FACE RECOGNITION IN REAL-WORLD IMAGES

Xavier Fontaine *Radhakrishna Achanta* *Sabine Süsstrunk*

School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

Face recognition systems are designed to handle well-aligned images captured under controlled situations. However real-world images present varying orientations, expressions, and illumination conditions. Traditional face recognition algorithms perform poorly on such images. In this paper we present a method for face recognition adapted to real-world conditions that can be trained using very few training examples and is computationally efficient. Our method consists of performing a novel alignment process followed by classification using sparse representation techniques. We present our recognition rates on a difficult dataset that represents real-world faces where we significantly outperform state-of-the-art methods.

Index Terms— Face recognition, sparse representation, alignment, mesh warping, facial landmarks.

1. INTRODUCTION

Face recognition is probably one of the most prominent areas of research in imaging and has a wide range of real-world applications including surveillance, access control, identity authentication [1], and photo-management. Face recognition systems either perform *face verification*, *i.e.,* classify a pair of pictures as belonging to the same individual or not, or perform *face identification*, *i.e.,* put a label on an unknown face with respect to some training set. In this paper, we address the latter problem of face identification.

During the last thirty years automatic face recognition has seen considerable progress. Despite this, face recognition is a very challenging problem when the training examples are few and the conditions of capture are unconstrained, resulting in face images varying widely in orientation, expression, and illumination.

In our work, we focus on the difficult problem of recognizing faces captured in uncontrolled environments. We impose additional constraints on the number of training samples and on computational efficiency without needing any specialized hardware. This rules out deep learning approaches which are data and computation hungry.

Our contribution is a face recognition scheme that performs automatic face alignment and recognition of detected faces with high accuracy and speed without employing any specialized hardware or parallel processing. Given detected faces and their landmarks, we present an algorithm to align these faces and use a modified version of a state-of-the-art algorithm to recognize faces. Our algorithm is able to identify pictures in almost real-time on a simple PC. Except for deep learning based schemes, our method outperforms all other schemes we are aware of in terms of recognition accuracy.

The rest of this paper is organized as follows: Section 2 briefly reviews prominent state-of-the-art methods for face recognition, Section 3 presents our approach, Section 4 shows the results of our experiments on datasets for face recognition, and Section 5 concludes the paper.

2. PREVIOUS WORK

The initial methods developed for face recognition used individual features on the faces, such as eyes, mouth or nose to perform identification [2]. However such methods did not lead to good results because of the variability of poses and the low amount of information used.

From the 90s, new methods that use global features of the faces were developed. For example, Turk and Pentland proposed EigenFaces [3] that uses Principal Component Analysis (PCA). Other methods like Fisherfaces [4] or Laplacianfaces [5] extract features from face images and perform nearest neighbor identification using Euclidean distance measure. Baback *et al.* [6] use a bayesian approach where a probabilistic similarity measure is used to perform classification.

Wright *et al.* applied the ideas of sparse coding to face recognition: they proposed the Sparse Representation based Classification (SRC) scheme [7], a dictionary learning based approach to recognize faces. This method, which can be seen as an improvement over the previous ones, is far more robust and is able to handle occlusions and corruption of face images. The SRC algorithm led to other approaches [8, 9, 10, 11] that use sparsity and improve the robustness in dealing with face alignment and pose variation issues.

Following the success of the use of sparsity in face recognition Zhang *et al.* [12] questioned if sparsity was the key to the success of the SRC algorithm. They concluded that it is the use of collaborative representation (*i.e.,* using an overcomplete dictionary) and not the sparsity constraint that improves

face recognition performance. This led to the development of other algorithms like MSPCRC or PCRC [13] that use this idea of collaborative representation.

Blanz and Vetter [14] use 3D scans of heads to learn a 3D model of faces and fit this model to 2D images of faces. Classification is done using the 3D representation. The advantage of such a method is its independence to face orientation or illumination. For instance, Zhu *et al.* [15] use 3D morphable models to eliminate pose and expression variations.

In the recent years, deep learning methods [16, 17] have been adapted to the face recognition problem. These methods achieve very good recognition rates and clearly outperform the "standard" algorithms. However they generally require a considerable amount of data and specialized hardware to train and deploy in practice. This makes them hard to train and less suited for embedded and low power devices.

3. OUR APPROACH

In order to be robust, computationally efficient, and use minimal training, we choose to use the Robust Sparse Coding (RSC) algorithm [9] with modifications. However, since the RSC algorithm represents an input image as a linear combination of the training images, all images used in this approach should be of similar size, well-aligned, and fully-frontal. It is mandatory to align the face images prior using RSC on real-world images. We thus describe our novel approach for automatically aligning faces and then the application of RSC on them for identification.

3.1. Automatic face alignment

To align a given face to a reference, we mesh-warp the input face image. The goal of mesh warping is to deform the input image to match its features with the corresponding features of a reference image based on a triangulation mesh. This is done in three steps - detection of face, detection of facial landmarks, and face warping.

First we detect faces on the picture using the Viola-Jones approach [18]. Then we detect landmarks, i.e. particular points on the face that are present in all face images and whose correspondences are supposed to be preserved, on the faces. For facial landmark detection we use the regression tree method of Kazemi and Sullivan [19] implemented in the Dlib [20] library. The 68 detected landmarks are mainly located on the eyes, nose, and mouth as shown in Fig. 1d.

