Using 3D Models to Recognize 2D Faces in the Wild

Iacopo Masi, Giuseppe Lisanti, Andrew D. Bagdanov, Pietro Pala and Alberto Del Bimbo
MICC - Media Integration and Communication Center, University of Florence
Viale Morgagni 65, Florence 50134, Italy
http://www.micc.unifi.it/vim/

Abstract

In this paper we consider the problem of face recognition in imagery captured in uncooperative environments using PTZ cameras. For each subject enrolled in the gallery, we acquire a high-resolution 3D model from which we generate a series of rendered face images of varying viewpoint. The result of regularly sampling face pose for all subjects is a redundant basis that over represents each target. To recognize an unknown probe image, we perform a sparse reconstruction of SIFT features extracted from the probe using a basis of SIFT features from the gallery. While directly collecting images over varying pose for all enrolled subjects is prohibitive at enrollment, the use of high speed, 3D acquisition systems allows our face recognition system to quickly acquire a single model, and generate synthetic views offline. Finally we show, using two publicly available datasets, how our approach performs when using rendered gallery images to recognize 2D rendered probe images and 2D probe images acquired using PTZ cameras.

1. Introduction

Automatic face recognition is one of the classic, fundamental problems in the computer vision community. In recent years even more effort has gone into studying techniques and systems for accurately modeling facial appearance and for recognizing faces in diverse environments [14]. A general statement of the automatic face recognition problem, from a computer vision standpoint, can be formulated as follows: given a *probe image or video* of a scene, verify the identity of one or more of the persons in it using stored *gallery of known individuals*. Despite its long history as a central problem in computer vision, face recognition remains a subject of great practical and theoretical interest [14].

The basic process of face recognition consists of:

- **Enrollment** of individuals in the gallery of known people. Enrollment usually takes the form of the capture of a sequence of high resolution images of each person, or a 3D model of each face if the system is designed for recognition of 3D probe images. A critical point for applicability of face recognition systems in practice is that enrollment be as efficient as possible.

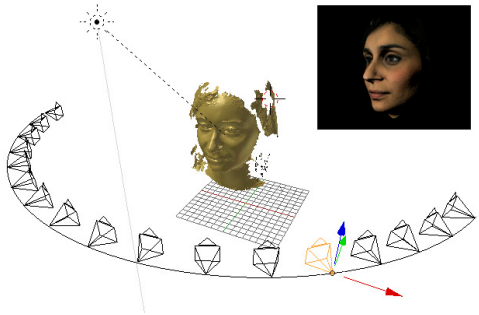

Figure 1. Synthetic data generation process: given a 3D model, we sample the yaw angle by rendering 25 poses. The highlighted camera gives the facial image shown in top right corner.

- **Learning** of discriminative or generative models of gallery subjects to be used for later recognition of faces in probe images. A variety of methods can be used for this stage, and in case of 3D face recognition the learning process often involves the estimation of an average 3D face that will be used to register probe image faces at recognition time.
- **Recognition** of unknown individuals in probe images. In this phase unknown faces in probe images are classified using models learned on the gallery image set. There are also myriad recognition scenarios, though they can be coarsely categorized into cooperative and uncooperative scenarios. In *cooperative* scenarios the unknown person is assumed to actively submit to facial image capture at recognition time and the resulting probe images are usually frontal and of very high quality. In *uncooperative* scenarios, recognition is passive and probe images must be captured using passive sensors in the environment. As with enrollment, it

is important that recognition be as efficient and non-intrusive as possible.

In this paper we take a hybrid approach that exploits 3D face models to recognize faces in PTZ camera imagery. From a high resolution 3D model of faces, we artificially generate multiple views of each subject by rendering the enrolled 3D models from varying viewpoints. The acquisition process for rendered 2D model views is illustrated in Fig. 1. From these rendered face images we extract SIFT descriptors at salient image positions, and, rather than quantizing these descriptors against a visual vocabulary, we then represent each individual as an unordered bag of SIFT features. By varying the viewpoint of subjects in the gallery, we reduce the need for frontal face imagery for use as probes. Probe images are also represented as unordered bags of SIFT features, and recognition is performed through sparse reconstruction of probe image features from gallery image features. The use of sparse reconstruction allows our approach to leverage the multiple views of each subject in the reconstruction of unknown probe images.

