, we utilize the multi-class SVM in Chang and Lin's LibSVM [120], which is based on the one-against-one approach in which we have one SVM for each pair of classes [121].

4. COMBINING IDENTITY LABELS FROM MULTIPLE VIEWPOINTS

This section addresses the question of combining identity labels of face images recorded from a collection of viewpoints. Combining the identity labels for the global approach to face classification, that is when all of the training data calculated from the RGBD images resides in a single low-dimensional subspace, is relatively straightforward. The most commonly used approach in the literature for this purpose is that of majority voting. That will be the method we will also use in this section for the global case.

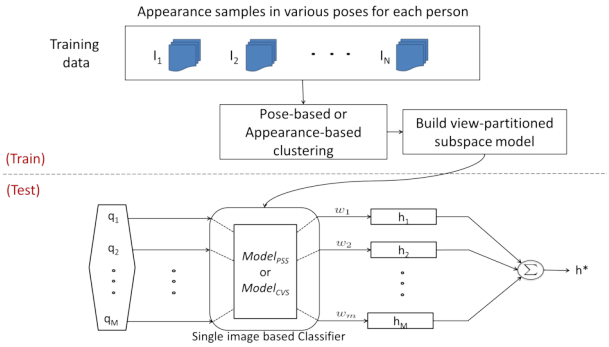

Figure 10: A weighted voting framework for multi-view inputs.

The view-partitioned subspaces open up the possibility of integrating the labels by giving greater weight to query images that can be associated with viewpoints that carry greater discriminatory power for determining the identity of a face. It should be intuitively obvious that frontal and near-frontal viewpoints are more important than images that are recorded from sharply lateral viewpoints. Hence, we investigate this weighted voting approach to combine identity labels in view-partitioned subspaces.

Figure 10 is a visual representation of our overall framework for training and testing the system for recognizing a face from a set of query images. The labels I_1, I_2, ..., I_N in the top box in the figure represent the different subjects in the population on which the system is trained. We assume we have access to a sin-

gle frontal RGBD image for the face of each subject. As explained in Section 3.1, we generate from each RGBD image a set of 2D images of the face from a large number of different viewpoints. These images are then partitioned into K clusters on the basis of either pose-based partitioning or appearance-based partitioning. Subsequently, we construct a subspace for each partition of the training images thus created.

For the testing phase, as shown below the dotted line in Figure 10, we are given M query images of the same individual, $\mathbf{q}_1, \cdots, \mathbf{q}_M$. As to how these images are processed depends on whether we use the view-partitioned subspaces where each subspace represents the data from all of the subjects ($Model_{CVS}$) or the view-partitioned subspaces in which each subject in the population gets his/her own view-partitioned subspace ($Model_{PSS}$). The specific classifiers for each model were described in the previous section.

In either case, the output of this step for each query image is an identity label. In general, we may associate a weight w_i with the identity label estimated for the i^{th} query image and then construct a weighted aggregation of the identity labels for the final recognition label. The weights reflect the degree of trust we place in a given query image. When the final identity label is calculated with simple majority voting, the weights w_i all become 1.

4.1. Weighted Voting by Normalized Reconstruction Error Distance

In the view-partitioned case, we consider the normalized reconstruction error distance as the weight to be assigned to a query image. That is, if a query image $\mathbf{q}$ is assigned to a subspace $S^{(k)}$ (or $S_h^{(k)}$ for the person-specific models), we compute the reconstruction error when $\mathbf{q}$ is projected into the subspace $S^{(k)}$ and normalize it by the mean value of the error between $\mathbf{q}$ and *all* the subspaces.[1] The inverse of this error then becomes the weight to be assigned to the classification label that is given to $\mathbf{q}$ by the subspace $S^{(k)}$.

For the PSS model, the least reconstruction error distance for the i^{th} query $\mathbf{q}_i$ is obtained by

$$\varepsilon(\mathbf{q}_i) = \min_{h,k} \left[d(\mathbf{q}_i, S_h^{(k)}) \right], \tag{12}$$

where $d(\mathbf{q}, S_h^{(k)})$ denotes the reconstruction error distance of $\mathbf{q}$ to the k^{th} subspace of the h^{th} person given by Eq. 8 (see Appendix B in [110] for more details). Similarly, for the CVS model, the minimum reconstruction error distance for a query $\mathbf{q}$ is obtained by

$$\varepsilon(\mathbf{q}_i) = \min_{k} \left[d(\mathbf{q}_i, S^{(k)}) \right]. \tag{13}$$

Then, the normalized minimum distance is given by

$$\widetilde{\varepsilon}(\mathbf{q}_i) = \frac{\varepsilon(\mathbf{q}_i)}{\frac{1}{H \cdot K} \sum_{h=1}^{H} \sum_{k=1}^{K} d(\mathbf{q}_i, S_h^{(k)})}, \tag{14}$$

for PSS, and

$$\widetilde{\varepsilon}(\mathbf{q}_i) = \frac{\varepsilon(\mathbf{q}_i)}{\frac{1}{K} \sum_{k=1}^{K} d(\mathbf{q}_i, S^{(k)})}, \tag{15}$$

for CVS. The weight for the i^{th} query $\mathbf{q}_i$ is determined in inverse proportion to $\widetilde{\varepsilon}(\mathbf{q}_i)$ as

$$w(\mathbf{q}_i) = \frac{1}{\widetilde{\varepsilon}(\mathbf{q}_i)}. \tag{16}$$

To summarize, given a query image $\mathbf{q}$, let the values for the reconstruction error between $\mathbf{q}$ and the subspaces $S^{(1)}, S^{(2)}, \ldots$ be denoted $\varepsilon_1, \varepsilon_2, \ldots$. For the purpose of class label calculation, we assign $\mathbf{q}$ to the subspace $S^{(i)}$ if $\varepsilon_i < \varepsilon_j$ for all $j \neq i$. Then, to combine the classifications returned for all the query images, the class label calculated for $\mathbf{q}$ is weighted in inverse proportion to ε_i (after normalization).

5. RESULTS

The preceding discussions have raised a number of important research questions which we now address with an extensive experimental evaluation. More specifically, these research questions deal with the issues of 1) whether view partitioning indeed improves

[1] Note that, since we need to calculate the reconstruction error between $\mathbf{q}$ and all the subspaces anyway in order to figure out which subspace is best for $\mathbf{q}$, no additional computations are involved in the normalization of the reconstruction errors.

Figure 11: Frontal faces of 10 human subjects in the RVL face dataset.

the performance of a classifier with respect to global approaches, 2) what the effect of the number of such partitions is on the performance, 3) whether such partitions should be carried out based on the subject pose or appearance, 4) what the impact of the dimensionality of the subspaces generated during partitioning is on the performance of the system, 5) whether a classification system can indeed benefit from aggregating multiple images from different perspectives, and if so 6) whether the proposed weighting mechanism can further improve the system performance with respect to simple majority voting aggregation.

In order to quantitatively assess the relative merits of the various classifying strategies, we use three RGBD datasets. The first is the RVL face dataset that we created, consisting of 10 human subjects. Some example images from the RVL dataset are shown in figures 11 and 13. The second is a public dataset consisting of 28 subjects from the Visual Analysis of People (VAP) lab at Aalborg University [122]. We will refer to this dataset as the VAP dataset in the rest of this paper. The third dataset is the ETH BIWI Kinect Dataset [123], another publicly available dataset consisting of 20 distinct test subjects. Finally, we also compare our methods to the state-of-the-art face classification approach proposed by Parkhi et al. in [124].

5.1. Comparison of the Discriminative Power of View-partitioned Subspaces

The goal of this section is to measure the class discriminatory information retained in the different subspace models by a 10-fold cross validation test. For the evaluation in this section, we generated 200 multi-view images for each of 10 human subjects from a single frontal RGBD image for each subject, as described in Section 3.1. Figure 11 illustrates frontal images of the 10 subjects. For 10-fold cross validation, we randomly shuffle the 200 images for each human subject. For each fold of the 10-fold test, we use 180 of these for training and the remaining 20 for testing.

