Computer Vision and Image Understanding 000 (2017) 1–19

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

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 C. Kak[a,*]

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

ARTICLE INFO

Article history:
Received 21 April 2016
Revised 18 December 2016
Accepted 16 April 2017
Available online xxx

Keywords:
Face recognition
Depth cameras
Manifold representations
Multi-view face recognition
RGBD models
Deep convolutional neural networks
Deep learning

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 the 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 face recognition approaches based on deep convolutional neural networks.

1. Introduction

Face recognition is now considered to be a reliable and non-intrusive 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 replicating such results in unconstrained scenarios by automating the face normalization step, 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 motion of the human subject is slow compared to the camera frame rate, and when high resolution cameras are employed. A solution to this general problem would be relevant in a number of applications, which include face verification and identification in static imagery (Abate et al., 2007; Phillips et al., 2011; Zhao et al., 2003), video (Krueger and Zhou, 2002; Lee et al., 2003), and with camera networks (An et al., 2012; Du et al., 2014).

The problem of recognizing faces under unconstrained conditions, also known as "face recognition in the wild", deals with assigning a face identity label to a set of face images collected by an assortment of cameras at random orientations with respect to the 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 face recognition from the images in a video or in other multi-view scenarios, there is no guarantee that any of the collected images would constitute a full frontal view of a face. In addition, we must also cope with other effects such as those caused 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, can multiple images from different viewpoints of the same face compensate for the lack of a single frontal image and controlled illumination? It is this

[*] Corresponding author. Tel.: +1 765 494 3551
E-mail addresses: zava@purdue.edu (D. Kim), bcomandu@purdue.edu (B. Comandur), henry.medeiros@marquette.edu (H. Medeiros), nelfiky@purdue.edu (N.M. Elfiky), kak@purdue.edu (A.C. Kak).

http://dx.doi.org/10.1016/j.cviu.2017.04.008
1077-3142/© 2017 Elsevier Inc. All rights reserved.

question that is the focus of this paper. If the reader accepts the validity of the question, the problem becomes one of how to pool together the visual evidence from the different viewpoints for classifying a face.

Some previous approaches have attempted to solve this problem by taking advantage of the machine learning algorithms made possible by the availability of large scale datasets of labeled faces in the wild (Huang et al., 2007; Lu and Tang, 2014; Taigman et al., 2014; Zhou et al., 2015). While these approaches have obtained impressive results, achieving accuracies as high as 99.5% and even surpassing the 97.53% accuracy obtained by human observers, they suffer from two main limitations. First, they rely on the existence of massive datasets for training purposes. While such datasets may be readily available for celebrities and other personalities, generating very large datasets for a broader population would be challenging. We are more interested, therefore, in a more practical scenario in which a 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 (Zhou et al., 2015), 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 for 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 for 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 these 2D images that are likely to reside on low-dimensional manifolds in high-dimensional measurement spaces (Okada and von der Malsburg, 2002; Seung and Lee, 2000; Wu and Souvenir, 2015); 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 frontal RGBD images of the human faces. We then view-partition the manifolds on which the data resides in order to identify the optimal subspaces in which groups of similar faces can be found together. We explore two different approaches for view-partitioning the training data, namely, pose based and appearance based.

Subsequently, we investigate how to best carry out multi-view classification by comparing view-partitioned approaches with global approaches. 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 separate person-specific global subspace for each human subject.

The view-partitioned approaches that we investigate create the possibility of carrying out weighted voting when combining the classification labels for a given set of test 2D images of the same face (as recorded from different viewpoints) into a single identity label. We do so by devising a weighting mechanism that uses the inverse of the normalized subspace reconstruction error for each test image as the weight that its classification label should carry in a multi-view aggregation of those labels.

This paper makes four main contributions. First, it presents a novel hierarchical approach for multi-view face recognition. Second, it proposes a weighted voting scheme for improved face recognition as obtained by combining the classification labels for the face images from different viewpoints. Third, it presents a new dataset of RGBD face images for the evaluation of multi-view face recognition algorithms. Finally, this paper includes an extensive evaluation and analysis of several approaches for carrying out data clustering and classification for the purpose of face recognition.

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

2. Prior work

Attempts at automatic recognition of faces using non-frontal imagery have generally involved constructing partial or full 3D models of the human head and then 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 (Beymer, 1994; Beymer and Poggio, 1995; Lando and Edelman, 1995). 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 (Blanz and Vetter, 2003; Georghiades et al., 2001; Niinuma et al., 2013; Vetter and Blanz, 1998; Zhao and Chellappa, 2000). 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 is morphed to fit 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 concurrently: a tracking/detection mechanism, a crucial alignment step, and a recognition algorithm, which generally attempts to exploit the availability of multiple image frames. Each of these three steps is complex and is an active subject of ongoing research (Choi et al., 2012; Hassner et al., 2015; Alabort-i Medina et al., 2014; Sagonas et al., 2013a; Sung et al., 2008; Tzimiropoulos, 2015; Tzimiropoulos and Pantic, 2013; Yoder et al., 2010). Regarding face tracking, a comprehensive survey of the existing approaches is presented in Chrysos et al. (2016). In a particularly relevant example, Marras et al. (2014) proposed a particle filtering method that uses the reconstruction error from learned subspaces to determine face orientation. As for face detection, although it is still a largely unsolved problem (especially if large variations in face poses are allowed), much progress has been made in this area during the last decade and a half (Hjelmås and Low, 2001; Viola and Jones, 2001; Yang et al., 2002; Zhang and Zhang, 2010). While face detection is generally regarded as the starting point for all face analysis tasks (Zafeiriou et al., 2015), face alignment is an essential intermediate step for many subsequent higher level tasks that range from biometric recognition to the interpretation of emotions. We discuss the issue of face alignment in more detail below.

2.1.1. Face alignment

The problem of face alignment is a well-studied topic in computer vision (Hassner et al., 2015; Matthews and Baker, 2004; Alabort-i Medina et al., 2014; Tzimiropoulos, 2015; Xiong and De la Torre, 2013). Face alignment is widely used by face recognition algorithms to improve their robustness against pose variations. Face recognition algorithms, such as those based on feature-based (structural) matching (Campadelli et al., 2003; Zhao et al., 2003), rely on accurate face alignment to establish correspondences for the local features (e.g. eyes, nose, mouth, etc.) used for matching.

Over the last two decades, numerous techniques have been developed for face alignment with varying degrees of success. Çeliktutan et al. (2013) have surveyed many traditional methods for face alignment for both 2D and 3D faces. For a more recent survey, see Yang et al. (2015). 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) that typically starts with a coarse initial shape and then proceeds by refining the shape estimate step by step until convergence. During the search process, two different sources of information are typically used: face appearance and face shape. Typically, faces are modeled as deformable objects that can vary in shape and appearance. Much of the early work along these lines was based on Active Shape Models (ASMs) and Active Appearance Models (AAMs) (Cootes et al., 2001; 1995; Tzimiropoulos and Pantic, 2013). In ASMs, face shape is expressed as a linear combination of shape bases learned through Principal Component Analysis (PCA), while appearance is modeled locally using (most commonly) discriminatively learned templates.

AAMs, first proposed by Cootes et al. (2001), are linear statistical models of both the shape and the appearance of a deformable object. Since AAM models can generate a variety of instances with only a small number of model parameters, they have been used widely in many computer vision tasks, such as face recognition (Lanitis et al., 1997), object tracking (Stegmann and Olsen, 2001), and medical image analysis (Stegmann et al., 2003). Despite their popularity and success, AAMs are generally considered to possess only limited representational power when used in unconstrained conditions. One possible way to overcome these drawbacks is to use part-based representations since local features are generally not as sensitive to lighting and occlusion as global features. ASMs are a notable example of part-based models (Cootes and Taylor, 1992; Cootes et al., 1995) that combine the generative appearance model for each face part with a Point Distribution Model for the global shape. More recently, the focus has shifted to a family of methods known as Constrained Local Models (CLMs) (Cristinacce and Cootes, 2006; Lucey et al., 2009; Saragih et al., 2011) that build upon ASM to model individual face parts using discriminatively trained local detectors (Asthana et al., 2013; Cristinacce and Cootes, 2007; Lindner et al., 2015; Saragih et al., 2011). In the training phase, a CLM learns an independent local detector for each face point and a prior shape model to characterize the deformation of the face shape. 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 (Belhumeur et al., 2013; Crabtree et al., 2013; Le et al., 2012; Rogers, 2011; Sagonas et al., 2013b; Zhu and Ramanan, 2012). These datasets have been used to develop a variety of cascaded regression-based techniques (Asthana et al., 2011; Cao et al., 2014; Kazemi and Sullivan, 2014; Ren et al., 2014; Sun et al., 2013; Tzimiropoulos and Pantic, 2014; Valstar et al., 2010; Xiong and De la Torre, 2013; Zhu et al., 2015) that have proved very successful in solving the face 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 (Liu et al., 2015; Yang and Ramanan, 2013), such methods are particularly successful when associated with generative deformable part models (Tzimiropoulos and Pantic, 2014). Despite the substantial progress made in face alignment in recent years, it is still unclear if the ability to determine the orientation of a face will translate into more accurate face recognition approaches for unconstrained scenarios in which face orientations may vary dramatically and frontal reconstructions are likely to be heavily distorted.

2.1.2. Exploiting multiple image frames for video-based face recognition

Rather than attempting to carry out face frontalization on each frame, most video-based face recognition approaches try to leverage the availability of multiple images of the same face in a video, often at different poses and under different illumination conditions.

Taking advantage of the presence of multiple frames showing the same face in a video does, however, come with its own challenges — overcoming the problems caused by sudden pose and illumination changes. A well-known approach to solving these problems consists of recording training videos of the human subjects in arbitrary motions and, subsequently, using the frames of the training videos as a gallery of images for each subject in the database. At test time, a query video recording is compared with all of the gallery images in the database. Generally, each frame of the query video is compared with all the gallery images for estimating a matching score for the query video (Chai et al., 2007; Howell and Buxton, 1996; Pentland et al., 1994; Shakhnarovich et al., 2002).

Instead of performing face recognition with a frame-by-frame comparison of the training and the 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. (2003) pioneered this kind of work by tracking human subjects in videos and extracting their faces to construct priors for the different views of the different faces. Lee et al. (2003) focused on automatically learning the transition probabilities between the different possible appearances of a face in a video. Along the same lines, Liu and Chen (2003) 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, known as the ensemble approach, 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 (Arandjelovic et al., 2005; Fan and Yeung, 2006; Hamm and Lee, 2008; Kim et al., 2007; Shakhnarovich et al., 2002; Yamaguchi et al., 1998; Zhou and Chellappa, 2006). In order to deal with this problem 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 frames. Subsequently, face recognition is based on comparing the query video with the virtual gallery videos for the different human subjects.

Please cite this article as: D. Kim et al., Multi-view face recognition from single RGBD models of the faces, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.04.008

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 problems because, with video, the variations in the viewpoints are bound to be localized to where the camera happens to be with respect to the human subject. On the other hand, when you have multiple cameras viewing the same subject, the cameras could be mounted at spatially dispersed locations that make for large variations in viewpoints vis-a-vis the subject. One typical example would be the cameras in an airport terminal 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 (Pnevmatikakis and Polymenakos, 2007; Xie et al., 2006; 2007). In Xie et al. (2007), for example, the reliability of each camera depends on how well both face detection and recognition can be carried out with the image captured by that camera. Alternatively, subspace learning methods can be used to compare the pose of a face as seen in a camera view, with the pose information needed for aggregating the multi-camera data. Li et al. (2005) proposed one of the first approaches for clustering faces into subspaces according to their poses. Their method is 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), their results are largely qualitative. Kan et al. (2012) 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 the one presented by Sharma et al. (2012), but their multi-view classification evaluation is restricted to the viewpoint range $[-45^\circ, +45^\circ]$ in azimuth. Note also that manifold learning approaches have been shown to be more robust and have better generalization capabilities than linear methods such as ISA (Lu et al., 2013; Tenenbaum et al., 2000; Zaki and Yin, 2015). Furthermore, none of the methods mentioned above is concerned with the problem of incorporating multiple query images recorded from different viewpoints in a classification algorithm.