In the next Section we discuss work related to face recognition and sparse discriminative classifiers. In Section 3 we describe how we acquire high-resolution 3D models of gallery subjects, generate rendered images from multiple viewpoints of each, and finally classify unknown probe images using these rendered views. We describe a series of experiments performed on two face datasets in Section 4, and finally conclude with a discussion of ongoing work in Section 5.

2. Related Work

In this section we briefly review the literature on hybrid recognition approaches, by which we mean automatic recognition systems using both 3D and 2D face data. For a more thorough survey of face recognition in general, the interested reader should consult the excellent reviews in [1, 17].

The method in [5] estimates 3D shape and texture of faces from single images. Rather than directly acquiring a 3D model from faces at enrollment, an estimate of a 3D face model is computed by fitting a morphable 3D model, learned from a set of textured 3D scans of faces, to images. Recognition is performed by matching the shape and texture information after fitting the 2D probe images to the 3D model.

In [9] the authors propose a method for view and pose invariant face recognition that combines component-based recognition and 3D morphable models. The approach first uses a 3D morphable model to generate 3D face models from only two input images of each person enrolled in the gallery database. By rendering the 3D models under varying pose and illumination conditions they create a large number of synthetic face images which are used to train a component-based face recognition system. Differently from our approach they generate a coarse 3D model from two 2D views of face and perform a two stage classification in which they first individuate the face component in the test image using an SVM classifier then detect the configuration of components to feed a geometric classifier.

The authors of [6] propose a face recognition solution combining both 2D and 3D face data. They develop a PCA-based approach tuned separately for 2D and for 3D. A multi-modal decision is obtained by first matching a 2D probe against the 2D gallery, and then the 3D probe against the 3D gallery. A confidence is computed for the 2D and 3D recognition scores and these confidences are used as weights in the sum of distances to obtain final classification score. Unlike our approach they use both 3D and 2D images in both the probe and gallery sets and only use the texture information of the 3D model as 2D views.

In [13] the authors propose a method to learn a person detector from synthetic data generated from virtual scenarios. More specifically, they record training sequences in virtual scenarios to learn an appearance-based pedestrian classifiers based on HOG and linear SVM. By testing the learned model on images containing real pedestrians they demonstrate that is possible to learn a model for detection also from synthetic data. One of the objectives of our work is to extend this approach from detection to recognition tasks by generating synthetic views from high resolution 3D models of faces.

The ℓ_1-regularized sparse basis expansion has been used in literature to perform person recognition on well-cropped 2D face images coming from the same source. In particular, Wright *et al.* [16] show how sparse representation can be used as a powerful classification tool for face recognition. This approach has been extended several times, integrating correntropy [8] and kernel-based sparse reconstruction [10]. Elhamifar and Vidal [7] extend the Sparse Discriminative Classifier of [16] by constraining the method to find a representation of a test example using the minimum number of blocks from the dictionary (each block corresponds to multiple instances of the same subject).

3. 2D Face Recognition from 3D Models

In this Section we describe our approach to hybrid 2D/3D face recognition. The first step in our approach is the acquisition of high resolution 3D models of each individual enrolled in the gallery, and then the synthetic generation of multiple 2D views of each individual. The final step is face recognition using the synthetic redundant basis to identify the probe.

3.1. 2D Face Synthesis and Feature Extraction

A high resolution 3D model for each individual is quickly acquired at enrollment using a 3D scanner. From each model we artificially generate n synthetic images across varying viewpoints. These images of the i-th person in the gallery are:

$$\mathcal{I}_i = \left\{u_i^1, u_i^2, \ldots, u_i^n\right\}, \text{ for } i \in [1, \ldots, P]. \quad (1)$$

In principle, the rendered images of each subject can be generated by varying both the yaw and pitch of each 3D model, and also by varying the illumination direction and illuminant. In this work, however, we consider only varying yaw angle for generating synthetic 2D images of each subject. We generate views by uniformly sampling 25 yaw angles in the range $[-90°, +90°]$. This process is illustrated in Fig. 1.

The final representation of individuals is an unordered bag of SIFT descriptors calculated at salient image points identified using a Harris-Laplace corner detector. The Bag of Features corresponding to the i-th person in the gallery:

$$\mathcal{X}_i = \left\{\mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^{n_i}\right\}, \text{ for } i \in [1, \ldots, P],$$

where each $\mathbf{x}_i^j$ is the j-th SIFT descriptor extracted from the images of the i-th gallery individual. To simplify notation we do not use an index on SIFT features to indicate from which image $\mathcal{X}$ comes.