During training, we generate the three models Pose-CVS, Pose-PSS, and App-PSS (see Figure 1). As previously mentioned, the Pose-CVS model has three variants: Pose-CVS-NN, Pose-CVS-LSVM, Pose-CVS-RKSVM according to the second layer classifier. For details about the classification logic of each model, the reader can refer to Figure 9. Our performance evaluation in each case is with respect to the number of clusters K and the dimensionality of the subspaces d. The baseline method is the RBF kernel-based SVM classifier with no partitioning because this type of classifier has been successfully used in practice.

Figure 12 shows the accuracy of each model with respect to the dimensionality d and the number of partitions K. In (a) of the figure, a comparison of all models is presented with the baseline when there is no partition. As the reader can see, the PSS model is not only better than the linear SVM and the global NN but is also comparable to the Nonlinear SVM. For the case of the Pose-CVS model, when we use the NN classifier, it approaches the global NN model as d increases. When we use the linear SVM and RBF kernel-based SVM, they converge to the baseline (although Pose-CVS-LSVM requires a substantially higher dimensionality than shown in the figure to do so). In terms of the number of view-partitions, as Figures 12 (b) to (f) show, as K increases, each of the models converges to its maximum accuracy in a smaller dimensional subspace. The Pose-CVS-LSVM model, shown in Figure 12 (c), for example, requires a dimension of approximately 200 to approach its saturated accuracy with $K = 1$, but for $K = 9$, it surpasses global linear SVM with as little as $d = 20$ dimensions.

5.2. Results on the RVL Face Dataset

In this section, we evaluate different mechanisms to combine information from multiple views in order to carry out face classification. We first demonstrate the application of the majority voting rule to the case when we use a single subspace for representing all of the training data (i.e., when $K = 1$). We then extend the majority voting approach to the case of

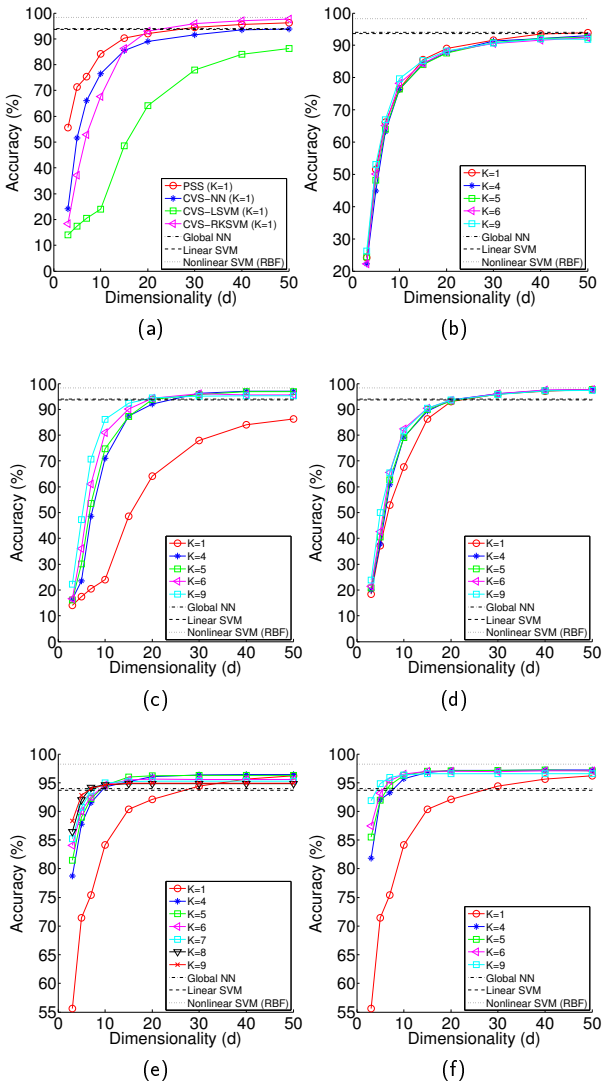

Figure 12: Accuracy vs. subspace dimensionality with respect to the number of partitions K for: (a) All models at $K = 1$, (b) Pose-CVS-NN, (c) Pose-CVS-LSVM, (d) Pose-CVS-RKSVM, (e) App-PSS, and (f) Pose-PSS. In the Global NN, Linear SVM and Nonlinear SVM approaches, all the samples are placed in a common global space without dimensionality reduction and classification is performed respectively with a Nearest Neighbor, Linear SVM, and Nonlinear SVM with an RBF kernel classifier.

Figure 13: Examples of all viewpoints of one subject in the RVL Face dataset.

view-partitioned subspaces (i.e., $K > 1$) and compare the results obtained with those of the non-partitioned approach. Finally, we consider the case of weighted voting for view-partitioned subspaces in which the weights depend upon the least reconstruction error distances.

All of our results in this section are based on the training data collected from the 10 human subjects whose 2D images are shown in Figure 11. For each subject, we record a single frontal RGBD image and from that image we generate 925 viewpoint variant images for that subject. The viewpoint variant images cover an angular range of $[-90°, 90°]$ in yaw and $[-60°, 60°]$ in pitch with respect to the frontal position of the face in steps of 5°. For the test data, we use a separate set of face images recorded from different viewpoints. To emphasize, the test data is NOT drawn from the RGBD based 2D images generated for each subject. We record separately a set of 17 images for each subject with different orientations of the face vis-à-vis the camera. Note that these are purely 2D images. No particular constraint is placed on the relationship of the head pose to the location/orientation of the camera — except for ensuring that the face is sufficiently visible in the camera images. Shown in Figure 13 are such test images for one of the subjects.

5.2.1. Majority Voting for a Non-partitioned Subspace

As stated earlier, we now place all of the training data in a single non-partitioned subspace. Although the main focus in this section is to show results with a single subspace, for the sake of completeness we also show results with an extension of the idea — we create person-specific subspaces *but with NO viewpoint partitioning*. While the former corresponds to the CVS model with $K = 1$, the latter is equivalent to

either of the PSS models also with $K = 1$. The results shown in this section demonstrate how the classification error varies as we change the dimensionality d of the single subspace and as we change the number M of query images available.

Figure 14 shows the classification accuracy as a function of the dimensionality of the subspace. Each datapoint in Figure 14 as well as in the remainder of this section corresponds to the average over 100 independent realizations of the experiment. The accuracy results plotted in Figure 14 indicate that the classification accuracy decreases rapidly when the dimensionality of the subspace is made larger than approximately 20. That indicates that the intrinsic dimensionality of the data is no more than 20 and we have mostly measurement noise beyond that. The most significant result in Figure 14 is that multi-view classification, that is, when M is greater than 1, definitely contributes to increases in overall classification accuracy.

In order to examine the results plotted in Figure 14 from a different perspective, shown in Figure 15 (a) are the same results for a fixed value of 20 for the subspace dimensionality and as a function of the number of views M. It is interesting to observe that, when the test images are drawn from a separate dataset, the PSS approach performs comparably to the nonlinear SVM for any number of query images. Figure 15 (b) shows the time performance of the classifiers for the same three cases as in (a).

5.2.2. Majority Voting for View-Partitioned Subspaces

This section presents the results obtained when the classification results generated by multiple views are combined by a simple majority voting approach in which the contributions from each view are equally weighted.

Figure 16 shows the multi-view classification accuracy as a function of dimensionality for the Pose-CVS model. In comparison with the non-partitioned case, the accuracy does not fall off as rapidly when we increase the dimensionality beyond 20. Instead, we see less pronounced peaks at a dimensionality of approximately 50, which indicates that the dimensionality of the data is dependent on the complexity of its subspace representation. The peak is slightly more pronounced on the linear SVM, indicating that the non-linear SVM is marginally more robust to the noise added by the extra dimensions. On average, when the dimensionality d is approximately 50 and both methods show their peak performance, the linear SVM performs about 5% better than the nonlinear SVM, which indicates that for properly modeled subspaces, the additional complexity of an RBF kernel is not justified. Figure 17 shows results similar to those in Figure 16 for the PSS model. As with the CVS model, here again the accuracy does not fall off as rapidly when we increase the dimensionality beyond 20. Instead we see a peak at a dimensionality of approximately 30, which again indicates the dependence of the dimensionality on the model. The peak is significantly more noticeable in the App-PSS model, which indicates that, similar to the Pose-CVS-LKSVM, it is less robust to noise at higher dimensionalities.