While not directly applicable to the multi-camera face recognition problem, another related non-linear subspace learning approach was proposed by Goudelis et al. (2007). In that work, the authors proposed a face verification (i.e., binary recognition) method that employs a kernelized discriminant for maximizing the impostor distance measures while minimizing the client (i.e., non-impostor) distance measure. Their method showed impressive single-digit equal error rates (EER) for several challenging datasets with varying face poses.

When there is no overlap between the fields of view of the cameras involved, person re-identification becomes a fundamental issue in multi-camera face recognition (Bąk et al., 2012; Bedagkar-Gala and Shah, 2014; Cai et al., 2008; Gong et al., 2014; Mazzon et al., 2012; de Oliveira and de Souza Pio, 2009; Satta et al., 2012; Zhu et al., 2014). 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 evidence aggregation is the main problem addressed in this paper.

Regarding previous contributions to such evidence aggregation, An et al. (2012) aggregate the information from different cameras with the help of a dynamic Bayesian network that contains a node for each camera and a node for each human subject that 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. (2014) 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 that is 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. (2009). In this approach, 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 from a set of randomly chosen viewpoints around the face. 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 is now well known, face data collected from different viewpoints is likely to reside on a manifold 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 best represent the training data for the different human subjects in a manifold-based 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 and the second is based on the similarity of appearance between the training images. We have previously used both of these approaches for solving the simpler problem of head pose estimation (Kim et al., 2013). 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

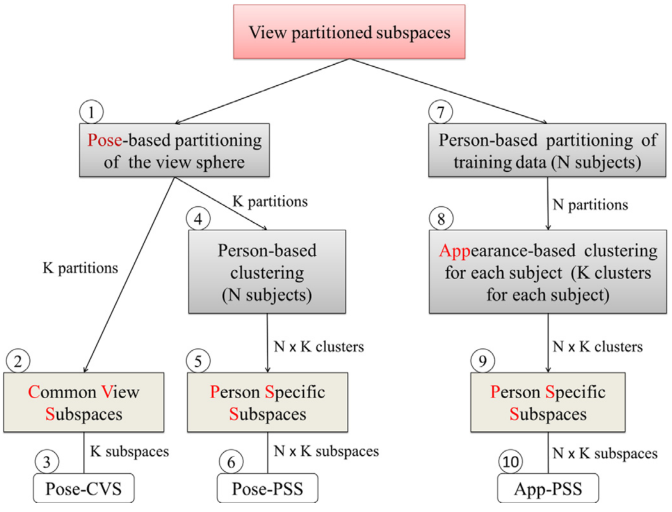

View partitioned subspaces
1
Pose-based partitioning of the view sphere
7
Person-based partitioning of training data (N subjects)
K partitions
N partitions
4
Person-based clustering (N subjects)
8
Appearance-based clustering for each subject (K clusters for each subject)
K partitions
N x K clusters
N x K clusters
2
Common View Subspaces
5
Person Specific Subspaces
9
Person Specific Subspaces
3
K subspaces
6
N x K subspaces
10
N x K subspaces
Pose-CVS
Pose-PSS
App-PSS

Fig. 1. Variations on the classifiers for face recognition with view-partitioned subspaces for multi-subject and multi-view face recognition scenarios.

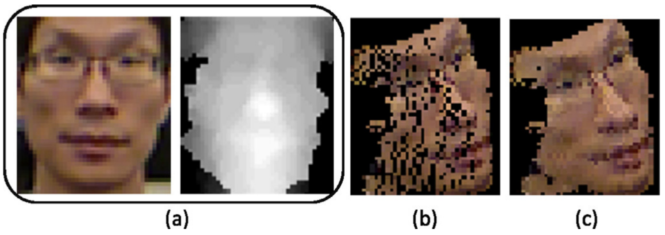


Fig. 2. The sequence of steps for generating a pose-transformed version of a frontal RGBD image: (a) the original RGBD image for the frontal pose (the RGB data is shown on the left and the depth is shown on the right); (b) the pose transformed and projected result from the data in (a); (c) the 2.5D interpolated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

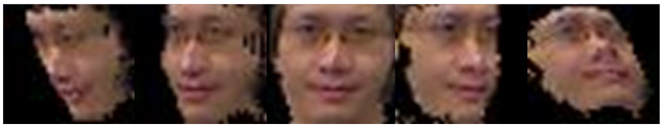


Fig. 3. Examples of the generated training images for one subject.

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. Fig. 1 illustrates these variations on the pose-based and the appearance-based subspace construction techniques. We will discuss each of the boxes in Fig. 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 collecting 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 frontal 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.

3.1. Creating viewpoint-variant face images from a single frontal RGBD scan of a human subject

As we have previously described in Kim et al. (2013), the 3D position (*X, Y, Z*) associated with an RGBD “pixel” at the raster coordinates (*x, y*) is given by:

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

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 (Otsu, 1975). 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 (Fig. 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 $\mathbf{x}_{2D}$ represents the *(x,y) pixel positions,* $\mathbf{Z}_D$ the depth values, $\mathbf{X}_{3D}$ the three spatial coordinates of the corresponding object points, and $\mathbf{V}_{RGB}$ the three color values recorded at the pixels. This cloud model includes the 2D pixel coordinates, the 3D coordinates of the corresponding object point, as well as the texture data in the form of RGB values at the object point. 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 into the camera image plane. The computation that generates a virtual view image I_t is described by

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

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, and $G(\cdot)$ is the 3D transformation involving 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 shown below:

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

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.

Generating 2D images projected from rotated 3D points has two general problems to be considered. First, after a pose transform is applied, it is possible for multiple 3D points in the point cloud to project to the same pixel in the camera image plane. To get around this problem, only the closest sample to the camera is projected into the camera image plane. Second, when a pose-transformed point cloud is projected into the camera image plane, one can end up with “holes” in the projected image on account of the variable depth resolution of the RGBD sensor. An example of this effect is shown in Fig. 2(b), which illustrates a pose-transformed version of the frontal RGBD image in Fig. 2(a). We eliminate such holes by applying bilinear interpolation to the neighboring points in the image plane using the constraint that the points used for bilinear interpolation possess roughly the same depth values. Fig. 2 (c) shows a projection when such interpolation is a part of the projection operator. Fig. 3 shows additional examples of the training images generated according to this process.

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 (Okada and von der Malsburg, 2002; Seung and Lee, 2000; Wu and Souvenir, 2015). This fact is responsible for much interest in topics such as manifold-based learning and data clustering (Fukunaga and Olsen, 1971; Ghahramani et al., 1996; Kambhatla and Leen, 1997; Roweis and Saul, 2000; Saul and Roweis, 2003; Tenenbaum et al., 2000; Verbeek, 2006). Much of this work is based on the intuition that if we could first create

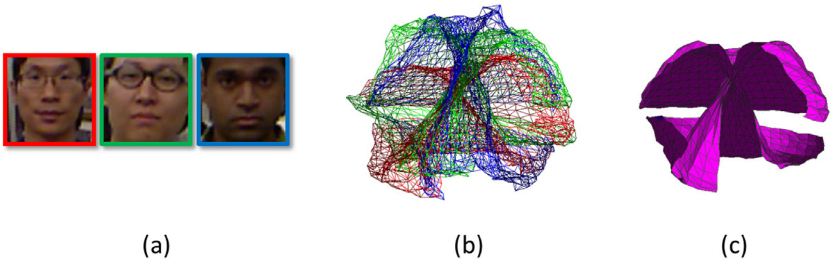


Fig. 4. Visualization of the manifolds corresponding to three subjects as obtained by ISOMAP: (a) Three subjects, (b) Visualization of person-specific manifold structure in the PCA space, (c) Mean manifold for the person-specific manifolds in (b).

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 that exist today for understanding the data on manifolds, namely: 1) Locally Linear Embedding (LLE) (Roweis and Saul, 2000); 2) ISOMAP (Tenenbaum et al., 2000); and 3) Representations that can be obtained by the Kambhatla and Leen algorithm (Kambhatla and Leen, 1997). Our study concluded that ISOMAP gives us the best partitioning of the data that minimizes the average reconstruction error in the subspaces in each of the view partitions of the data (Kim, 2015). 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 a set of appropriate projection transforms to the scan. The clustering results we show in this section are obtained on the image data collected in this manner. These results are based on the training images collected from the RGBD scans for the three subjects shown in Fig. 4(a).

The manifold structure shown in Fig. 4(b) for each of the three subjects in Fig. 4(a) is in the space spanned by the three leading eigenvectors when all of the data for all three subjects is subject to a PCA based dimensionality 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 Fig. 4(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 Fig. 4(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 Fig. 4(a)–(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 using ISOMAP, the extent to which the algorithm can capture both the global shape variations in the manifolds shown in Fig. 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. Fig. 5(a)–(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 Fig. 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 Fig. 5 (c). This implies that small values of γ are to be preferred since the class discriminatory information between the different human subjects is likely to reside primarily in the local variations on the manifold. When we apply the KMeans algorithm to the ISOMAP representation with clusters$K = 9$, the corresponding appearance-based clustering results that we get are as shown in Fig. 5 for the three different values of γ.

We now show that appearance-based clustering of the multi-subject data represented by the results in Fig. 5 does NOT yield a usable partitioning of the view sphere. Shown in Fig. 6 is a random sampling of the images in each of the clusters in Fig. 5. What is even more important with regard to the results shown in Fig. 6 are the triple of data entries, with each entry of the form $\texttt{SI} : \texttt{X}$ where $\texttt{SI}$ is one of $\{\texttt{S1}, \texttt{S2}, \texttt{S3}\}$ and where $\texttt{X}$ 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 Fig. 4(a). The integer $\texttt{X}$ in $\texttt{SI} : \texttt{X}$ stands for the number of images for the subject $\texttt{SI}$ in the cluster. Given this notation, out of 9 clusters, we have 5 clusters that consist exclusively of images for 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 partitioning we achieve automatically with such clustering does not correspond to an even distribution of the different face poses for the three subjects. As shown in Kim (2015), however, this algorithm does typically give us good viewspace partitioning of the images *as long as they belong to a single human subject*. We will take advantage of this fact later in this paper when we consider person-specific appearance-based partitioning of the manifold data for constructing a set of locally optimal subspaces for each subject.

What works best for the case when *multi-subject images are considered together* is pose based partitioning of the viewspace , which is accomplished trivially since the training images generated from the RGBD data are tagged with the face poses. Fig. 7(a) illustrates nine partitions that are manually delineated in the pitch and yaw space. Shown in Fig. 7(b) is a visualization of all the images in the space spanned by the three leading eigenvectors extracted with PCA from all of the images. Fig. 7(c) shows the partitioning applied to the mean of the images of the three human subjects, the means being computed for the same pose parameters.

In the next subsection, we discuss locally optimum subspace construction for the individual clusters in the data on the manifold. However, before launching into the material presented next , we must first point out that there has been much research in the past in fitting locally linear subspaces to data that resides on nonlinear manifolds (Hu and Huang, 2008; Lee et al., 2003; Morency et al., 2008; Pentland et al., 1994).