In Fig. 2 we illustrate some of the rendered images derived from a model in the Florence 2D/3D face dataset. Note the high quality of the resulting images, which is due to the very high resolution of the models in the dataset (each model has around 70,000 facets, and a 4MPixel texture, on average).

Feature extraction from probe images is performed in a similar fashion, though of course without the synthesis process from 3D models. Assume we have a probe image that contains a face region corresponding to a single individual. We use the Viola-Jones face detector [15] to identify frontal and profile faces [2]; then we extract SIFT descriptors at salient points identified with the Harris-Laplace corner detector in the detected face region. The probe image is represented as a bag of SIFT features:

$$\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m\}, \quad (2)$$

where $\mathbf{z}_j$ is the j-th SIFT descriptor extracted from the probe image.

3.2. Face Recognition by Sparse Reconstruction

Given the gallery representation as bags of unordered SIFT features $\mathcal{X}_i$ and a probe image $\mathcal{Z}$, also represented as a bag of SIFT features, we perform face recognition using a sparse discriminative classifier, similar to that of [16]. We start by computing a ℓ_1-regularized sparse basis expansion of each probe SIFT $\mathbf{z}_i$ as a sparse linear combination of SIFT descriptors in $\mathcal{X}$:

$$\widehat{\mathbf{A}} = \arg\min_{\mathbf{A}} ||\mathbf{Z} - \mathbf{X}\mathbf{A}||_2 + \lambda||\mathbf{A}||_1, \quad (3)$$

where $\mathbf{X}$ is a column-wise concatenation of all gallery SIFT features from all $\mathbf{X}_i$, and thus is a matrix of size $128 \times \sum_{i=1}^{P} n_i$, and $\mathbf{Z}$ is similarly a column-wise concatenation of the m SIFT points from the probe image (and thus a matrix of size $128 \times m$). Despite the potentially large number of SIFT points (especially in the gallery), there exist very efficient techniques for solving these types of ℓ_1-regularized reconstruction problems [12]. We discovered a good value for λ to be 0.1 and fixed this value for all the experiments.

To perform classification, we examine the reconstruction error obtained by limiting the basis expansion to SIFT points extracted from gallery images corresponding to a single individual:

$$\varepsilon_i = ||\mathbf{Z} - \mathbf{X}\mathbf{I}_i\widehat{\mathbf{A}}||_2, \text{ for } i \in \{1, \ldots, P\}, \quad (4)$$

where $\mathbf{I}_i$ is a diagonal matrix with ones on the diagonal corresponding to SIFT descriptors in $\mathbf{X}$ extracted from images of subject i, and zeros everywhere else. This matrix effectively selects only those coefficients in the solution matrix $\widehat{\mathbf{A}}$ that correspond to the i-th person in the gallery. The identity of the probe image is classified as the one yielding the lowest overall error ε_i.

If we have multiple probe images of each subject, we apply the method described above for each image and accumulate the reconstruction errors across all probe images. Then we assign the identity to the person by taking the minimum of the ratio between the probe image yielding minimum reconstruction error and the probe image yielding the second best reconstruction error.

4. Experimental Results

In this section we report on a variety of experiments we performed on two face datasets. For each experiment we define the number of tested images per subject as N, while the number of images per subject in the gallery is M. We evaluate our approach using two test modalities:

- *Single image vs Multi image*: considering each single image in the probe tested independently ($N = 1$), and having multiple images per subject in the gallery ($M > 1$).
- *Multi image vs Multi image*: using multiple images in the probe ($N > 1$) in addition to multiple gallery images ($M > 1$), modeling scenarios in which multiple face images of the same subject can be reliably associated.

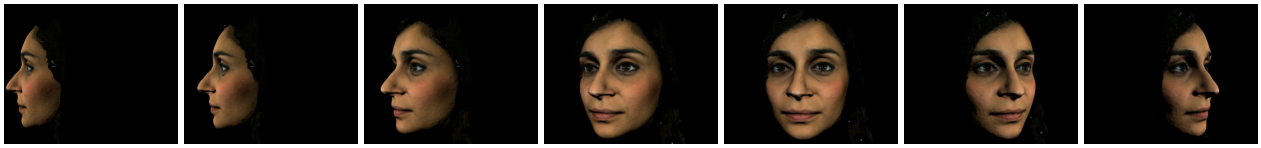

Figure 2. 2D face views synthesized from the 3D model. Images are generated by varying the yaw angle of the 3D model rendering a 2D image.