Figure 18 (a) shows how the classification accuracy depends on the number M of query images for a fixed value of the subspace dimensionality $d = 20$ and view partitions $K = 25$. For the RVL dataset, the PSS models show marginally higher performance than the CVS models with either a nonlinear or a linear kernel for seven or more views. Shown in (b) of the same figure are the time performance comparisons for the four approaches shown in (a). As can be seen from the plots in (b), CVS based classification with pose partitioned subspaces gives the best time performance.

Figure 19 shows a comparison of the non-partitioned approaches of section 5.2.1 with the view-partitioned approaches of this section. As is evident from this figure, the multi-view classification approaches with view-partitioned subspaces tend to significantly outperform the non-partitioned subspace methods, particularly when the number of views is greater than 5. Shown in (b) is a comparison of the time performance numbers associated with all cases in (a). This figure tells us that there is a cost associated with the superior classification accuracies one achieves with person-specific view-partitioned multi-view classification — increased time to arrive at the results. As we increase the number of views, the

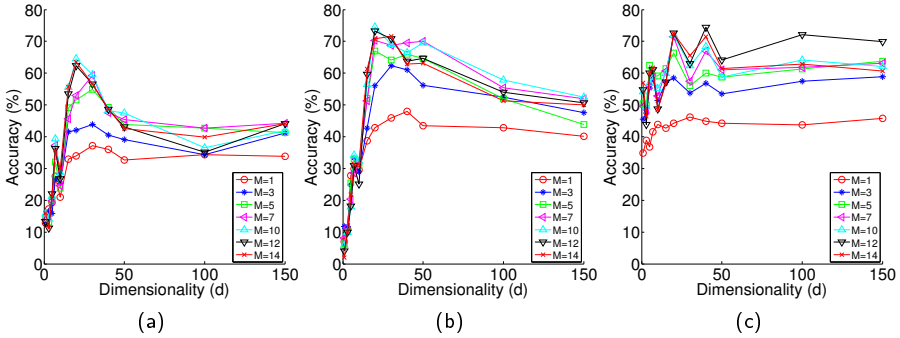

Figure 14: Multi-view classification accuracy with a single non-partitioned subspace and majority voting as a function of the subspace dimensionality d and the number M of query images. (a) Results with a linear SVM classifier (CVS-LSVM), (b) Results with RBF-kernel based SVM classifier (CVS-RKSVM), and (c) a single subspace for each individual separately (PSS).

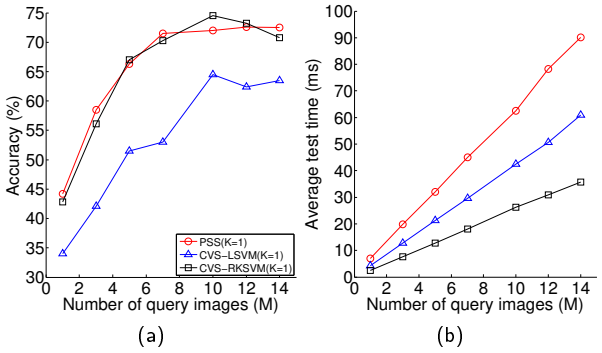

Figure 15: Classification performance as a function of the number of query images M for a single non-partitioned subspace with the dimensionality $d = 20$. (a) Classification accuracies. (b) Time performance of the classifiers for the three cases in (a).

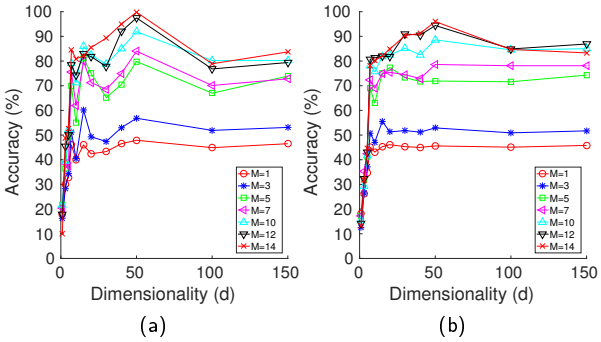

Figure 16: Multi-view classification accuracy versus subspace dimensionality for the Pose-CVS model with $K = 25$. (a) Linear SVM (Pose-CVS-LSVM). (b) Nonlinear SVM (Pose-CVS-RKSVM).

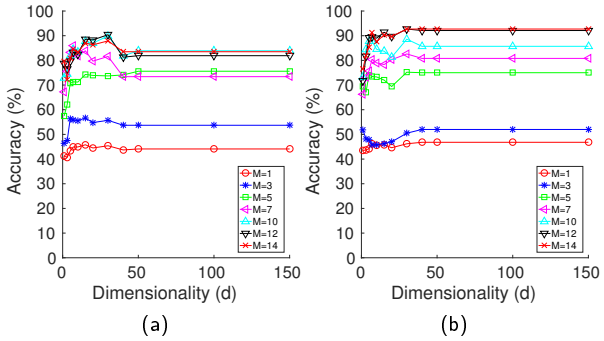

Figure 17: Multi-view classification accuracy versus the subspace dimensionality for the PSS model with $K = 25$. (a) Appearance based clustering (App-PSS). (b) Pose based clustering (Pose-PSS).

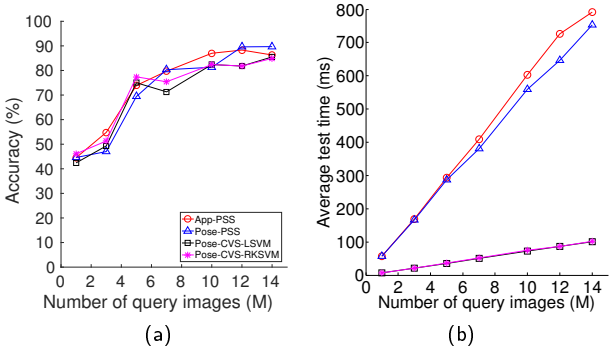

Figure 18: Classification performance as a function of the number of query images M with the dimensionality $d = 20$ and $K = 25$. (a) Classification accuracies. (b) Time performance of the classifiers in (a).

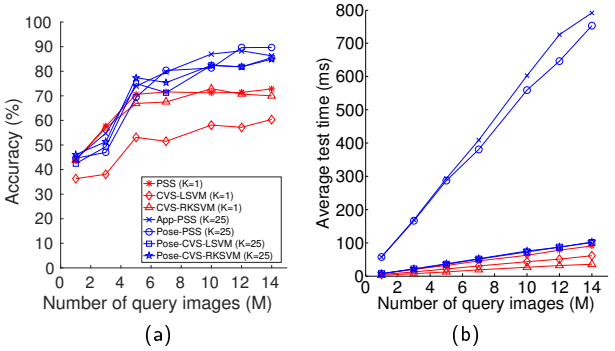

Figure 19: Comparison of the multi-view classification results for a single subspace and view-partitioned subspaces for the RVL dataset. The plots in red are for the case when single non-partitioned subspaces are used and the plots in blue are for the case when view-partitioning is applied to the training data. The subspace dimensionality d is fixed as 20 for both the red and the blue plots. The value of K is 1 for the red plots (since they correspond to the case of a single global subspace) and 25 for the blue plots. (a) Classification accuracies. (b) Time performance of the classifiers in (a).

time it takes to arrive at a classification decision by a person-specific view-partitioned classifier goes up linearly with M. On the other hand, this time increases sub-linearly for both the common-view view-partitioned classifier and the non-partitioned classifier.