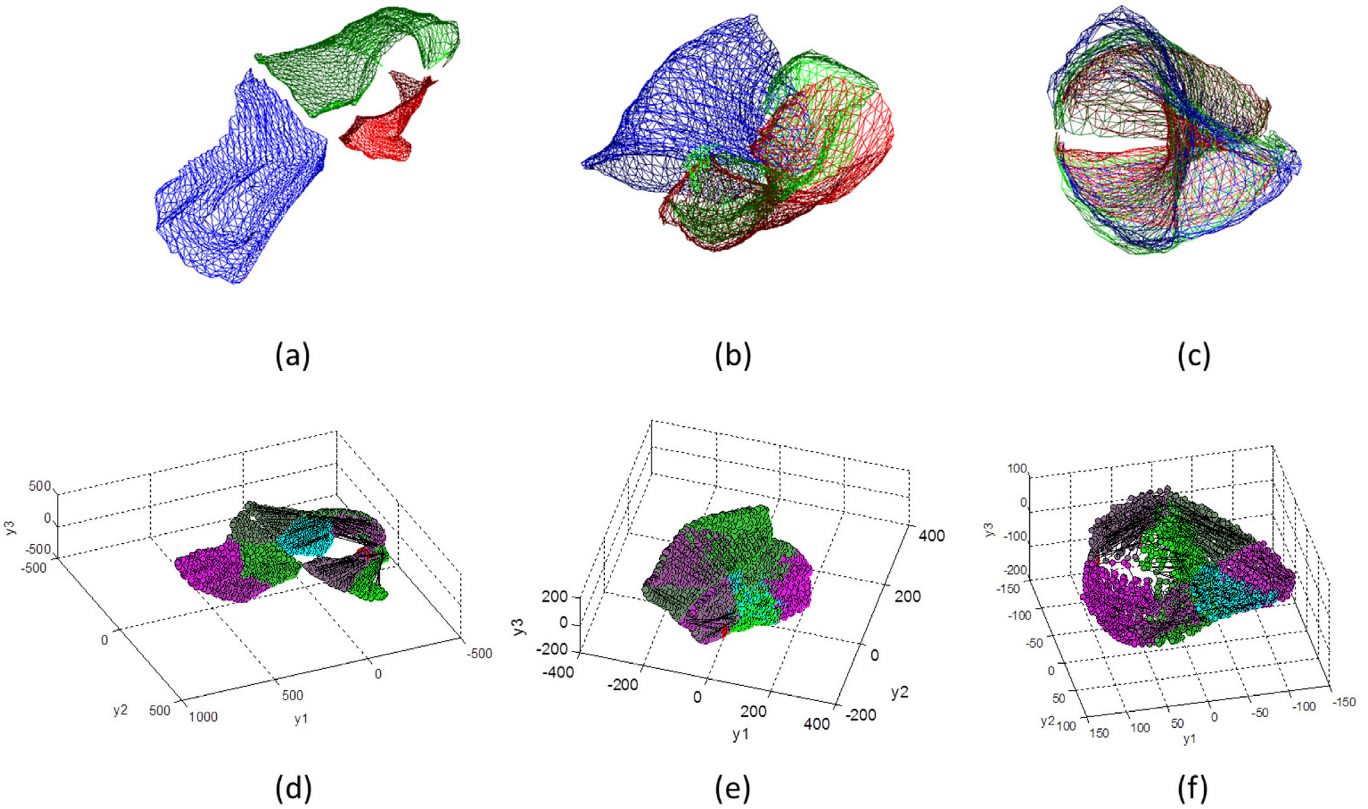


Fig. 5. Top row: ISOMAP-based representation of multi-subject face images with (a) $\gamma = 6$, (b) $\gamma = 10$, (c) $\gamma = 27$. Bottom row: Clustering results using KMeans applied to the ISOMAP representation with (d) $\gamma = 6$, (e) $\gamma = 10$, and (f) $\gamma = 27$. The parameter γ controls the size of the immediate neighborhood of a data point that ISOMAP uses for calculating point-to-point geodesic distances.

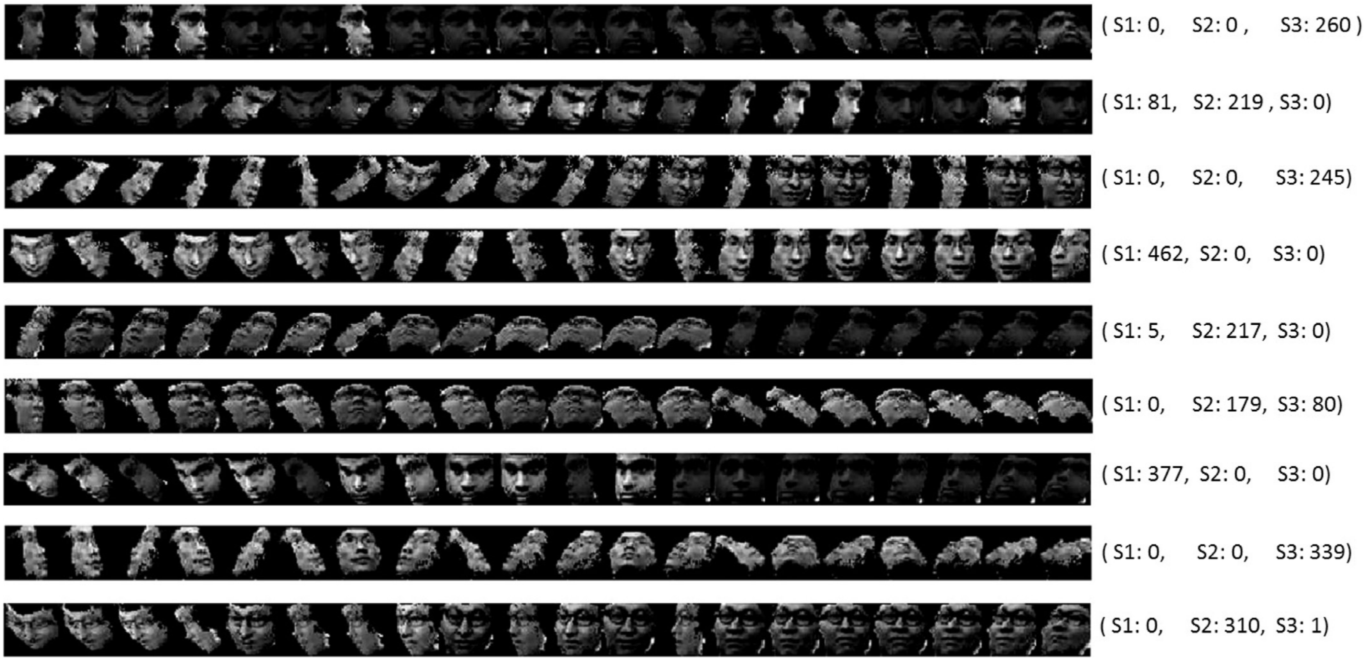


Fig. 6. Clustered image samples that correspond to the result shown in Fig. 5 (d) with $K = 9$ and $\gamma = 6$ for the three subjects in Fig. 4(a).

3.3. Constructing subspaces from view-partioned clusters

Before we can construct optimal subspaces for the individual clusters on the manifold, 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 Fig. 1, or through a finer person-specific decomposition as at nodes 5 and 9. For the common-view case, each pose-partitioned subspace contains training data from all the subjects. For the case of person-specific

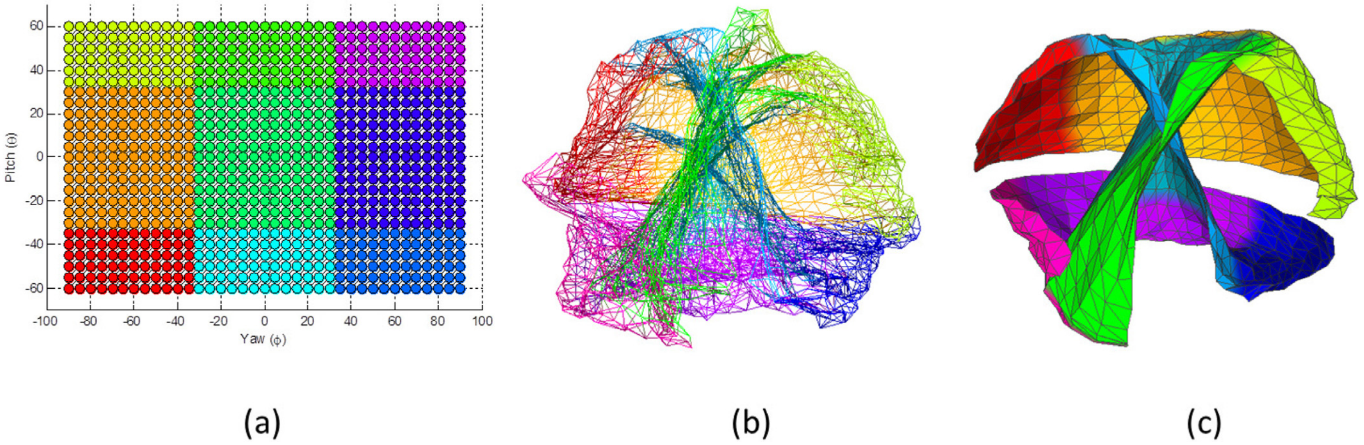


Fig. 7. Visualization of pose-based clustering for $K = 9$ the three subjects shown in Fig. 4(a): (a) Manual pose partition in the pitch and yaw space, (b) Partitioned subject-specific manifolds in the PCA space, (c) A partitioned mean manifold in the PCA space.

subspaces, the pose-partitioned subspaces are made specific to each individual subject. In addition to pose partitioning, we consider appearance-based partitioning for the case of person-specific subspaces.

Recent literature in face recognition suggests that we are likely to achieve higher recognition accuracies if we construct person-specific subspaces (Belhumeur et al., 1997; Lee et al., 2003; Lee and Kriegman, 2005; Luo et al., 2007; Sivic et al., 2009; Wang et al., 2012). The reason has to do with the fact that the fine details on the manifold structure for each individual subject are likely to get 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 corresponding to each human subject 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 in our investigation is for the pose-based partitioning criterion as shown at node 3 of Fig. 1 (as demonstrated by the clustering results shown in Fig. 6 in Section 3.2, partitioning the viewspace for the global case based on subject appearance does not provide a useful representation). We call this model Pose-CVS; it is created by first pose-partitioning the view sphere and then placing the relevant 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 the 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 and the total number of human subjects H (elsewhere in this paper, especially in Fig. 1, we have used the symbol N for the total number of human subjects in the training data), the CVS model is represented by

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

$$= \{L_{cvs,h}\}_{h=1}^{H}, \tag{5}$$

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

Again, as shown in Fig. 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 in each pose partition separately. We call this approach Pose-PSS. On the other hand, at 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 KMeans clustering (see Kim, 2015 for more details). This approach is called App-PSS. Consequently, both PSS models can be expressed in the following form:

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

$$= \left\{ L_{pss,h} \right\}_{h=1}^{H}, \tag{7}$$

where K is the number of clusters formed for each human subject (based on pose or appearance), and H denotes the number of human subjects (recall from the earlier note in this section that the symbol N is synonymous with the symbol H in this paper). The kth cluster-based subspace for the hth subject is represented by $S_h^{(k)} =< \mathbf{r}_h^{(k)}, \mathbf{U}_h^{(k)}, \mathbf{\Lambda}_h^{(k)} >$ where $\mathbf{r}_h^{(k)}$ is the center of the kth cluster, $\mathbf{U}_h^{(k)}$ the eigenvector matrix, and the eigenvalues matrix$\mathbf{\Lambda}_h^{(k)}$. Also, $\mathbf{Y}_h^{(k)}$ is the set of points for the hth human subject 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 calculates the orthogonal distance from a query to the hyperplane obtained by PCA. Fig. 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

Please cite this article as: D. Kim et al., Multi-view face recognition from single RGBD models of the faces, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.04.008

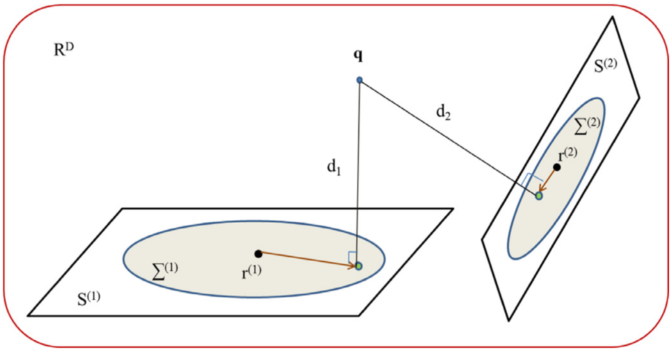


Fig. 8. Geometric interpretation of the reconstruction error distance for two subspaces $S^{(1)}$and $S^{(2)}$ in R^D.

distance is given by:

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

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

Referring back to Fig. (1), App-PSS and Pose-PSS return the face identity using only the NS classifier. On the other hand, Pose-CVS requires two-layered classifiers in order to determine the face label. First we must select a subspace and then figure out the applicable face label in that subspace. We consider two different classifiers in the second layer of Pose-CVS: a nearest neighbor (NN) classifier and an SVM classifier (Cortes and Vapnik, 1995; Vapnik, 1963).