Figure 3. Some of the 2D face views obtained from a PTZ camera at different level of zoom.

We express the performance figures of our approach in term of ROC (Receiver Operating Characteristic) curves and by reporting the Recognition Rate at First Rank.

4.1. Experiments on 2D Images

The first set of experiments we performed was on the FacePix dataset of 2D face images [11]. This dataset is particularly appropriate for testing the central idea of our approach since each subject has been directly imaged under a variety of poses and illumination conditions. In particular, the FacePix dataset provides facial poses for each subject from +90°to -90°at increments of one degree. This results in 181 images per subject considering only pose variations.

In this experiment both the gallery and the probe sets contain real 2D images from the same dataset. The objective of this experiment is to show the ability of our approach to scale with respect to the number of images present in the gallery and to validate our belief that a redundant gallery can provide excellent recognition performance. We perform 2-fold cross validation considering all the images per subject where the pose ranges from -90°to +90°. After selecting these poses, we vary the number of images in the gallery by sampling the pose.

The results of our approach are shown in Fig. 4. The ROC curves represent the improvement in performance over varying numbers of images in the gallery $M = \{3, 6, 12\}$. The probe images are tested independently of each other using the *Single image vs Multi image* modality, and thus $N = 1$. Considering the number of images in the gallery, our method achieves recognition rates at first rank of 75.9% with $M = 3$, 92.2% with $M = 6$ and 98.5% with $M = 12$. These results indicate, as expected, that given enough variety in samples of each individual, high classification accuracy can be achieved.

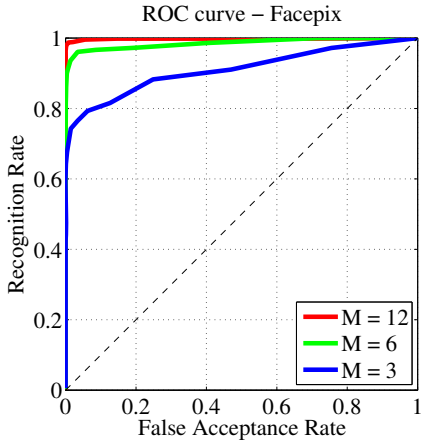

Figure 4. ROC curves for face recognition on FacePix. The different curves represent different numbers of gallery images per subject (M).

4.2. Experiments on Rendered 2D Images

For these experiments we use the 3D models from the Florence 2D/3D Face Dataset [3]. The models in the database are raw 3D meshes along with associated textures.

In order to asses the potential of our approach, we duplicate the FacePix experimental scenario with face imagery renderd using the 3D models from the Florence dataset. We rendered images from 22 of the subjects of this dataset using the approach described in Section 3.1. Sampling 25 yaw angles per subject, we obtain a gallery of 550 images. On this dataset we perform again 2-fold cross validation by varying the number of images per subject in the gallery in the range $M = \{2, 3, 13\}$. The rendered images are very similar to each other and face recognition performance saturates quickly. We achieve excellent recognition accuracy when considering half of the images ($M = 13$) in the gallery set per subject. The probe images are tested independently of

each other in the *Single image vs Multi images* modality, and thus $N = 1$. The ROC curves for these experiments are given in Fig. 5(a). Varying the number of images in the gallery, we obtain recognition rates of 66.6% with $M = 2$, 84.0% with $M = 3$ and 100% with $M = 13$.

4.3. Rendered 2D Gallery versus 2D Probes

In this Section we report preliminary experimental results on face recognition in a video streams from the Florence 2D/3D Face Dataset captured from a PTZ camera viewing one person, as shown in Fig. 3. Recognition is performed using a gallery of rendered images from 3D models as described in Section 3.1. This scenario is very challenging considering that subjects were told to act naturally and we are basically comparing multi-modal data: probes imaged by the PTZ camera, and gallery images rendered using 3D models.

In these experiments, we tried using both a single probe image for test ($N = 1$), and multiple probe images ($N > 1$). In all these experiments we used the synthetic rendered images as described in Section 3.1 as gallery, thus each subject has 25 images rendered across with varying yaw as shown in Fig. 2.