5.2.3. Weighted Voting for View-Partitioned Subspaces

Figure 20 (a) compares the classification accuracy obtained with the weighted voting approach of section 4.1 to that of simple majority voting. In the figure, blue lines correspond to weighted voting and red lines to majority voting. Weighted voting improves the classification accuracy for all models. For example, when $M = 7$, weighted voting yields an overall accuracy about 14% higher than majority voting. Regarding computational time, Figure 20 (b) shows a comparison of weighted voting with the majority voting method. The average time does not change much by calculating the weights for each query. Therefore, weighted voting by normalized reconstruction error distance improves the classification accuracy without

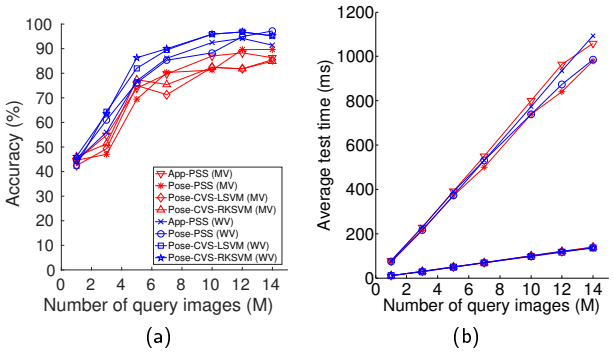

Figure 20: Comparison of weighted voting (by average normalized reconstruction error distance) with the majority voting approach for the view partitioned multi-view classification methods with $d = 20$ and $K = 25$ in terms of (a) accuracy and (b) average test time. The plots in red are for majority voting and the plots in blue are for weighted voting.

additional computational cost when compared to majority voting.

5.3. Results on the VAP Dataset

In this section, we evaluate the various classification strategies presented above on the publicly available VAP database [122]. This database has RGB images at a resolution of 1280×960 pixels and depth data at a resolution of 640×480 pixels for 31 subjects. For each subject, there are 17 different face poses. Note that in this dataset, the authors use the term 'face pose' to refer to both different face orientations vis-à-vis the sensor as well as different facial expressions. For each subject, 14 poses correspond to different orientations and 3 correspond to different expressions. We maintain this notation in the rest of our discussion. Each pose has been recorded 3 times resulting in a total of 51 images per person. More details about the dataset can be found in [122].

Two pre-processing steps are required to use this dataset for evaluating our classification strategies. First the RGB images and depth maps are not co-registered, so simple downsampling of the RGB image is not sufficient to align the two data sources. We used the Microsoft Kinect SDK to co-register them. The second step is face detection. The Haar feature based cascade classifier from OpenCV was used to detect faces in the images. We rejected false detections by using the observation that the position of the test subjects does not vary much with respect to the sensor.

Using the procedure described in Section 3.1, we use one frontal image to generate 925 viewpoint variant training images of each subject. Regarding the testing dataset, it is sufficient to ensure that each of the different views is present once in the test dataset. That can be accomplished by simply making sure that the test dataset for each individual subject contains 17 images. Using this process, we were able to extract all the required (1 training + 17 different test) face images automatically for 28 people.[2] The rest of our discussion is based on this dataset of 28 people.

One observation that we made while using these frontal images to generate the viewpoint variant images is that this dataset suffers from stronger shadow and occlusion effects when compared to our RVL Face dataset. Visualizing the point clouds in MeshLab showed that the holes in the point clouds create holes in the projected images, even after applying the depth-constrained bilinear interpolation described in Section 3.1. Some examples to illustrate these artifacts are shown in Figure 21. Such effects can impair the performance of any classifier. More robust 3-D surface reconstruction algorithms are needed to fill these holes. This is a part of our ongoing research. Despite these challenges, our method still shows very high accuracy, as demonstrated in the following sections.

5.3.1. Majority Voting

We first show results using majority voting for view partitioned subspaces. Adhering to the discussion in Section 5.2.2, we use $K = 25$ for the number of partitions. Figure 22 depicts the classification accuracy versus dimensionality for different numbers of query images for the Pose-CVS model. In the figure, as well

[2] For the remaining three individuals, we were unable to automatically extract the faces for all views using a face detector. Instead of manually processing the missing views, we chose not to include these three individuals in this evaluation.

Figure 21: Examples of holes in the viewpoint variant training images for the publicly available VAP dataset.

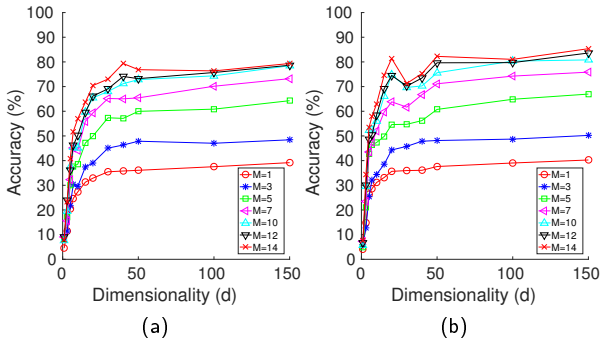

Figure 22: Multi-view classification accuracy versus subspace dimensionality for the Pose-CVS model with $K = 25$ partitions on the VAP dataset. (a) Linear SVM (Pose-CVS-LSVM). (b) Nonlinear SVM (Pose-CVS-RKSVM).

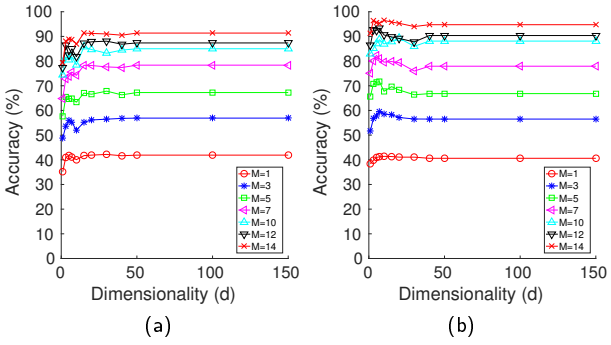

Figure 23: Multi-view classification accuracy versus the subspace dimensionality for the PSS model with $K = 25$ partitions on the VAP dataset. (a) Appearance based clustering (App-PSS). (b) Pose based clustering (Pose-PSS).

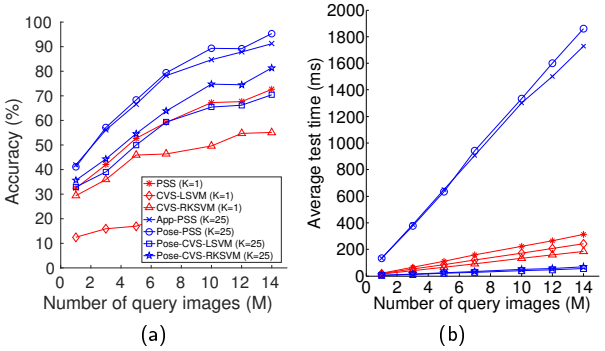

Figure 24: Comparison of multi-view classification approaches when we use majority voting as a function of the number of query images M with the dimensionality $d = 20$ in the VAP dataset. The plots in red are for the case when single global subspaces are used and the plots in blue are for the case when view-partitioning is applied to the training data. The value of K is 1 for the red plots (since they correspond to the case of a single global subspace) and 25 for the blue plots. (a) Classification accuracies. (b) Time performance of the classifiers in (a).

as in the remainder of this section, each datapoint corresponds to the average over 17 independent realizations of the experiment. Again we notice that additional query images increase accuracy and that peak accuracy is obtained at a dimensionality of 50, although less noticeably so for the RKSVM model. The performance difference between linear and nonlinear kernel SVMs is less evident in this dataset, possibly due to its additional complexity. Figure 23 illustrates the performance when we use the PSS models. Similarly, the dependence on the dimensionality of the data is less noticeable for the PSS models in this dataset.

To illustrate the relative performances of the different classifiers while using majority voting, we fix the dimensionality d as 20 and plot the accuracies for the non-partitioned and view-partitioned classifiers for different numbers of query images M in part (a) of Figure 24. Part (b) of the figure shows the corresponding time performances. As in the previous results, view partitioning significantly improves performance. Also, we notice that the PSS approaches tend outperform the CVS approaches even when a

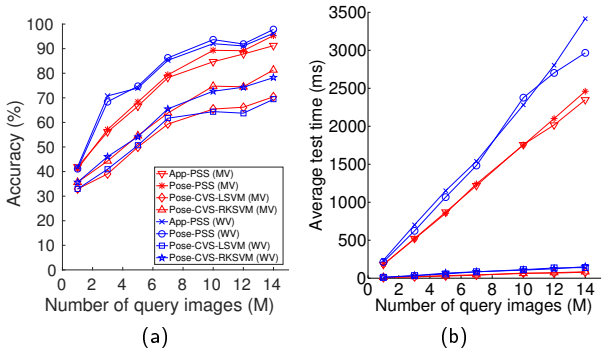

Figure 25: Comparison of weighted voting (by average normalized reconstruction error distance) with the majority voting approach for the view partitioned multi-view classification methods on the VAP dataset with $d = 20$ and $K = 25$ in terms of (a) accuracy and (b) average test time. The plots in red are for majority voting and the plots in blue are for weighted voting.

nonlinear kernel is used. This result is expected since the number of individuals in this dataset is larger and classification within a common subspace would be more challenging.