Fig. 9 illustrates the classification logic for each of the models we consider. Fig. 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

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

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

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

$$j = argmin_k \, d\left(\mathbf{q}, S^{(k)}\right), \qquad \text{k=1,} \cdots \text{,K.} \tag{10}$$

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

For the NN classifier, let the training samples $\mathbf{x}_i^{(j)}$ in the jth subspace have their local-subspace representations given by the vectors $\mathbf{y}_i^{(j)} = \mathbf{F}^{(j)T}(\mathbf{x}_i^{(j)} - \mathbf{r}^{(j)})$ 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 jth 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

$$l^* = argmin_i \, \|\mathbf{y}_i^{(j)} - \mathbf{F}^{(j)T}\left(\mathbf{q}-\mathbf{r}^{(j)}\right)\|^2, \qquad i = 1, \cdots, T_j. \tag{11}$$

The person label returned for the query image $\mathbf{q}$ is the label h associated with the nearest training sample 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 with the local-subspace representation of the training samples $\mathbf{x}_i^{(j)}$ in the jth 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 $\mathbb{K}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$ and a nonlinear kernel with radial basis function, $\mathbb{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$. In this paper, we utilize the multi-class SVM in Chang and Lin's LibSVM (Chang and Lin, 2011), which is based on the one-against-one approach in which we have one SVM for each pair of classes (Hsu and Lin, 2002).

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 is the method we use in this section for the global approaches.

On the other hand, 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 carry greater discriminatory power than the other viewpoints. That then provides us with motivation for investigating a weighted voting approach to combine identity labels generated by view-partitioned subspaces.

For the case of view-partitioned subspaces, Fig. 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 N different subjects in the population on which the system is trained. We assume we have access to a single 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 as would be seen 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 Fig. 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 common view-partitioned subspaces where each subspace represents the data from all of the subjects ($Model_{CVS}$) or the person-specific 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 ith 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.

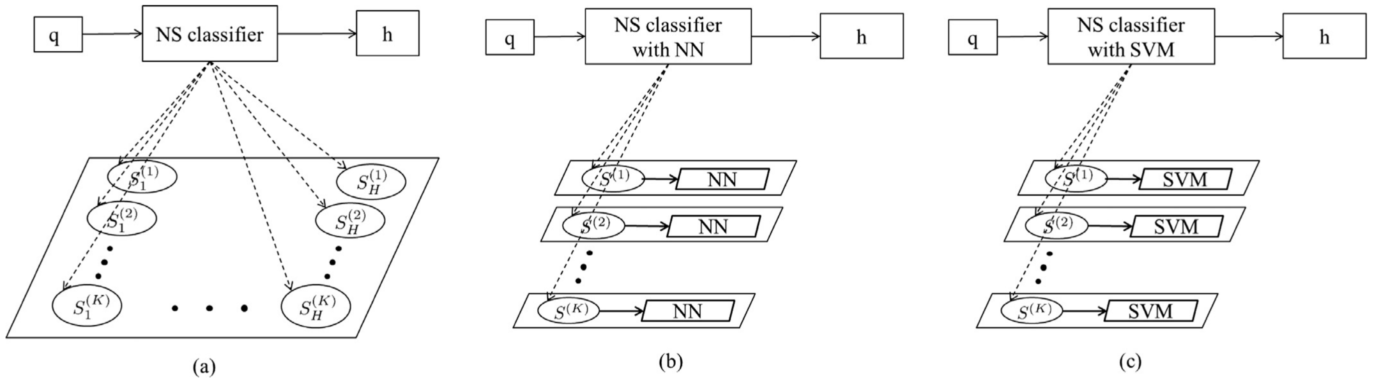


Fig. 9. Classification logic for: (a) App-PSS and Pose-PSS, (b) Pose-CVS-NN, (c) Pose-CVS-LSVM and Pose-CVS-RKSVM. See Fig. 1 for what is meant by App-PSS, Pose-PSS, and Pose-CVS. The additional qualifiers used with Pose-CVS stand for the second-layer classification strategy used. The symbol H in the figure stands for the total number of human subjects in the training data (which is also represented by N in this paper). The symbol K stands for the total number of partitioned subspaces for CVS and for the total number of partitioned subspaces *per person* for PSS.

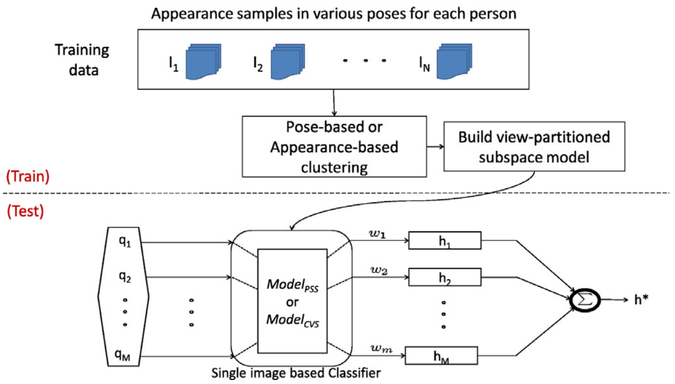

Appearance samples in various poses for each person
Training data
I_1
I_2
I_N
Pose-based or Appearance-based clustering
Build view-partitioned subspace model
(Train)
(Test)
q_1
q_2
q_M
$Model_{PSS}$ or $Model_{CVS}$
w_1
w_2
w_m
h_1
h_2
h_M
Σ
h*
Single image based Classifier

Fig. 10. A weighted voting framework for multi-view inputs.

4.1. Weighted voting by normalized reconstruction error distance

For 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 as we explain below.[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 ith query $\mathbf{q}_i$ is obtained by

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

where $d(\mathbf{q}, S_h^{(k)})$ denotes the reconstruction error distance of $\mathbf{q}$ to the kth subspace of the hth person given by Eq. (8) (see Appendix B in Kim, 2015 for more details). Similarly, for the CVS model, the minimum reconstruction error distance for a query $\mathbf{q}$ is obtained by

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

Then, the normalized minimum distance is given by

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

for PSS, and

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

for CVS. In Eqs. (14) and (15), the symbol H stands for the total number of human subjects in the training data. (Recall, from Section 3.3.1, this paper uses the symbols N and H synonymously.) The weight for the ith 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 , respectively$\varepsilon_1, \varepsilon_2, \ldots$. 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

Our discussion so far has raised a number of important research questions that we now address with an extensive experimental evaluation. In particular, we focus on the following research questions: 1) Does viewspace partitioning improve the performance of a classifier in comparison with that of the global approaches? 2) What is the effect of the number of such partitions on the classification performance? 3) Should viewspace partitioning be carried out on the basis of face pose or face appearance? 4) What is the impact of the dimensionality of the subspaces that represent the different partitions on the performance of the system? 5) Does a classification system benefit from aggregating multiple images from different viewpoints? And if so 6) Does the proposed weighted voting method further improve the system performance in comparison with simple majority voting?

In order to quantitatively assess the relative merits of the different classification strategies, we use three RGBD datasets. The first is the RVL face dataset consisting of 10 human subjects that we created. Some example images from the RVL dataset are shown in Figs. 11 and 13. The second is a public dataset consisting of 28 subjects from the Visual Analysis of People (VAP) lab at Aalborg

[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.

Please cite this article as: D. Kim et al., Multi-view face recognition from single RGBD models of the faces, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.04.008

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

University (Høg et al., 2012). 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 (Fanelli et al., 2011), another publicly available dataset consisting of 20 distinct test subjects. Finally, we also compare our methods with the state-of-the-art face classification approach proposed by Parkhi et al. (2015).

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 the 10 human subjects from a single frontal RGBD image for each subject in the RVL dataset, as described in Section 3.1. Fig. 11 displays the frontal images of the 10 subjects. For 10-fold cross validation, we randomly shuffle the 200 images for each human subject. For each run of the 10-fold test, we use 180 of these for training and the remaining 20 for testing. (Compared to the 200 multi-view images generated from the RGBD model for each subject in this section, we generate a much larger number of views — 925 — for the training required for multi-view face recognition as reported in the next section.)

During training, we generate three models: Pose-CVS, Pose-PSS, and App-PSS (see Fig. 1). As previously mentioned, the Pose-CVS model has three variants, Pose-CVS-NN, Pose-CVS-LSVM and Pose-CVS-RKSVM, each with a different second-layer classifier. For details regarding the classification logic in each model, see Fig. 9 and the associated explanations. We investigate the discriminatory power of the subspaces in each model as we vary the number of clusters (which is the same as the number of subspaces) K and the dimensionality of the subspaces d. The baseline method is the SVM classifier with the RBF kernel and no viewspace partitioning because this type of classifier has had a rich history of success in the past.

Fig. 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 viewspace partitioning. As the reader can see, the PSS model is not only better than the linear SVM and the global NN models but is also comparable to the nonlinear SVM model. For the Pose-CVS model, when we use the NN classifier, its performance approaches that of 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 what is shown in the figure in order to do so). In terms of the number of view-partitions, as Fig. 12(b) – (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 Fig. 12 (c), for example, requires a subspace dimensionality of around 200 (not shown in the figure) to approach its saturated accuracy with $K = 1$, but for $K = 9$, it surpasses the global linear SVM with only $d = 20$ dimensions.

5.2. Results on the RVL face dataset

Starting with this subsection, we present our multi-view face recognition results in this and the next two subsections. As stated earlier in Section 5, our overall evaluation of the classification framework presented in this paper is based on three different

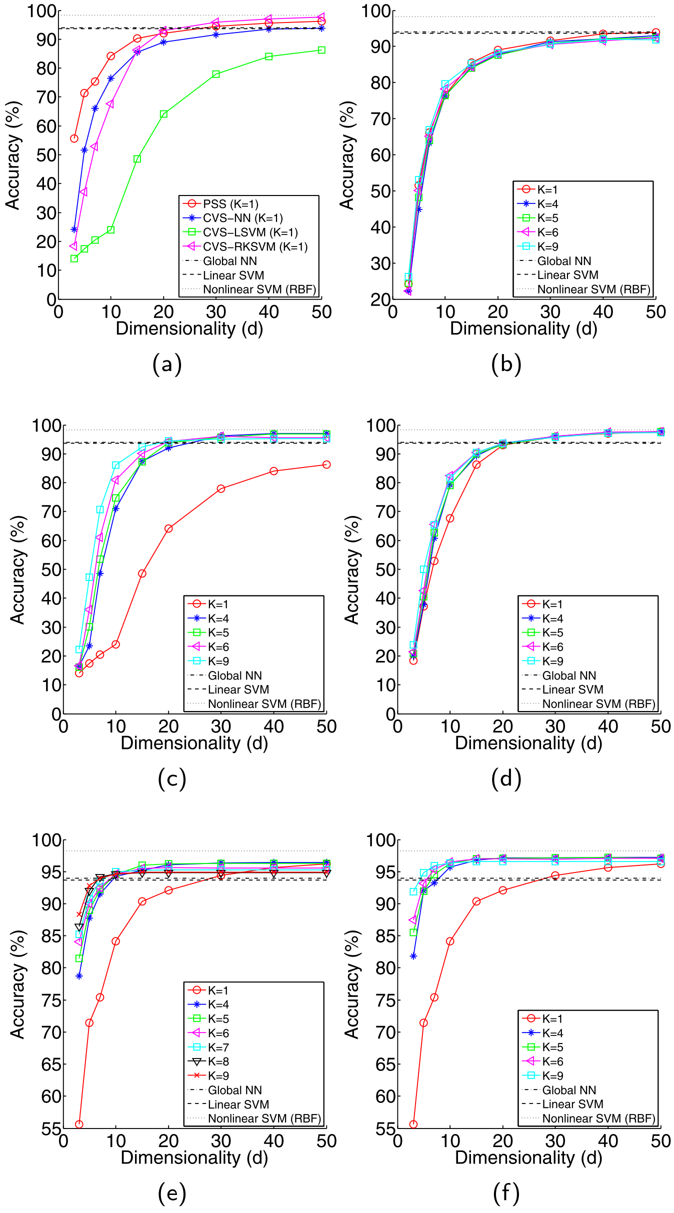


Fig. 12. We investigate the extent to which the different subspace models retain the class discriminatory information by measuring the accuracy with which individual image samples are classified in a 10-fold cross-validation test. This figure shows accuracy vs. subspace dimensionality with respect to 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.