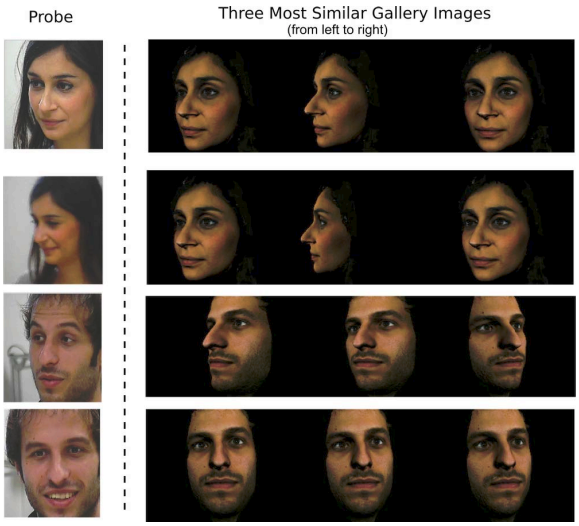

Figure 6. Face recognition results. *Left:* the probe image to identify. *Right:* the most similar images in the gallery (from left to right) in terms of the coefficient energy used the reconstruction ($\mathbf{I}_i\widehat{\mathbf{A}}$). Note that the face pose of the image with highest coefficient energy tends to be very similar to the pose of the subject in the 2D image.

Single image vs Multi image. The performance using single probe image per subject is shown in Fig. 5(b). In this experiment we also attempted to quantify how the system performs under zoom variation (given that zoom variations affect the imaged face size) by sampling faces uniformly across the entire PTZ sequence and hence including testing probe images at difference zoom levels. In the legend of Fig. 5(b) we report the average size of the faces in probe images. In these ROC curves, there is little difference between the three sets of zoom levels, each achieving a recognition rate between 20% and 28%. This is likely due to the fact that other factors, such as facial expression and extreme pose variation, affect accuracy more than variations in face size.

In Fig. 6 we show four cases of true positive along with the three most similar images from the gallery from left to right. Note the probe images are captured "in the wild" with expressions, large pose variations and motion blur. It is interesting to note that most similar face image in the probe usually has a face pose similar to that of the face in the image.

Multi image vs Multi image. In this experiment we evaluate the performance of our approach using multiple images in the probe. This assumption of multiple images in the probe is well known in literature for person re-identification [4] and it seems also a reasonable assumption in real world scenarios if we consider a tracker that can track and schedule a PTZ camera to follow the target face [2].

In Fig. 5(c) we report the performance of our approach over varying number of images used in the probe ($N > 1$). From this figure we see that using more than one image to describe an unknown person improves overall accuracy. In particular just considering $N = 7$ we outperform the *Single image vs Multi image* approach, with a recognition rate of **31.8**%. If we continue to add images from the video stream, the chance of get the right person goes up to **36.3**% with $N = 10$ and to **45.5**% with $N = 15$.

5. Conclusions and Future Work

In this paper we described a hybrid approach to face recognition that uses rendered images of 3D models to form a gallery of images with varying pose for each enrolled subject. SIFT feature descriptors are extracted from these images and form a bag of features representing each gallery image. Probe images are similarly represented as unordered bags of SIFT descriptors. An ℓ_1-regularized reconstruction of probe image descriptors is used to derive a sparse discriminative classifier that effectively incorporates the information present in multiple views into the recognition process. An advantage of our approach is that no discriminative model is learned and adding new subjects to the gallery requires only concatenation of SIFT features to the existing gallery.

Experiments on a standard 2D face dataset demonstrate that our approach is very effective when very many views of each subject are incorporated into the gallery, and similar experiments on rendered 2D images for both gallery and