5.3.2. Weighted Voting

We now use the weighted voting approach of section 4.1 to combine the classification results from different views. The results are shown in Figure 25. For the VAP dataset weighted voting also shows a performance improvement over majority voting. Also, the PSS approaches outperform the CVS approaches in both voting schemes.

5.4. Results on the BIWI Dataset

We also tested our framework on the public BIWI Kinect Dataset [123]. Although this dataset was originally created and used for head pose estimation in real time, it can be used for our purposes as well. The dataset consists of 24 sequences collected using the Microsoft Kinect sensor. There are 20 distinct test subjects consisting of 6 women and 14 men. This dataset is different from and more challenging than the VAP and RVL datasets in a number of aspects. First, in the other datasets, all human subjects looked at a fixed number of points on a wall sequentially, so that the images were recorded by the same Kinect sensor at roughly the same angles for each person. On the other hand, for the BIWI dataset, the test subjects sat in front of the sensor and moved their heads randomly in different directions in a continuous fashion while simultaneously changing facial expressions. Moreover, the calibration for the sensor can be different for different test subjects. The number of recorded images per test subject can vary between 395 to 946. For each frame, we are provided with the RGB data as a PNG image and the depth data as a binary file. Both of these have dimensions of 640 x 480 pixels. More details about this dataset can be found in [123]. Since our focus is face recognition, we use data for 20 unique test subjects.

Before testing our framework on this dataset, we need to first align the depth images and the RGB images. Each sequence is provided with the calibration information of the RGB sensor and the depth sensor. Please note that the calibration can be different for different sequences. This calibration information can be used to align the depth and RGB images. To be more precise, for each pixel in the depth image, we use the calibration information of the depth sensor to backproject the depth image to the 3D point and then use the calibration information of the RGB sensor to find the corresponding color values. After this step, we need to detect faces in the images. The BIWI dataset also contains mask images that indicate the location of faces in the image. For convenience, we use these masks to extract small windows around faces in our aligned RGB and depth images.

We use one frontal image per sequence for generating the 925 viewpoint variant training images. Some of them are shown in Figure 26. All the remaining images are used as test images. From Figure 26 we can observe that even with 2.5D interpolation, the quality of these training images is not as good as compared to our RVL dataset. Just as in the VAP dataset, there are holes in the viewpoint variant images. Moreover, a couple of human subjects were wearing spectacles and these created holes even in the frontal image. These artifacts can affect the accuracy of any classification paradigm. Better calibration, surface reconstruction and image alignment strategies are possible solutions to address these problems.

As with the VAP dataset, nonetheless, we were able to achieve high classification accuracies despite these difficulties, as described in detail below.

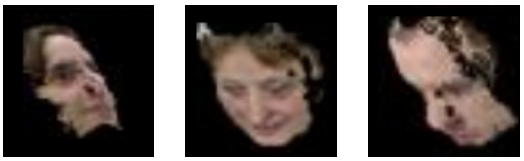

Figure 26: Examples of viewpoint variant training images for the BIWI dataset.

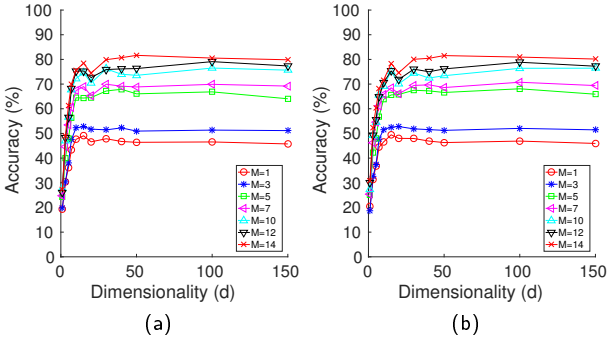

Figure 27: Multi-view classification accuracy versus subspace dimensionality for the Pose-CVS model with $K = 25$ partitions for the BIWI dataset. (a)Linear SVM (Pose-CVS-LSVM). (b) Nonlinear SVM (Pose-CVS-RKSVM).

5.4.1. Majority Voting

We first show results with the majority voting scheme. In Figure 27 we show classification accuracy versus dimensionality for the Pose-CVS model. We use $K = 25$ partitions. Similar to our observations for the RVL and VAP datasets, accuracy increases with more query images. Corresponding plots for the PSS models are shown in Figure 28.

Fixing the dimensionality d to 20, we compare the accuracies and time performances of the different models as a function of the number of query images in Figure 29. Again, the view-partitioned models perform better than the single subspace models and the PSS models outperform the CVS models. Note that the performance difference between App-PSS and Pose-PSS is more pronounced in this dataset than in the VAP dataset. This result is expected since the subjects moved their heads continuously and pose parameters were not available, making manual partitioning inevitably substantially less accurate.

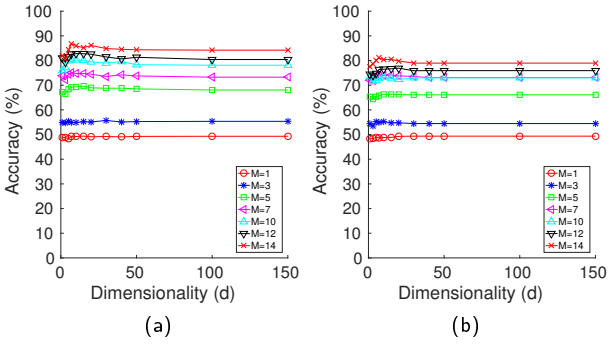

Figure 28: Multi-view classification accuracy versus subspace dimensionality for the PSS model with $K = 25$ partitions for the BIWI dataset. (a)Appearance based clustering (App-PSS). (b) Pose based clustering (Pose-PSS).

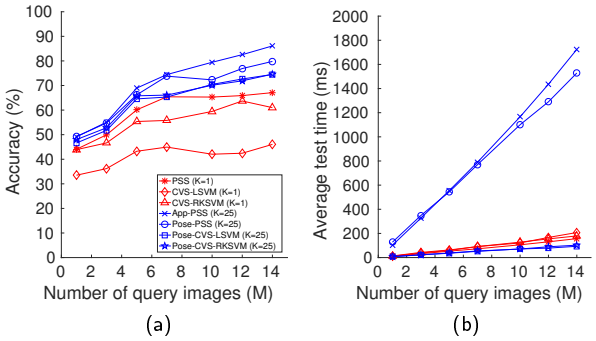

Figure 29: Comparison of multi-view classification approaches for the BIWI dataset when we use majority voting, as a function of the number of query images M with the dimensionality $d = 20$. The plots in red are for the case when single global subspaces are used and the plots in blue are for the case when view-partitioning is applied to the training data. The value of K is 1 for the red plots (since they correspond to the case of a single global subspace) and 25 for the blue plots. (a) Classification accuracies. (b) Time performance of the classifiers in (a).

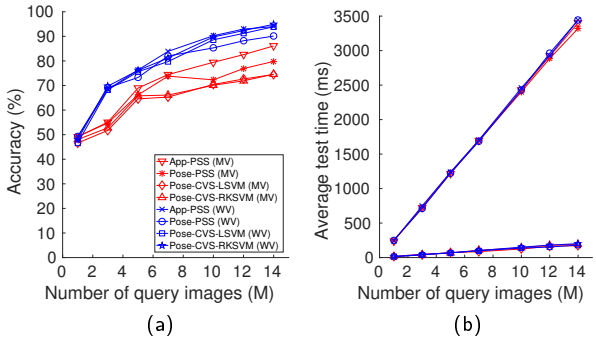

Figure 30: Comparison of weighted voting (by average normalized reconstruction error distance) with the majority voting approach for the view partitioned multi-view classification methods on the BIWI dataset with $d = 20$ and $K = 25$ in terms of (a) accuracy and (b) average test time. The plots in red are for majority voting and the plots in blue are for weighted voting.