Fig. 13. These are the 17 purely 2D test images collected for one of the subjects in the RVL dataset. To the extent possible, the pose of the face is random with respect to the camera viewpoint.

Please cite this article as: D. Kim et al., Multi-view face recognition from single RGBD models of the faces, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.04.008

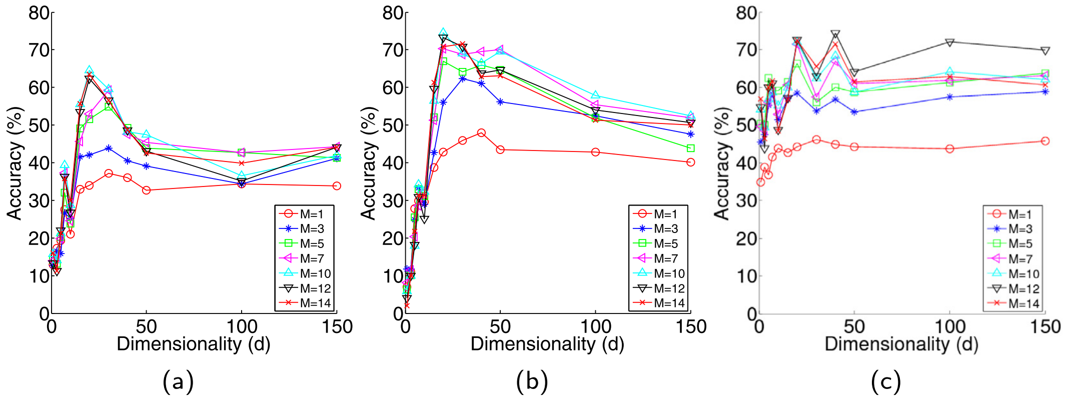


Fig. 14. Multi-view classification accuracy with a single non-partitioned subspace and majority voting as a function of the subspace dimensionality d and the number M of query images for the RVL dataset. (a) Results with a linear SVM classifier (CVS-LSVM), (b) Results with an RBF kernel based SVM classifier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

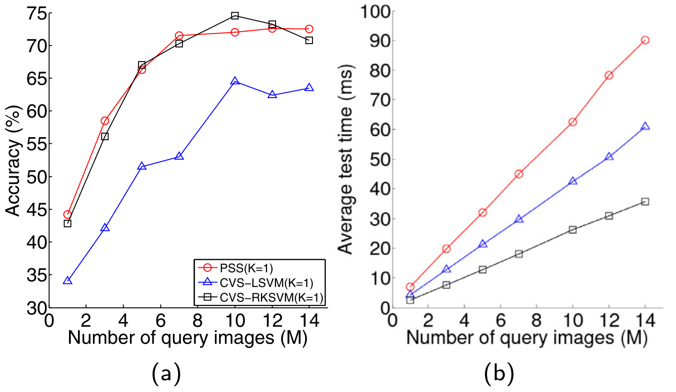


Fig. 15. Classification performance as a function of the number of query images M for a single non-partitioned subspace with the dimensionality for the RVL datasetd $= 20$. (a) Classification accuracies. (b) Time performance of the classifiers for the three cases in (a).

datasets. In this section, we show results on the home-brewed RVL face dataset.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 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 Fig. 11. For each subject, we record a single frontal RGBD image and from that image we generate 925 viewpoint variant images for the subject. The viewpoint variant images cover an angular range of $[-90^\circ, 90^\circ]$ in yaw and $[-60^\circ, 60^\circ]$ in pitch with respect to the frontal view 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 training images generated for each subject. We separately record 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 face pose to the location/orientation of the camera — except for ensuring that the face is sufficiently visible in the camera images. Shown in Fig. 13 are such test images for one of the subjects.

5.2.1. *Majority voting for a non-partitioned subspace*

This study is for the case when we 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.

Fig. 14 shows the classification accuracy as a function of the dimensionality of the subspace. Each datapoint in Fig. 14 as well as in the remainder of this section corresponds to the average over 100 independent realizations of the experiment, with each realization consisting of query images drawn randomly from the testing dataset. The accuracy results plotted in Fig. 14 indicate that the classification accuracy decreases rapidly when the dimensionality of the subspace is made larger than approximately 20. The most significant result in Fig. 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 Fig. 14 from a different perspective, shown in Fig. 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. Fig. 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.

Fig. 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 for 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,

Please cite this article as: D. Kim et al., Multi-view face recognition from single RGBD models of the faces, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.04.008

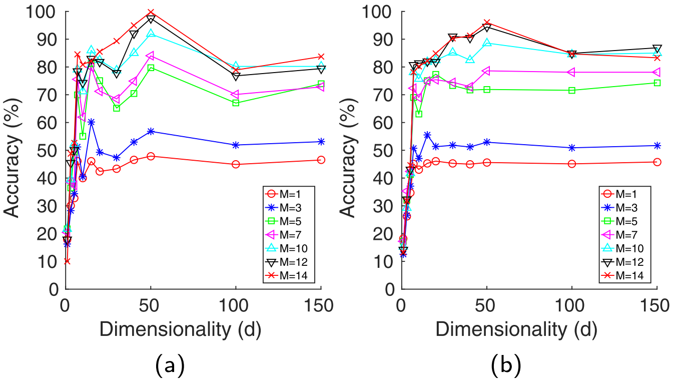


Fig. 16. Multi-view classification accuracy versus subspace dimensionality for the Pose-CVS model with $K = 25$ for the RVL dataset. (a) Linear SVM (Pose-CVS-LSVM). (b) Nonlinear SVM (Pose-CVS-RKSVM).

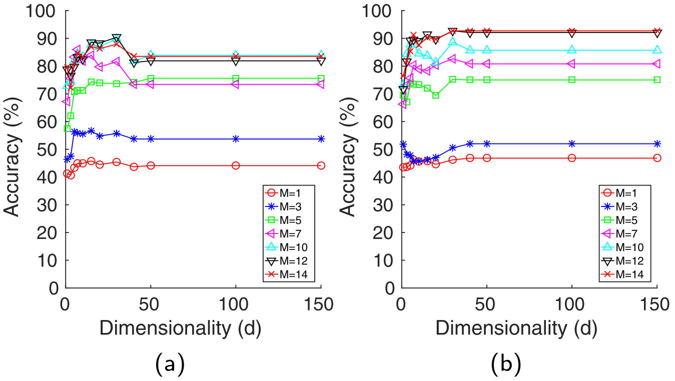


Fig. 17. Multi-view classification accuracy versus the subspace dimensionality for the PSS model with $K = 25$ for the RVL dataset. (a) Appearance based clustering (App-PSS). (b) Pose based clustering (Pose-PSS).

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. Fig. 17 shows results similar to those in Fig. 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-LSVM, it is less robust to noise at higher dimensionalities.

Fig. 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.

Fig. 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 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.

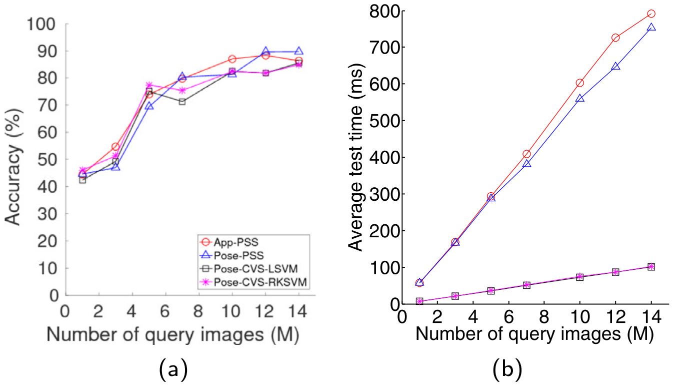


Fig. 18. Classification performance as a function of the number of query images M with the dimensionality fixed at $d = 20$ and the number of view partitions fixed at $K = 25$ for the RVL dataset. (a) Classification accuracies. (b) Time performance of the classifiers in (a).

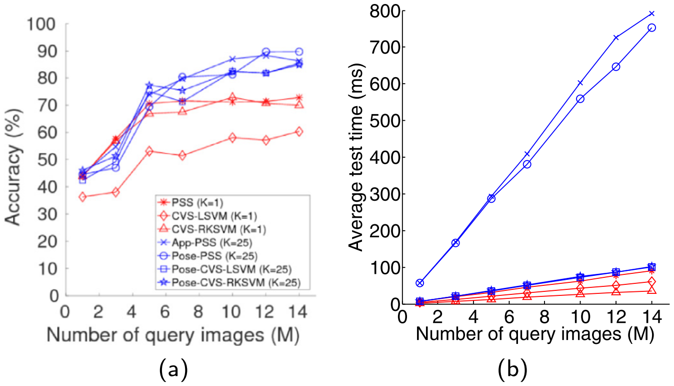


Fig. 19. Comparison of the multi-view classification results for a single subspace with those obtained using view-partitioned subspaces for the RVL dataset. The plots in red are for the case when single non-partitioned subspaces are used and the plots in blue are for the case when view-partitioning is 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). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.2.3. *Weighted voting for view-partitioned subspaces*

Fig. 20(a) compares the classification accuracy obtained with the weighted voting approach of Section 4.1 to that of simple majority voting. In this 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, Fig. 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 additional computational cost when compared to majority voting.

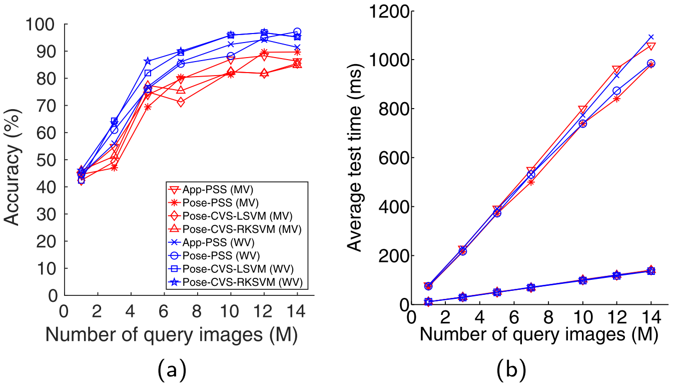


Fig. 20. Comparison of weighted voting with majority voting for the view-partitioned multi-view classification methods with $d = 20$ and $K = 25$ for the RVL dataset. Shown in (a) are the accuracy results and in (b) the average time taken by the classifier to return the result. The plots in red are for majority voting and the plots in blue are for weighted voting. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

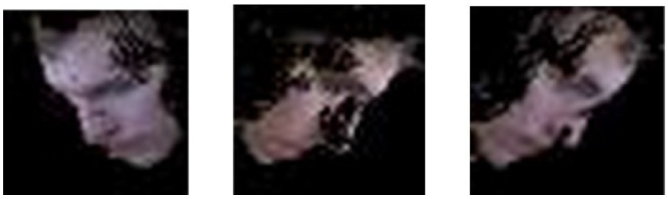


Fig. 21. Examples of holes in the viewpoint variant training images for the publicly available VAP dataset.