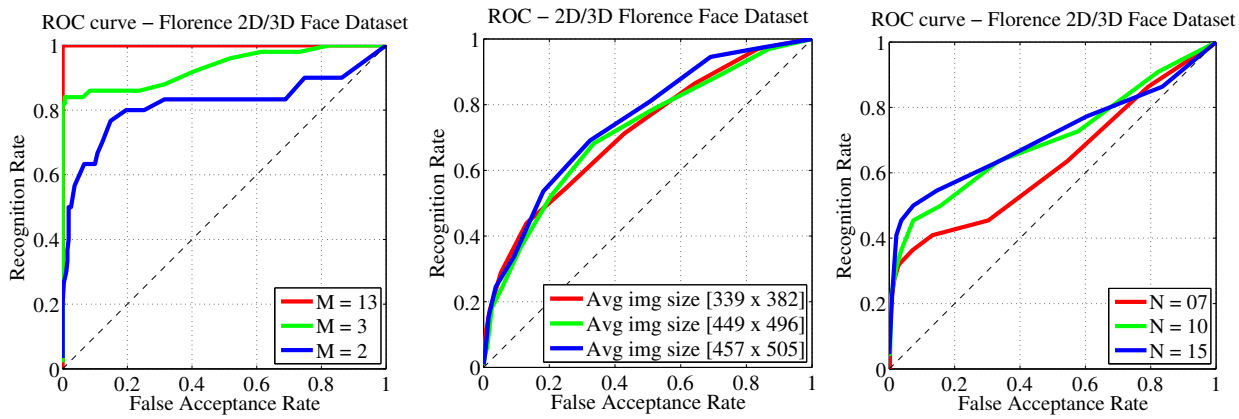

Figure 5. ROC curves showing performance on the Florence 2D/3D Face Dataset: each curve represents a result in function of (a) number of images (M) per subject in the gallery when recognizing 2D rendered images (b) the image size when recognizing face imagery from a PTZ camera (c) and of the number of images (N) present in the probe when recognizing face imagery from a PTZ camera

probe show that the approach generalizes to synthetic imagery as well. Experiments on recognizing real 2D face imagery using rendered gallery images show promising results, particularly when incorporating multiple probe images per subject.

Our ongoing work is related to determining the best face images to extract from PTZ sequences and quantifying more conclusively how performance is affected by varying face resolution and quality. We are also looking at better ways of structuring SIFT descriptors in the gallery (for example according to pose) and of structuring sparse solutions in discriminative ways (for example using the group lasso).

References

[1] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino. 2D and 3D face recognition: A survey. *Pattern Recogn. Lett.*, 28(14):1885–1906, Oct. 2007. 2

[2] A. D. Bagdanov, A. Del Bimbo, F. Dini, G. Lisanti, and I. Masi. Posterity logging of face imagery for video surveillance. *IEEE Multimedia*, 19(4):48–59, oct-dec 2012. 3, 5

[3] A. D. Bagdanov, A. Del Bimbo, and I. Masi. Florence faces: a dataset supporting 2d/3d face recognition. In *Proc. of Int. Symposium on Communication Control and Signal Processing (ISCCSP)*, Rome, Italy, 2012. 4

[4] L. Bazzani, M. Cristani, A. Perina, and V. Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. *Pattern Recognition Letters*, 33(7):898 – 903, 2012. 5

[5] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:2003, 2003. 2

[6] K. I. Chang, K. W. Bowyer, and P. J. Flynn. Multi-modal 2d and 3d biometrics for face recognition. In *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 187–, 2003. 2

[7] E. Elhamifar and R. Vidal. Robust classification using structured sparse representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1873–1879, 2011. 2

[8] R. He, W.-S. Zheng, and B.-G. Hu. Maximum correntropy criterion for robust face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8):1561 – 1576, aug. 2011. 2

[9] J. Huang, V. Blanz, and B. Heisele. Face recognition with support vector machines and 3d head models. In *Workshop on Pattern Recognition with Support Vector Machines*, pages 334–341, 2002. 2

[10] C. Kang, S. Liao, S. Xiang, and C. Pan. Kernel sparse representation with local patterns for face recognition. In *ICIP*, pages 3009–3012, 2011. 2

[11] G. Little, S. Krishna, J. Black, and S. Panchanathan. A methodology for evaluating robustness of face recognition algorithms with respect to variations in pose angle and illumination angle. In *International Conference on Acoustics, Speech, and Signal 2005*, volume 2, pages 89–92, 2005. 4

[12] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, 11:19–60, Mar. 2010. 3

[13] J. Marn, D. Vzquez, D. Gernimo, and A. M. Lpez. Learning appearance in virtual scenarios for pedestrian detection. In *CVPR*, pages 137–144, 2010. 2

[14] P. Phillips, W. Scruggs, A. O'Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale experimental results. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):831–846, 2010. 1

[15] P. A. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. 3

[16] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210 –227, feb. 2009. 2, 3

[17] X. Zhang and Y. Gao. Face recognition across pose: A review. *Pattern Recogn.*, 42(11):2876–2896, Nov. 2009. 2