5.4.2. *Weighted Voting*

Fixing dimensionality d as 20, we compare the majority and weighted voting schemes in Figure 30. As in the previous sections, the weighted voting scheme clearly outperforms the majority voting scheme. In this dataset, however, CVS models tend to benefit more from weighted voting than the view-partitioned approaches, possibly due to the different calibration parameters used for different subjects as well as the large variability in number of image frames for different subjects both of which would affect the quality of the reconstruction error metric.

5.5. *Comparison with Face Recognition based on a Convolutional Neural Network*

We compare our method with the state-of-the-art facial recognition approach proposed by Parkhi et al. in [124], which is based on the VGG deep convolutional network. This model was trained on face images of 2,622 different human subjects. For our experiments, we used the MatConvNet framework and a NVIDIA Tesla K20 GPU. The training and testing procedures are briefly described below.

5.5.1. *Training and Testing*

The viewpoint variant training images from all subjects (925 per subject) are normalized to zero mean and resized to be of size $224 \times 224 \times 3$ as per the requirements of the VGG net. In order to retrain the network for our purposes, we first remove the last two layers of the neural network. These correspond to the last fully connected layer (denoted as 'fc8' in the literature) and the final softmax layer. Since the original VGG net was trained on 2,622 classes, its 'fc8' layer produces 2,622 outputs. We replace this layer with a new 'fc8' layer that has as many outputs as the number of classes (10 for the RVL dataset, 28 for the VAP dataset, and 20 for the BIWI dataset). The weights of this new 'fc8' layer are randomly initialized. The

final softmax layer of the original VGG net is also replaced with a new softmax layer for the correct number of classes. We retrain the neural network using the gradient descent approach for 30 epochs. The 925 images per person are randomly split into training (90% of the images) and validation sets (10%). We use the trained neural net to classify the test images. Similar to the procedure used for our CVS and PSS approaches, we evaluate the performance of the deep learning approach by varying both dimensionality and number of query images. We use majority voting to combine labels from multiple views.

5.5.2. Results and Comparison

For the discussion in this section, we will denote the deep learning classifier as VGG-NET. We fixed $K = 25$ and dimension $d = 20$ for our approaches. In Figure 31 we compare the classification performance of VGG-NET with that of our framework. We observe that for all three datasets our PSS approaches, when used with the weighted voting strategy, outperform VGG-NET when the number of query images is larger than 7. It is interesting to note that the CVS approaches also outperform VGG-NET when used in conjunction with majority voting for the RVL and the BIWI datasets.

6. CONCLUSION

This paper answers the following question: "To what extent can face recognition be carried out using images from multiple arbitrary viewpoints if each human subject in a population is represented by a single frontal RGBD image?" No constraints are placed on the orientation of the camera vis-à-vis that of the face, except, of course, for the underlying assumption that a face can be seen with sufficient clarity from each viewpoint.

Towards answering the question stated above, this paper started out by first investigating the issue of how to generate multi-view training data from the individual RGBD images of the faces. Once the training data was available, we then dealt with how to best partition the multi-subject multi-view data for the construction of subspaces. Subsequently, we finally confronted our main research problem — multi-view recognition from images collected from a random selection of viewpoints. We compared global methods with view-partitioned methods, and, for each case, we experimented with common-view subspaces and person-specific subspaces. In the context of using view-partitioned subspaces, we also investigated the possibility of carrying out weighted voting in which the different query images are given different weights in the final classification depending on which subspace that query image was assigned to.

Three important conclusions were drawn. First, methods based on view-partitioned subspaces showed superior performance relative to global subspace methods. Second, person-specific subspaces, when used in a majority voting framework, are significantly more effective than common-view subspaces, although in most cases common-view subspaces also provided highly satisfactory results. Finally, weighted voting based on the normalized reconstruction error distance outperforms simple majority voting for multi-view classification. In particular, the App-PSS approach with weighted voting proved more flexible and robust than the other methods with a maximal accuracy of approximately 95% in all three datasets. The Pose-PSS approach with weighted voting performed only slightly worse in most cases, except for the BIWI dataset, in which the CVS methods benefited substantially from the weighted voting scheme. The App-PSS approach outperforms the state-of-the-art recognition method presented in [124] by as much as 7% when at least 7 views are available.

With regard to future directions, perhaps the most important goal would be to investigate the effect of noise and labeling errors when collecting 2D images of a face in a crowded environment. This paper made a simplifying assumption that all the query images on which the final decision is to be based belong to the same individual. That is highly unlikely to be the case in real life scenarios. Other issues that will certainly be present in a real-world application of our algorithms and hence need to be investigated in the future are the effect of query images of variable resolution caused by different distances between the targets and the cameras and the presence of motion blur in the images. At the moment it is not clear how a

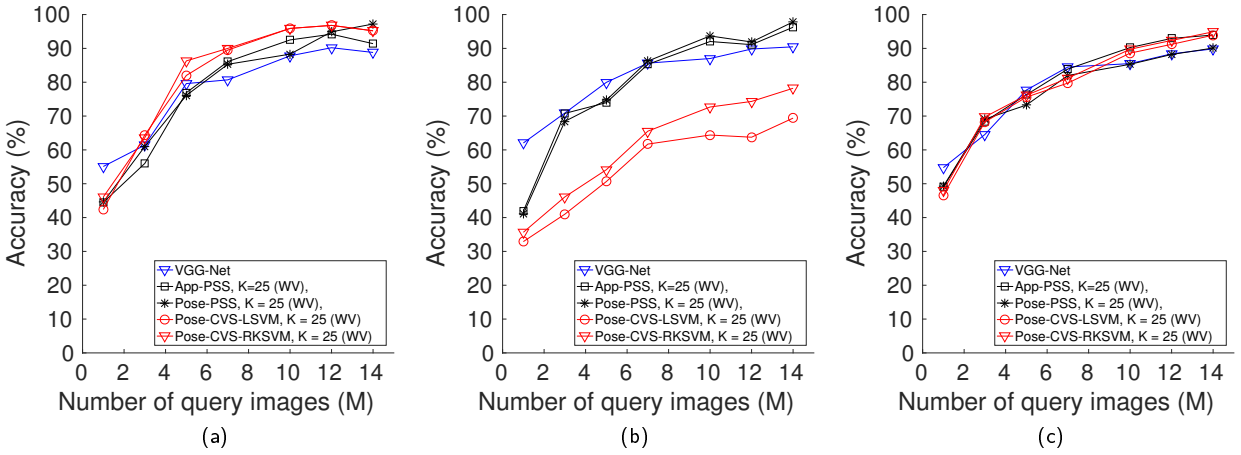

Figure 31: Comparison of our proposed approaches with the deep-learning based face recognition system presented in [124] for the a) RVL dataset, b) VAP dataset, and c) BIWI dataset.

large variability in photo resolution in the cameras or modest amounts of motion blur would affect the final classification outcome. Finally, another challenging issue for any facial recognition method are appearance modifiers such as facial hair and glasses. Since such modifiers can be seen as different kinds of partial occlusion, robust dimensionality reduction approaches such as IGO-PCA [125] which are specifically designed to handle these kinds of scenarios could be employed to alleviate this problem. Since our subspace construction methods necessarily involve a dimensionality reduction step, incorporating such robust algorithms should be relatively simple.

7. REFERENCES

[1] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino, "2D and 3D face recognition: A survey," *Pattern Recognition Letters*, vol. 28, no. 14, pp. 1885–1906, 2007.

[2] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *Acm Computing Surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[3] P. J. Phillips, P. Grother, and R. Micheals, *Evaluation methods in face recognition*. Springer, 2011.

[4] V. Krueger and S. Zhou, "Exemplar-based face recognition from video," in *European Conference on Computer Vision*. Springer, 2002, pp. 732–746.

[5] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2003, pp. 313–320.

[6] M. Du, A. Sankaranarayanan, and R. Chellappa, "Robust face recognition from multi-view videos," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1105–1117, March 2014.