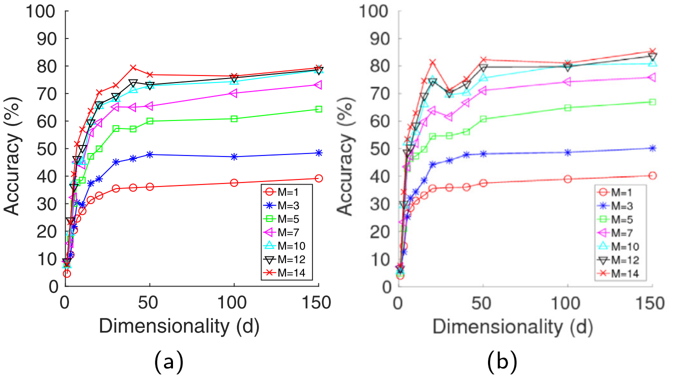


Fig. 22. Multi-view classification accuracy versus subspace dimensionality for the Pose-CVS model with $K = 25$ partitions for the VAP dataset. (a) Linear SVM (Pose-CVS-LSVM). (b) Nonlinear SVM (Pose-CVS-RKSVM).

5.3. Results on the VAP dataset

In this section, we evaluate the various classification strategies presented in our paper on the publicly available VAP database (Høg et al., 2012). 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. Each pose was recorded 3 times resulting in a total of 51 RGBD images per person. More details about the dataset can be found in Høg et al. (2012).

Two pre-processing steps are required to use this dataset for evaluating our classification strategies. First, since the RGB images and depth maps are not co-registered, simple downsampling of the RGB images 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, our goal is to parallel the test data in the RVL dataset to the maximum extent possible. Recall that the testing segment of the RVL dataset consists of 17 2D images of the face of each subject taken from random orientations. For the testing portion drawn from the VAP dataset, we select 17 2D shots randomly from the RGB data associated with the 51 RGBD images for each subject. In making this selection we make sure that no two of the 17 views are for the same pose of the subject. Note that whereas the original dataset is for 31 subjects, we use the data for 28 of them.[2]

It is interesting to note 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 some of the holes in the projected images are caused by the holes in the point clouds and that the holes in the former persist notwithstanding the application of the depth-constrained bilinear interpolation described in Section 3.1. Some examples to illustrate these artifacts are shown in Fig. 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. Fig. 22 depicts the classification accuracy versus dimensionality for different numbers of query images for the Pose-CVS model. In the figure, as well as in the remainder of this section, each datapoint corresponds to the average over 100 independent realizations of the experiment, with each realization consisting of query images drawn randomly from the 2D testing dataset. Again we notice that additional query images increase the accuracy and that the peak accuracy is obtained at a dimensionality of 50, although less noticeably so for the RKSVM model. The performance difference between linear and non-linear kernel SVMs is slightly less evident for this dataset. Fig. 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 Fig. 24. Part (b) of the figure shows the corresponding time performances. As for the RVL dataset, view partitioning significantly improves the performance for the VAP dataset.. Also, we notice that the PSS approaches tend to outperform the CVS approaches even when a nonlinear kernel is used. This result is to be expected since the number of human subjects in this dataset is larger, which would make classification within a common subspace to be more challenging.

[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.

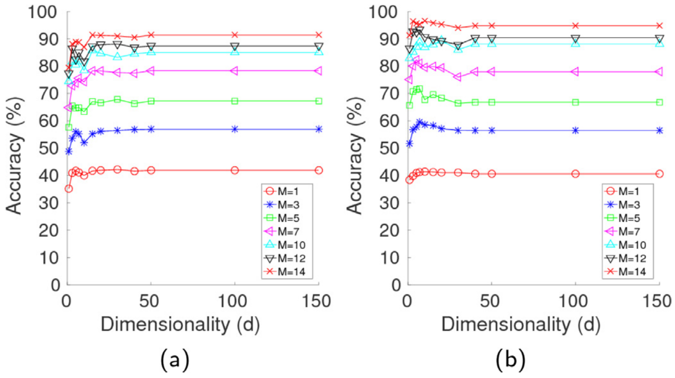


Fig. 23. Multi-view classification accuracy versus the subspace dimensionality for the PSS model with $K = 25$ partitions for the VAP dataset. (a) Appearance based clustering (App-PSS). (b) Pose based clustering (Pose-PSS).

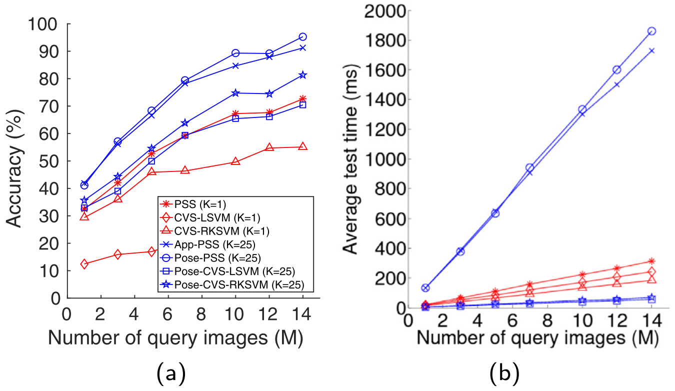


Fig. 24. Comparison of multi-view classification approaches when we use majority voting as a function of the number of query images M with the dimensionality $d = 20$ for the VAP dataset. The plots in red are for the case when single global subspaces are used and the plots in blue are for the case when 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). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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 Fig. 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 publicly available BIWI Kinect Dataset (Fanelli et al., 2011). 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 a total of 24 RGBD image sequences collected for 20 human subjects with the Microsoft Kinect sensor, implying that some of the subjects were recorded more than once. Given that face recognition is the main focus of our work, and that it is desirable to have roughly the same amount of data for each subject, we chose to keep 20 RGBD image sequences, one for each human subject.. This dataset is different from and more challenging than the VAP and RVL datasets in a number of aspects. First, in the other datasets, for generating the 2D images for testing purposes, the human subjects looked at a fixed number of points on a wall so that the images were recorded by the same Kinect sensor at roughly the same angles vis-a-vis the frontal face pose 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 subjects. The number of data frames in each of the 20 retained RGBD image sequences varies between 395 and 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 × 480 pixels. More details about this dataset can be found in Fanelli et al. (2011).

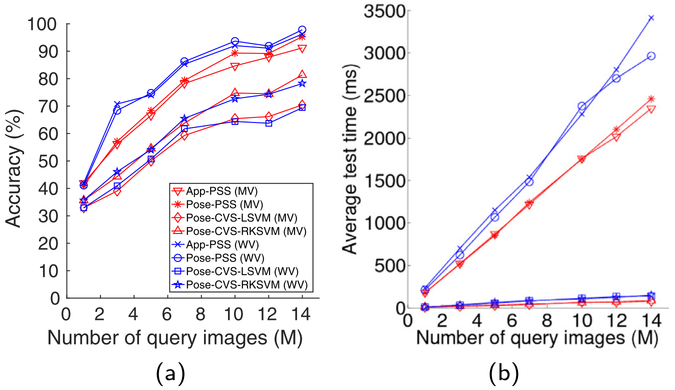


Fig. 25. Comparison of weighted voting with majority voting for the view-partitioned multi-view classification methods for the VAP dataset with $d = 20$ and $K = 25$ in terms of (a) accuracy and (b) average test time. The plots in red are for majority voting and the plots in blue are for weighted voting. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

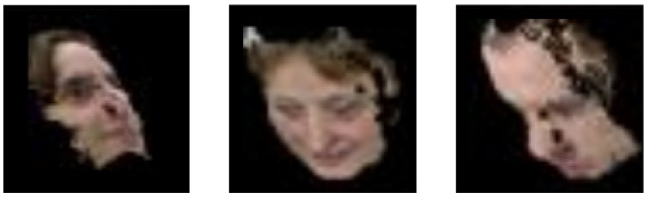


Fig. 26. Examples of viewpoint-variant training images for the BIWI dataset.

Before testing our framework on this dataset, we needed to first align the depth images and the RGB images. Each RGBD recording is provided with its own calibration information for the RGB sensor and the depth sensor, which we used to align the depth and the corresponding RGB images. Specifically, for each pixel in the depth image, we used the calibration information of the depth sensor to backproject the depth value to a 3D point and then used the calibration information of the RGB sensor to find the corresponding color values in the forward projection of that 3D point. After this step, we needed to detect faces in the images. The BIWI dataset also contains mask images that can be used to localize the faces in the RGB projections.

We used one frontal RGBD image for each human subject for generating the 925 viewpoint variant training images. All the remaining RGBD images in each sequence were used for extracting the 2D test images needed for evaluating our algorithms. It is interesting to note that the training data generated from the BIWI dataset contains some of the same artifacts as the training data generated from the VAP dataset. Fig. 26 shows holes in the viewpoint variant images generated from the BIWI dataset after the application of 2.5D interpolation. These artifacts can affect the accuracy of any classifier. Better calibration, surface reconstruction and image alignment strategies are possible solutions to address these

Please cite this article as: D. Kim et al., Multi-view face recognition from single RGBD models of the faces, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.04.008

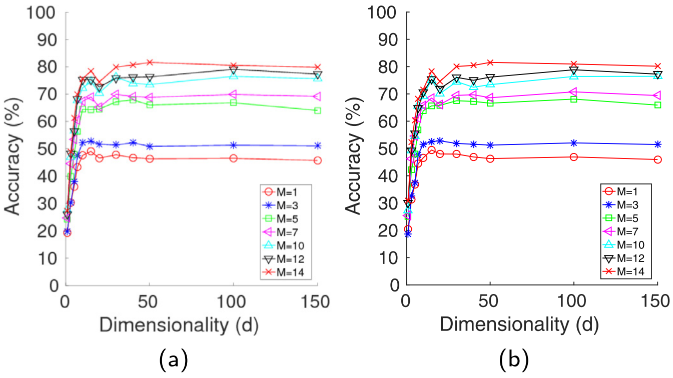


Fig. 27. Multi-view classification accuracy versus subspace dimensionality for the Pose-CVS model with $K = 25$ partitions for the BIWI dataset. (a) Linear SVM (Pose-CVS-LSVM). (b) Nonlinear SVM (Pose-CVS-RKSVM).

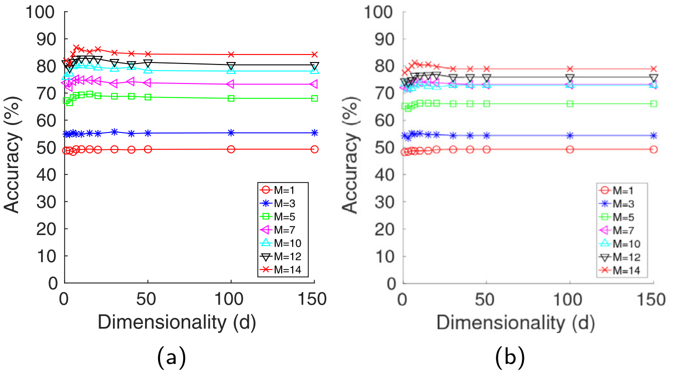


Fig. 28. Multi-view classification accuracy versus subspace dimensionality for the PSS model with $K = 25$ partitions for the BIWI dataset. (a) Appearance based clustering (App-PSS). (b) Pose based clustering (Pose-PSS).

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

5.4.1. Majority voting

We first show results with the majority voting scheme. In Fig. 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 Fig. 28.

Fixing the dimensionality d at 20, we compare the accuracy and time performance of the different models as a function of the number of query images in Fig. 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.

5.4.2. Weighted voting

Fixing dimensionality d as 20, we compare the majority and weighted voting schemes in Fig. 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 the different subjects as well as the large variability in the number of image frames available for the different subjects. Both of these would affect the value of the reconstruction error metric.

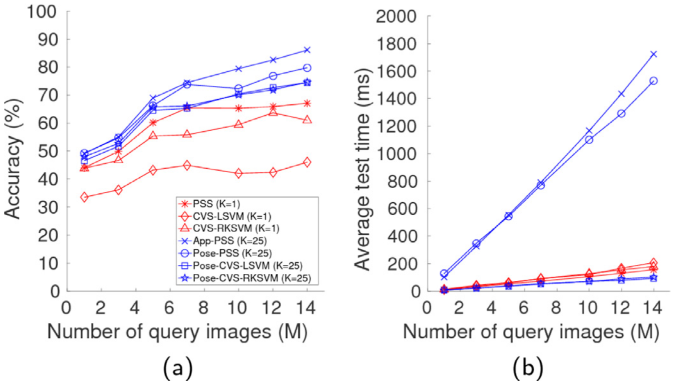


Fig. 29. Comparison of multi-view classification approaches for the BIWI dataset when we use majority voting as a function of the number of query images M with the dimensionality $d = 20$. The plots in red are for the case when single global subspaces are used and the plots in blue are for the case when 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). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

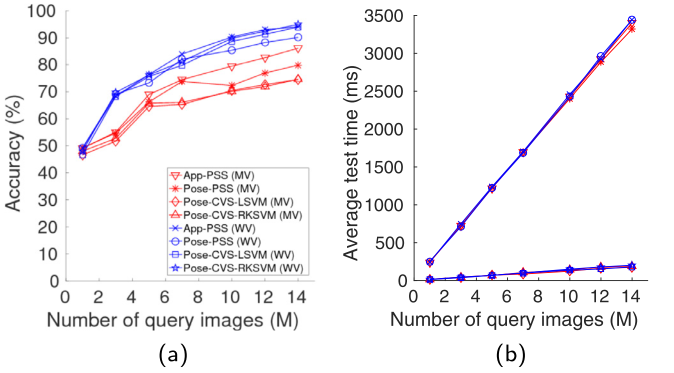


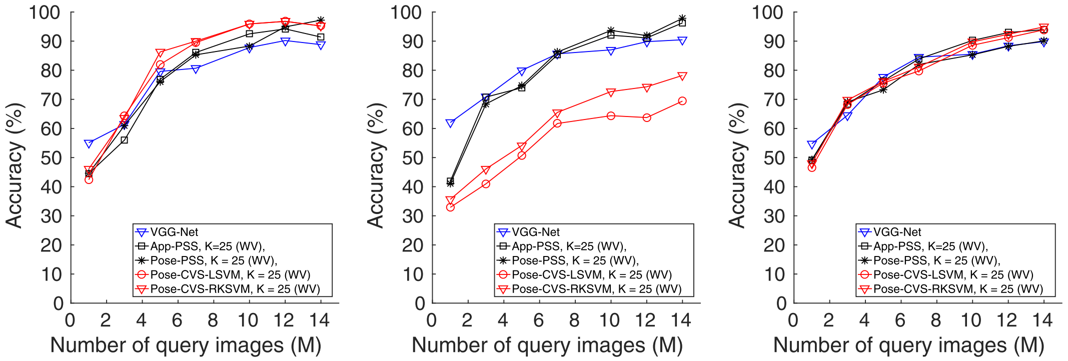
Fig. 30. Comparison of weighted voting with majority voting for the view-partitioned multi-view classification methods for the BIWI dataset with $d = 20$ and $K = 25$ in terms of (a) accuracy and (b) average test time. The plots in red are for majority voting and the plots in blue are for weighted voting. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.5. Comparison with multi-view face recognition using a deep convolutional neural network

We now compare our algorithms with a state-of-the-art face recognition approach proposed by Parkhi et al. (2015), which is based on the VGG deep convolutional network that was originally trained on face images of 2622 different human subjects. For our experiments, we used the MatConvNet framework and an 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 224 × 224 × 3 as per the requirements of the VGG network. In order to retrain the network for our purpose, we first removed 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 2622 classes, its ‘fc8’ layer produces an equal number of outputs. We replaced 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 in this new ‘fc8’



(a) (b) (c)

Fig. 31. Comparison of our proposed approaches with the deep-learning based face recognition system presented in Parkhi et al. (2015) for the (a) RVL dataset, (b) VAP dataset, and (c) BIWI dataset.

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

5.5.2. Results and comparison

For presenting the results in this section, we denote the deep learning classifier by VGG-NET. We fixed $K = 25$ and dimension $d = 20$ for our approaches used in the comparison. In Fig. 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 frontal 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 each query image is given a different weight in the final classification depending on how accurately the query image can be represented in the subspace to which it is assigned.

Here are our three important conclusions: 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, were 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 outperformed 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 case the CVS methods benefited substantially from the weighted voting scheme. The App-PSS approach outperformed the state-of-the-art deep-learning based face recognition method presented in Parkhi et al. (2015) 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 would need to be investigated in the future are the effect of variable resolution query images (variability in the resolution caused by the cameras being at different distances from the human subject) and the presence of motion blur in the images. At the moment it is not clear how a 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 face recognition method are appearance modifiers such as facial hair and eyeglasses. Since such modifiers can be seen as different kinds of partial occlusion, robust dimensionality reduction approaches such as IGO-PCA (Tzimiropoulos et al., 2012) 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.

References

Abate, A.F., Nappi, M., Riccio, D., Sabatino, G., 2007. 2D And 3D face recognition: a survey. Pattern Recognit. Lett. 28 (14), 1885–1906.
An, L., Bhanu, B., Yang, S., 2012. Face recognition in multi-camera surveillance videos. In: International Conference on Pattern Recognition. IEEE, pp. 2885–2888.

Please cite this article as: D. Kim et al., Multi-view face recognition from single RGBD models of the faces, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.04.008

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

Asthana, A., Marks, T., Jones, M., Tieu, K., Rohith, M., 2011. Fully automatic pose-invariant face recognition via 3D pose normalization. In: IEEE International Conference on Computer Vision, pp. 937–944. doi:10.1109/ICCV.2011.6126336.

Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M., 2013. Robust discriminative response map fitting with constrained local models. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Bąk, S., Corvee, E., Bremond, F., Thonnat, M., 2012. Boosted human re-identification using Riemannian manifolds. Image Vis. Comput. 30 (6), 443–452.

Bedagkar-Gala, A., Shah, S.K., 2014. A survey of approaches and trends in person re-identification. Image Vis. Comput. 32 (4), 270–286.

Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J., 1997. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. 19 (7), 711–720.

Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N., 2013. Localizing parts of faces using a consensus of exemplars. IEEE Trans. Pattern Anal. Mach. Intell. 35 (12), 2930–2940. doi:10.1109/TPAMI.2013.23.

Beymer, D., 1994. Face recognition under varying pose. In: Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on, pp. 756–761. doi:10.1109/CVPR.1994.323893.

Beymer, D., Poggio, T., 1995. Face recognition from one example view. In: IEEE International Conference on Computer Vision, pp. 500–507. doi:10.1109/ICCV.1995.466898.

Blanz, V., Vetter, T., 2003. Face recognition based on fitting a 3D morphable model. IEEE Trans. Pattern Anal. Mach. Intell. 25 (9), 1063–1074.

Cai, Y., Huang, K., Tan, T., 2008. Human appearance matching across multiple non-overlapping cameras. In: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. IEEE, pp. 1–4.

Campadelli, P., Lanzarotti, R., Savazzi, C., 2003. A feature-based face recognition system. In: Image Analysis and Processing, 2003.Proceedings. 12th International Conference on, pp. 68–73. doi:10.1109/ICIAP.2003.1234027.

Cao, X., Wei, Y., Wen, F., Sun, J., 2014. Face alignment by explicit shape regression. Int. J. Comput. Vis. 107 (2), 177–190. doi:10.1007/s11263-013-0667-3.

Çeliktutan, O., Ulukaya, S., Sankur, B., 2013. A comparative study of face landmarking techniques. EURASIP J. Image Video Process. 2013 (1), 1–27. doi:10.1186/1687-5281-2013-13.

Chai, X., Shan, S., Chen, X., Gao, W., 2007. Locally linear regression for pose-invariant face recognition. Image Process. IEEE Trans. 16 (7), 1716–1725.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 27:1–27:27.

Choi, J., Dumortier, Y., Choi, S.-I., Ahmad, M., Medioni, G., 2012. Real-time 3-D face tracking and modeling from awebcam. In: Applications of Computer Vision (WACV), 2012 IEEE Workshop on, pp. 33–40. doi:10.1109/WACV.2012.6163031.

Chrysos, G.G., Antonakos, E., Snape, P., Asthana, A., Zafeiriou, S., 2016. A comprehensive performance evaluation of deformable face tracking "in-the-wild". CoRR. abs/1603.06015.

Cootes, T.F., Edwards, G.J., Taylor, C.J., 2001. Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. 23 (6), 681–685.

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

Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1995. Active shape models-their training and application. Comput. Vis. Image Understanding 61 (1), 38–59.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Crabtree, A., Chamberlain, A., Davies, M., Glover, K., Reeves, S., Rodden, T., Tolmie, P., Jones, M., 2013. Doing innovation in the wild. In: Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI. ACM, New York, NY, USA, pp. 25:1–25:9. doi:10.1145/2499149.2499150.

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

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

Du, M., Sankaranarayanan, A., Chellappa, R., 2014. Robust face recognition from multi-view videos. IEEE Trans. Image Process. 23 (3), 1105–1117. doi:10.1109/TIP.2014.2300812.

Fan, W., Yeung, D.-Y., 2006. 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, 2. IEEE, pp. 1384–1390.

Fanelli, G., Gall, J., Gool, L.V., 2011. Real time head pose estimation with random regression forests. In: Computer Vision and Pattern Recognition (CVPR), pp. 617–624.

Fukunaga, K., Olsen, D.R., 1971. An algorithm for finding intrinsic dimensionality of data. Comput. IEEE Trans. 100 (2), 176–183.

Georghiades, A.S., Belhumeur, P.N., Kriegman, D., 2001. From few to many: illumination cone models for face recognition under variable lighting and pose. Pattern Anal. Mach. Intell. IEEE Trans. 23 (6), 643–660.

Ghahramani, Z., Hinton, G.E., et al., 1996. The EM Algorithm for Mixtures of Factor Analyzers. Technical Report. Technical Report CRG-TR-96-1, University of Toronto.

Gong, S., Cristani, M., Loy, C.C., Hospedales, T.M., 2014. The re-identification challenge. In: Person Re-Identification. Springer, pp. 1–20.

Goudelis, G., Zafeiriou, S., Tefas, A., Pitas, I., 2007. Class-specific kernel-discriminant analysis for face verification. IEEE Trans. Inf. Forensics Secur. 2 (3), 570–587. doi:10.1109/TIFS.2007.902915.

Hamm, J., Lee, D.D., 2008. Grassmann discriminant analysis: a unifying view on subspace-based learning. In: Proceedings of the 25th International Conference on Machine Learning. ACM, pp. 376–383.

Harguess, J., Hu, C., Aggarwal, J., 2009. Fusing face recognition from multiple cameras. In: Applications of Computer Vision (WACV), 2009 Workshop on, pp. 1–7. doi:10.1109/WACV.2009.5403055.

Hassner, T., Harel, S., Paz, E., Enbar, R., 2015. Effective face frontalization in unconstrained images. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

Hjelmås, E., Low, B.K., 2001. Face detection: a survey. Comput. Vis. Image Understanding 83 (3), 236–274.

Høg, R., Jasek, P., Rofidal, C., Nasrollahi, K., Moeslund, T., 2012. 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.

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

Hsu, C.-W., Lin, C.-J., 2002. A comparison of methods for multiclass support vector machines. Neural Netw. IEEE Trans. 13 (2), 415–425.

Hu, Y., Huang, T., 2008. Subspace learning for human head pose estimation. In: IEEE International Conference on Multimedia and Expo, pp. 1585–1588.

Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E., 2007. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report. Technical Report 07-49, University of Massachusetts, Amherst.

Kambhatla, N., Leen, T., 1997. Dimension reduction by local principal component analysis. Neural Comput. 9 (7), 1493–1516.

Kan, M., Shan, S., Zhang, H., Lao, S., Chen, X., 2012. Multi-View Discriminant Analysis. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 808–821.

Kazemi, V., Sullivan, J., 2014. One millisecond face alignment with an ensemble of regression trees. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Kim, D., 2015. Pose and Appearance Based Clustering of Face Images on Manifolds and Face Recognition Applications Thereof. Purdue University Ph.D. thesis.