[7] L. An, B. Bhanu, and S. Yang, "Face recognition in multi-camera surveillance videos," in *International Conference on Pattern Recognition*. IEEE, 2012, pp. 2885–2888.

[8] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report

07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.

[9] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, June 2014, pp. 1701–1708.

[10] C. Lu and X. Tang, "Surpassing human-level face verification performance on LFW with gaussianface," *CoRR*, vol. abs/1404.3840, 2014. [Online]. Available: http://arxiv.org/abs/1404.3840

[11] E. Zhou, Z. Cao, and Q. Yin, "Naive-deep face recognition: Touching the limit of LFW benchmark or not?" *CoRR*, vol. abs/1501.04690, 2015. [Online]. Available: http://arxiv.org/abs/1501.04690

[12] H. S. Seung and D. D. Lee, "The manifold ways of perception," *Science*, vol. 290, no. 5500, pp. 2268–2269, 2000.

[13] K. Okada and C. von der Malsburg, "Pose-invariant face recognition with parametric linear subspaces," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on.* IEEE, 2002, pp. 64–69.

[14] H. Wu and R. Souvenir, "Robust regression on image manifolds for ordered label denoising," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[15] D. Beymer, "Face recognition under varying pose," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, Jun 1994, pp. 756–761.

[16] D. Beymer and T. Poggio, "Face recognition from one example view," in *IEEE International Conference on Computer Vision*, 1995, pp. 500–507.

[17] M. Lando and S. Edelman, "Receptive field spaces and class-based generalization from a single view in face recognition," *Network: Computation in Neural Systems*, vol. 6, no. 4, pp. 551–576, 1995.

[18] T. Vetter and V. Blanz, "Estimating coloured 3D face models from single images: An example based approach," in *European Conference on Computer Vision.* Springer, 1998, pp. 499–513.

[19] W. Zhao and R. Chellappa, "SFS based view synthesis for robust face recognition," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on.* IEEE, 2000, pp. 285–292.

[20] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 6, pp. 643–660, 2001.

[21] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003.

[22] K. Niinuma, H. Han, and A. K. Jain, "Automatic multi-view face recognition via 3d model based pose regularization," in *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, Sept 2013, pp. 1–8.

[23] J. Choi, Y. Dumortier, S.-I. Choi, M. Ahmad, and G. Medioni, "Real-time 3-D face tracking and modeling from awebcam," in *Applications of Computer Vision (WACV), 2012 IEEE Workshop on*, Jan 2012, pp. 33–40.

[24] J. Yoder, H. Medeiros, J. Park, and A. Kak, "Cluster-based distributed face tracking in camera networks," *Image Processing, IEEE Transactions on*, vol. 19, no. 10, pp. 2551–2563, Oct 2010.

[25] J. Sung, T. Kanade, and D. Kim, "Pose robust face tracking by combining active appearance models and cylinder head models," *International Journal of Computer Vision*, vol. 80, no. 2, pp. 260–274, 2008.

[26] J. Alabort-i Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou, "Menpo: A comprehensive platform for parametric image alignment and visual deformable models," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 679–682. [Online]. Available: http://doi.acm.org/10.1145/2647868.2654890

[27] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[28] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, June 2013.

[29] G. Tzimiropoulos, "Project-out cascaded regression with an application to face alignment," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3659–3667.

[30] G. Tzimiropoulos and M. Pantic, "Optimization problems for fast aam fitting in-the-wild," in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 593–600.

[31] G. G. Chrysos, E. Antonakos, P. Snape, A. Asthana, and S. Zafeiriou, "A comprehensive performance evaluation of deformable face tracking "in-the-wild"," *CoRR*, vol. abs/1603.06015, 2016. [Online]. Available: http://arxiv.org/abs/1603.06015

[32] I. Marras, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Online learning and fusion of orientation appearance models for robust rigid object tracking," *Image and Vision Computing*, vol. 32, no. 10, pp. 707 – 727, 2014, best of Automatic Face and Gesture Recognition 2013.

[33] E. Hjelmås and B. K. Low, "Face detection: A survey," *Computer vision and image understanding*, vol. 83, no. 3, pp. 236–274, 2001.

[34] M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 1, pp. 34–58, 2002.

[35] C. Zhang and Z. Zhang, "A survey of recent advances in face detection," Tech. rep., Microsoft Research, Tech. Rep., 2010.

[36] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–511.

[37] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: Past, present and future," *Computer Vision and Image Understanding*, vol. 138, pp. 1 – 24, 2015.

[38] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[39] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[40] P. Campadelli, R. Lanzarotti, and C. Savazzi, "A feature-based face recognition system," in *Image Analysis and Processing, 2003.Proceedings. 12th International Conference on*, Sept 2003, pp. 68–73.

[41] O. Çeliktutan, S. Ulukaya, and B. Sankur, "A comparative study of face landmarking techniques," *EURASIP Journal on*

Image and Video Processing, vol. 2013, no. 1, pp. 1–27, 2013. [Online]. Available: http://dx.doi.org/10.1186/1687-5281-2013-13

[42] H. Yang, X. Jia, C. C. Loy, and P. Robinson, "An empirical study of recent face alignment methods," *CoRR*, vol. abs/1511.05049, 2015. [Online]. Available: http://arxiv.org/abs/1511.05049

[43] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[44] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[45] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic interpretation and coding of face images using flexible models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 743–756, Jul 1997.

[46] M. B. Stegmann and S. Olsen, "Object tracking using active appearance models," in *Proc. 10th Danish Conference on Pattern Recognition and Image Analysis*, 2001, pp. 54–60.

[47] M. B. Stegmann, B. K. Ersboll, and R. Larsen, "Fame-a flexible appearance modeling environment," *IEEE Transactions on Medical Imaging*, vol. 22, no. 10, pp. 1319–1331, Oct 2003.

[48] T. F. Cootes and C. J. Taylor, *Active Shape Models — 'Smart Snakes'*. Springer London, 1992, pp. 266–275.

[49] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models," in *Proc. BMVC*, 2006, pp. 95.1–95.10, doi:10.5244/C.20.95.

[50] S. Lucey, Y. Wang, M. Cox, S. Sridharan, and J. F. Cohn, "Efficient constrained local model fitting for non-rigid face alignment," *Image and Vision Computing*, vol. 27, no. 12, pp. 1804 – 1813, 2009, visual and multimodal analysis of human spontaneous behaviour:.

[51] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011. [Online]. Available: http://dx.doi.org/10.1007/s11263-010-0380-4

[52] C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes, "Robust and accurate shape model matching using random forest regression-voting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1862–1874, Sept 2015.

[53] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[54] D. Cristinacce and T. F. Cootes, "Boosted regression active shape models," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2007, pp. 79.1–79.10, doi:10.5244/C.21.79.

[55] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2930–2940, Dec 2013.

[56] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, *Interactive Facial Feature Localization*. Springer Berlin Heidelberg, 2012, pp. 679–692.

[57] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2013.

[58] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the

wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012, pp. 2879–2886.

[59] A. Crabtree, A. Chamberlain, M. Davies, K. Glover, S. Reeves, T. Rodden, P. Tolmie, and M. Jones, "Doing innovation in the wild," in *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, ser. CHItaly '13. New York, NY, USA: ACM, 2013, pp. 25:1–25:9. [Online]. Available: http://doi.acm.org/10.1145/2499149.2499150

[60] Y. Rogers, "Interaction design gone wild: Striving for wild theory," *interactions*, vol. 18, no. 4, pp. 58–62, Jul. 2011. [Online]. Available: http://doi.acm.org/10.1145/1978822.1978834

[61] G. Tzimiropoulos and M. Pantic, "Gauss-newton deformable part models for face alignment in-the-wild," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[62] A. Asthana, T. Marks, M. Jones, K. Tieu, and M. Rohith, "Fully automatic pose-invariant face recognition via 3d pose normalization," in *IEEE International Conference on Computer Vision*, 2011, pp. 937–944.

[63] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[64] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[65] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014. [Online]. Available: http://dx.doi.org/10.1007/s11263-013-0667-3

[66] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[67] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June 2010, pp. 2729–2736.

[68] S. Zhu, C. Li, C. Change Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[69] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, Dec 2013.

[70] J. Liu, Y. Li, P. K. Allen, and P. N. Belhumeur, "Articulated pose estimation using hierarchical exemplar-based models," *CoRR*, vol. abs/1512.04118, 2015. [Online]. Available: http://arxiv.org/abs/1512.04118

[71] A. J. Howell and H. Buxton, "Towards unconstrained face recognition from image sequences," in *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*. IEEE, 1996, pp. 224–229.

[72] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1994, pp. 84–91.

[73] G. Shakhnarovich, J. W. Fisher, and T. Darrell, "Face recognition from long-term observations," in *European Conference on Computer Vision*. Springer, 2002, pp. 851–865.

[74] X. Chai, S. Shan, X. Chen, and W. Gao, "Locally linear regression for pose-invariant face recognition," *Image Processing, IEEE Transactions on*, vol. 16, no. 7, pp. 1716–1725, 2007.

[75] S. Zhou, V. Krueger, and R. Chellappa, "Probabilistic recognition of human faces from video," *Computer Vision and Image Understanding*, vol. 91, no. 1, pp. 214–245, 2003.

[76] X. Liu and T. Chen, "Video-based face recognition using adaptive hidden markov models," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 1. IEEE, 2003, pp. I–340.

[77] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, "Face recognition with image sets using manifold density divergence," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 581–588.

[78] S. K. Zhou and R. Chellappa, "From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 6, pp. 917–929, 2006.

[79] W. Fan and D.-Y. Yeung, "Locally linear models on face appearance manifolds with application to dual-subspace based classification," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 1384–1390.

[80] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 1005–1018, 2007.

[81] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: a unifying view on subspace-based learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 376–383.

[82] O. Yamaguchi, K. Fukui, and K. Maeda, "Face recognition using temporal image sequence," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, 1998, pp. 318–323.

[83] B. Xie, T. Boult, V. Ramesh, and Y. Zhu, "Multi-camera face recognition by reliability-based selection," in *Computational Intelligence for Homeland Security and Personal Safety, Proceedings of the 2006 IEEE International Conference on*. IEEE, 2006, pp. 18–23.

[84] A. Pnevmatikakis and L. Polymenakos, "Far-field multi-camera video-to-video face recognition," *Face Recognition*, pp. 467–486, 2007.

[85] B. Xie, V. Ramesh, Y. Zhu, and T. Boult, "On channel reliability measure training for multi-camera face recognition," in *Applications of Computer Vision, 2007. WACV'07. IEEE Workshop on*. IEEE, 2007, pp. 41–41.

[86] S. Z. Li, X. Lu, X. Hou, X. Peng, and Q. Cheng, "Learning multiview face subspaces and facial pose estimation using independent component analysis," *IEEE Transactions on Image Processing*, vol. 14, no. 6, pp. 705–712, June 2005.

[87] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, *Multi-view Discriminant Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 808–821.

[88] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012, pp. 2160–2167.

[89] J. Tenenbaum, V. De Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[90] S. M. Zaki and H. Yin, *Multi-manifold Approach to Multi-view Face Recognition*. Springer International Publishing, 2015, pp. 370–377.

[91] J. Lu, Y. P. Tan, and G. Wang, "Discriminative multimanifold analysis for face recognition from a single training sample per person," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 39–51, Jan 2013.

[92] G. Goudelis, S. Zafeiriou, A. Tefas, and I. Pitas, "Class-specific kernel-discriminant analysis for face verification," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 570–587, Sept 2007.

[93] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, no. 4, pp. 270–286, 2014.

[94] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales, "The re-identification challenge," in *Person Re-Identification*. Springer, 2014, pp. 1–20.

[95] R. Satta, G. Fumera, and F. Roli, "Fast person re-identification based on dissimilarity representations," *Pattern Recognition Letters*, vol. 33, no. 14, pp. 1838–1848, 2012.

[96] S. Bąk, E. Corvee, F. Bremond, and M. Thonnat, "Boosted human re-identification using riemannian manifolds," *Image and Vision Computing*, vol. 30, no. 6, pp. 443–452, 2012.

[97] R. Mazzon, S. F. Tahir, and A. Cavallaro, "Person re-identification in crowd," *Pattern Recognition Letters*, vol. 33, no. 14, pp. 1828–1837, 2012.

[98] I. O. de Oliveira and J. L. de Souza Pio, "People reidentification in a camera network," in *Dependable, Autonomic and Secure Computing, 2009. DASC'09. Eighth IEEE International Conference on*. IEEE, 2009, pp. 461–466.

[99] Y. Cai, K. Huang, and T. Tan, "Human appearance matching across multiple non-overlapping cameras," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[100] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Multi-view perceptron: a deep model for learning face identity and view representations," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 217–225.

[101] J. Harguess, C. Hu, and J. Aggarwal, "Fusing face recognition from multiple cameras," in *Applications of Computer Vision (WACV), 2009 Workshop on*, 2009, pp. 1–7.

[102] D. Kim, J. Park, and A. C. Kak, "Estimating head pose with an RGBD sensor: a comparison of appearance-based and pose-based local subspace methods," in *IEEE International Conference on Image Processing*, 2013.

[103] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[104] K. Fukunaga and D. R. Olsen, "An algorithm for finding intrinsic dimensionality of data," *Computers, IEEE Transactions on*, vol. 100, no. 2, pp. 176–183, 1971.

[105] Z. Ghahramani, G. E. Hinton *et al.*, "The EM algorithm for mixtures of factor analyzers," Technical Report CRG-TR-96-1, University of Toronto, Tech. Rep., 1996.

[106] N. Kambhatla and T. Leen, "Dimension reduction by local principal component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1493–1516, 1997.

[107] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[108] L. K. Saul and S. T. Roweis, "Think globally, fit locally: unsupervised learning of low dimensional manifolds," *The Journal of Machine Learning Research*, vol. 4, pp. 119–155, 2003.

[109] J. Verbeek, "Learning nonlinear image manifolds by global alignment of local linear models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 8, pp. 1236–1250, 2006.

[110] D. Kim, "Pose and appearance based clustering of face images on manifolds and face recognition applications thereof," Ph.D. dissertation, Purdue University, 2015.

[111] Y. Hu and T. Huang, "Subspace learning for human head pose estimation," in *IEEE International Conference on Multimedia and Expo*, 2008, pp. 1585–1588.

[112] L. Morency, J. Whitehill, and J. Movellan, "Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation," in *IEEE International Conference on Automatic Face & Gesture Recognition*, 2008, pp. 1–8.

[113] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[114] K.-C. Lee and D. Kriegman, "Online learning of probabilistic appearance manifolds for video-based recognition and tracking," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 852–859.

[115] J. Luo, Y. Ma, E. Takikawa, S. Lao, M. Kawade, and B.-L. Lu, "Person-specific SIFT features for face recognition," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 2. IEEE, 2007, pp. II–593.

[116] J. Sivic, M. Everingham, and A. Zisserman, "'Who are you?' - learning person specific classifiers from video," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1145–1152.

[117] R. Wang, S. Shan, X. Chen, Q. Dai, and W. Gao, "Manifold–manifold distance and its application to face recognition with image sets," *IEEE Transactions on Image Processing*, vol. 21, no. 10, pp. 4466–4479, 2012.

[118] V. Vapnik, "Pattern recognition using generalized portrait method," *Automation and remote control*, vol. 24, pp. 774–780, 1963.

[119] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[120] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[121] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *Neural Networks, IEEE Transactions on*, vol. 13, no. 2, pp. 415–425, 2002.

[122] R. Høg, P. Jasek, C. Rofidal, K. Nasrollahi, and T. Moeslund, "An RGB-D database using Microsoft's Kinect for Windows for face detection," in *IEEE 8th International Conference on Signal Image Technology & Internet Based Systems*, 2012.

[123] G. Fanelli, J. Gall, and L. V. Gool, "Real time head pose estimation with random regression forests," in *Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 617–624.

[124] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[125] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Subspace learning from image gradient orientations," *IEEE Transactions on Pattern Anal-*

ysis and Machine Intelligence, vol. 34, no. 12, pp. 2454–2466, Dec 2012.