Kim, D., Park, J., Kak, A.C., 2013. 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.

Kim, T.-K., Kittler, J., Cipolla, R., 2007. Discriminative learning and recognition of image set classes using canonical correlations. Pattern Anal. Mach. Intell. IEEE Trans. 29 (6), 1005–1018.

Krueger, V., Zhou, S., 2002. Exemplar-based face recognition from video. In: European Conference on Computer Vision. Springer, pp. 732–746.

Lando, M., Edelman, S., 1995. Receptive field spaces and class-based generalization from a single view in face recognition. Netw. 6 (4), 551–576.

Lanitis, A., Taylor, C.J., Cootes, T.F., 1997. Automatic interpretation and coding of face images using flexible models. IEEE Trans. Pattern Anal. Mach. Intell. 19 (7), 743–756. doi:10.1109/34.598231.

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

Lee, K.-C., Ho, J., Yang, M.-H., Kriegman, D., 2003. Video-based face recognition using probabilistic appearance manifolds. In: IEEE Conference on Computer Vision and Pattern Recognition, 1, pp. 313–320.

Lee, K.-C., Kriegman, D., 2005. 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, 1. IEEE, pp. 852–859.

Li, S.Z., Lu, X., Hou, X., Peng, X., Cheng, Q., 2005. Learning multiview face subspaces and facial pose estimation using independent component analysis. IEEE Trans. Image Process. 14 (6), 705–712. doi:10.1109/TIP.2005.847295.

Lindner, C., Bromiley, P.A., Ionita, M.C., Cootes, T.F., 2015. Robust and accurate shape model matching using random forest regression-voting. IEEE Trans. Pattern Anal. Mach. Intell. .37 (9), 1862–1874. doi:10.1109/TPAMI.2014.2382106.

Liu, J., Li, Y., Allen, P.K., Belhumeur, P.N., 2015. Articulated pose estimation using hierarchical exemplar-based models. CoRR. abs/1512.04118.

Liu, X., Chen, T., 2003. Video-based face recognition using adaptive hidden markov models. In: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, 1. IEEE, pp. I–340.

Lu, C., Tang, X., 2014. Surpassing human-level face verification performance on LFW with Gaussianface. CoRR. abs/1404.3840.

Lu, J., Tan, Y.P., Wang, G., 2013. Discriminative multimanifold analysis for face recognition from a single training sample per person. IEEE Trans. Pattern Anal. Mach. Intell. 35 (1), 39–51. doi:10.1109/TPAMI.2012.70.

Lucey, S., Wang, Y., Cox, M., Sridharan, S., Cohn, J.F., 2009. Efficient constrained local model fitting for non-rigid face alignment. Image Vis. Comput. 27 (12), 1804–1813. Visual and multimodal analysis of human spontaneous behaviour.

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

Marras, I., Tzimiropoulos, G., Zafeiriou, S., Pantic, M., 2014. Online learning and fusion of orientation appearance models for robust rigid object tracking. Image Vis. Comput. 32 (10), 707–727. Best of Automatic Face and Gesture Recognition 2013.

Matthews, I., Baker, S., 2004. Active appearance models revisited. Int. J. Comput. Vis. 60 (2), 135–164.

Mazzon, R., Tahir, S.F., Cavallaro, A., 2012. Person re-identification in crowd. Pattern Recognit. Lett. 33 (14), 1828–1837.

Alabort-i Medina, J., Antonakos, E., Booth, J., Snape, P., Zafeiriou, S., 2014. Menpo: a comprehensive platform for parametric image alignment and visual deformable models. In: Proceedings of the 22Nd ACM International Conference on Multimedia. ACM, New York, NY, USA, pp. 679–682. doi:10.1145/2647868.2654890.

Morency, L., Whitehill, J., Movellan, J., 2008. Generalized adaptive view-based appearance model: integrated framework for monocular head pose estimation. In: IEEE International Conference on Automatic Face & Gesture Recognition, pp. 1–8.

Niinuma, K., Han, H., Jain, A.K., 2013. Automatic multi-view face recognition via 3D model based pose regularization. In: Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on, pp. 1–8. doi:10.1109/BTAS.2013.6712735.

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

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

Otsu, N., 1975. A threshold selection method from gray-level histograms. Automatica 11 (285–296), 23–27.

Parkhi, O.M., Vedaldi, A., Zisserman, A., 2015. Deep face recognition. In: British Machine Vision Conference.

Pentland, A., Moghaddam, B., Starner, T., 1994. View-based and modular eigenspaces for face recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 84–91.

Phillips, P.J., Grother, P., Micheals, R., 2011. Evaluation methods in face recognition. Springer.

Pnevmatikakis, A., Polymenakos, L., 2007. Far-field multi-camera video-to-video face recognition. Face Recognit. 467–486.

Ren, S., Cao, X., Wei, Y., Sun, J., 2014. Face alignment at 3000 fps via regressing local binary features. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Rogers, Y., 2011. Interaction design gone wild: striving for wild theory. Interactions 18 (4), 58–62. doi:10.1145/1978822.1978834.

Roweis, S., Saul, L., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290 (5500), 2323–2326.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M., 2013a. 300 Faces in-the-wild challenge: the first facial landmark localization challenge. In: The IEEE International Conference on Computer Vision (ICCV) Workshops.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M., 2013b. A semi-automatic methodology for facial landmark annotation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

Saragih, J.M., Lucey, S., Cohn, J.F., 2011. Deformable model fitting by regularized landmark mean-shift. Int. J. Comput. Vis. 91 (2), 200–215. doi:10.1007/s11263-010-0380-4.

Satta, R., Fumera, G., Roli, F., 2012. Fast person re-identification based on dissimilarity representations. Pattern Recognit. Lett. 33 (14), 1838–1848.

Saul, L.K., Roweis, S.T., 2003. Think globally, fit locally: unsupervised learning of low dimensional manifolds. J. Mach. Learn. Res. 4, 119–155.

Seung, H.S., Lee, D.D., 2000. The manifold ways of perception. Science 290 (5500), 2268–2269.

Shakhnarovich, G., Fisher, J.W., Darrell, T., 2002. Face recognition from long-term observations. In: European Conference on Computer Vision. Springer, pp. 851–865.

Sharma, A., Kumar, A., Daume, H., Jacobs, D.W., 2012. Generalized multiview analysis: a discriminative latent space. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 2160–2167. doi:10.1109/CVPR.2012.6247923.

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

Stegmann, M.B., Ersboll, B.K., Larsen, R., 2003. Fame-a flexible appearance modeling environment. IEEE Trans. Med. Imaging 22 (10), 1319–1331. doi:10.1109/TMI.2003.817780.

Stegmann, M.B., Olsen, S., 2001. Object tracking using active appearance models. In: Proc. 10th Danish Conference on Pattern Recognition and Image Analysis, pp. 54–60.

Sun, Y., Wang, X., Tang, X., 2013. Deep convolutional network cascade for facial point detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Sung, J., Kanade, T., Kim, D., 2008. Pose robust face tracking by combining active appearance models and cylinder head models. Int. J. Comput. Vis. 80 (2), 260–274. doi:10.1007/s11263-007-0125-1.

Taigman, Y., Yang, M., Ranzato, M., Wolf, L., 2014. Deepface: closing the gap to human-level performance in face verification. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pp. 1701–1708. doi:10.1109/CVPR.2014.220.

Tenenbaum, J., De Silva, V., Langford, J., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290 (5500), 2319–2323.

Tzimiropoulos, G., 2015. Project-out cascaded regression with an application to face alignment. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3659–3667. doi:10.1109/CVPR.2015.7298989.

Tzimiropoulos, G., Pantic, M., 2013. Optimization problems for fast AAM fitting in-the-wild. In: 2013 IEEE International Conference on Computer Vision, pp. 593–600.

Tzimiropoulos, G., Pantic, M., 2014. Gauss–Newton deformable part models for face alignment in-the-wild. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Tzimiropoulos, G., Zafeiriou, S., Pantic, M., 2012. Subspace learning from image gradient orientations. IEEE Trans. Pattern Anal. Mach. Intell. 34 (12), 2454–2466. doi:10.1109/TPAMI.2012.40.

Valstar, M., Martinez, B., Binefa, X., Pantic, M., 2010. Facial point detection using boosted regression and graph models. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 2729–2736. doi:10.1109/CVPR.2010.5539996.

Vapnik, V., 1963. Pattern recognition using generalized portrait method. Autom. Remote Control 24, 774–780.

Verbeek, J., 2006. Learning nonlinear image manifolds by global alignment of local linear models. Pattern Anal. Mach. Intell. IEEE Trans. 28 (8), 1236–1250.

Vetter, T., Blanz, V., 1998. Estimating coloured 3D face models from single images: an example based approach. In: European Conference on Computer Vision. Springer, pp. 499–513.

Viola, P., Jones, M., 2001. 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, 1. IEEE, pp. I–511.

Wang, R., Shan, S., Chen, X., Dai, Q., Gao, W., 2012. Manifold–manifold distance and its application to face recognition with image sets. IEEE Trans. Image Process. 21 (10), 4466–4479.

Wu, H., Souvenir, R., 2015. Robust regression on image manifolds for ordered label denoising. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Xie, B., Boult, T., Ramesh, V., Zhu, Y., 2006. 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, pp. 18–23.

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

Xiong, X., De la Torre, F., 2013. Supervised descent method and its applications to face alignment. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Yamaguchi, O., Fukui, K., Maeda, K., 1998. Face recognition using temporal image sequence. In: Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on, pp. 318–323. doi:10.1109/AFGR.1998.670968.

Yang, H., Jia, X., Loy, C.C., Robinson, P., 2015. An empirical study of recent face alignment methods. CoRR. abs/1511.05049.

Yang, M.-H., Kriegman, D., Ahuja, N., 2002. Detecting faces in images: a survey. Pattern Anal. Mach. Intell. IEEE Trans. 24 (1), 34–58.

Yang, Y., Ramanan, D., 2013. Articulated human detection with flexible mixtures of parts. IEEE Trans. Pattern Anal. Mach. Intell. 35 (12), 2878–2890. doi:10.1109/TPAMI.2012.261.

Yoder, J., Medeiros, H., Park, J., Kak, A., 2010. Cluster-based distributed face tracking in camera networks. Image Process. IEEE Trans. 19 (10), 2551–2563. doi:10.1109/TIP.2010.2049179.

Zafeiriou, S., Zhang, C., Zhang, Z., 2015. A survey on face detection in the wild: past, present and future. Comput. Vision Image Understanding 138, 1–24. http://dx.doi.org/10.1016/j.cviu.2015.03.015.

Zaki, S.M., Yin, H., 2015. Multi-Manifold Approach to Multi-view Face Recognition. Springer International Publishing, pp. 370–377.

Zhang, C., Zhang, Z., 2010. A Survey of Recent Advances in Face Detection. Technical Report. Tech. rep., Microsoft Research.

Zhao, W., Chellappa, R., 2000. SFS based view synthesis for robust face recognition. In: Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on. IEEE, pp. 285–292.

Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A., 2003. Face recognition: a literature survey. ACM Comput. Surv. (CSUR) 35 (4), 399–458.

Zhou, E., Cao, Z., Yin, Q., 2015. Naive-deep face recognition: touching the limit of LFW benchmark or not? CoRR. abs/1501.04690.

Zhou, S., Krueger, V., Chellappa, R., 2003. Probabilistic recognition of human faces from video. Comput. Vis. Image Understanding 91 (1), 214–245.

Zhou, S.K., Chellappa, R., 2006. From sample similarity to ensemble similarity: probabilistic distance measures in reproducing kernel hilbert space. Pattern Anal. Mach. Intell. IEEE Trans. 28 (6), 917–929.

Zhu, S., Li, C., Change Loy, C., Tang, X., 2015. Face alignment by coarse-to-fine shape searching. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Zhu, X., Ramanan, D., 2012. Face detection, pose estimation, and landmark localization in the wild. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 2879–2886. doi:10.1109/CVPR.2012.6248014.

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