Apart from the 68 detected landmarks, we add equally-spaced points on the border of the face square. This allows us to compute the Delaunay triangulation mesh that covers the entire face image, as in Figure 1b. For the landmarks of each input face image we copy this reference face triangulation mesh. We now simply need to warp each of these triangles to map to the corresponding triangle on the reference face mesh. This is done by an affine transformation consisting of a rotation, a scaling, and a translation in order to map a point $[x\ y]^T$ on the input face to a point $[x'\ y']^T$ on the reference face as:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{1}$$

where a, b, c, d are the rotation and scaling parameters while t_x, t_y are the translation parameters. The warped image thus obtained is now aligned such that the inter-eye distance and chin-eye distance are roughly the same in all images. This is done with the following transformations:

1. Rotation to force the eyes to be horizontally aligned.
2. Rescaling to obtain fixed inter-eye and eye-chin distances.
3. Translation to set the position of the left eye to a fixed predefined value.
4. Cropping the relevant 30×30 pixel part of the face image.

This completes the face alignment process. The faces for training are pre-processed as explained above and stored. In order to perform identification, every candidate image is subject to the same alignment process. The mapping of all the triangles to the reference results in an image whose landmark points coincide with the landmarks of the reference image. This deformation results in a kind of "frontalization" of the input image, as shown in Figures 1e and 2c. The goal of this phase is not to obtain a visually accurate version of the input image, but to prepare the image for the recognition step. Notably, there exist automatic face alignment techniques [21, **?**] but they are computationally too expensive to allow real-time applications. Our pre-processing, as described in Section 3.1, takes roughly 0.1s per image with our Python implementation, leaving ample room for an even more efficient implementation.

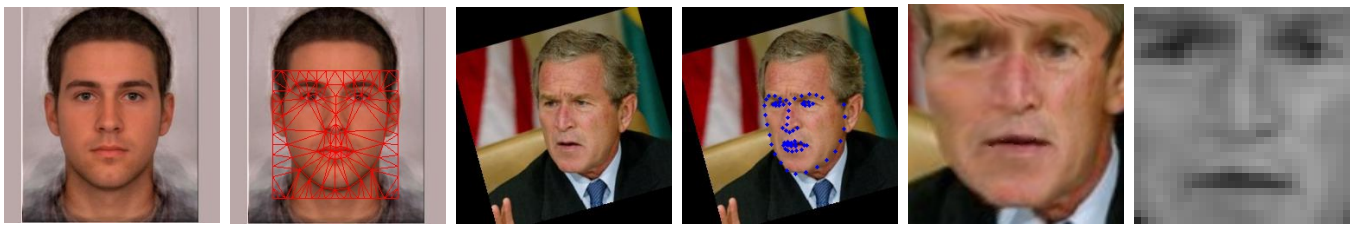
3.2. Face recognition

We use a modified version of the Robust Sparse Coding (RSC) algorithm [9] for recognizing faces. The RSC algorithm is an improvement of the SRC [7] method. The SRC method creates a dictionary matrix D containing all training images. For a given unknown image y the goal is to find a vector of weights x such that $y = Dx$. By using the ℓ^1-norm of x as the regularizer, x is forced to be sparse.

RSC [9] differs from SRC in the use of a diagonal weight matrix W for improved robustness to occlusions and lighting changes, and in the use of a Maximum Likelihood Estimator to solve the sparse coding problem. In summary, the RSC algorithm solves the weighted-LASSO problem:

$$\min_x \left\| W^{1/2}(y - Dx) \right\|_2^2 \text{ s.t. } \|x\|_1 \leqslant \varepsilon \tag{2}$$

[1] Source: http://lanimeshvariousarticles.blogspot.fr/2014/05/the-science-of-sex-appeal-face.html



(a) Reference face [1] (b) Triangle mesh (c) LFW sample (d) Face landmarks (e) Mesh warping (f) Aligned image

Fig. 1: *The steps of alignment using our method on LFW sample image of G.W. Bush. Output (f) is cropped and resized after alignment.*

where ε is a constant representing the noise level. If we use the ℓ^2-norm instead of the ℓ^1-norm we obtain the regularized least squares problem

$$\min_x \left\| W^{1/2}(Dx - y) \right\|_2^2 + \lambda \|x\|_2^2 \tag{3}$$

with $\lambda > 0$, which has the analytical solution

$$x = (D^T W D + \lambda I)^{-1} D^T W y \tag{4}$$

Even though this solution requires inverting a matrix of size $\#(\text{training\_samples}) \times \#(\text{training\_samples})$, it is computationally much more efficient than the ℓ^1-regularized version. On the AR database [22], using the procedure described in [9], we obtain a recognition rate of 95.0% in $2.4s$ with the ℓ^1-norm and a rate of 94.1% in $0.6s$ with the ℓ^2-norm. That is, using the ℓ^2-norm we lose less than 1% in accuracy for a 4-fold speed-up. We thus use the modified ℓ^2-norm version of the RSC algorithm.

4. RESULTS

On well-aligned images of the AR dataset [22], our modified RSC algorithm achieves 95.0% recognition rate. So, for our experiments we use the more challenging Labeled Faces in the Wild (LFW) database [23, 24], which consists of unaligned real-world images.

The full LFW dataset contains more than 13,000 images of 5749 individuals in unconstrained environments. Among these, 158 individuals have at least 10 distinct images. To be able to compare our method with the state-of-the-art we use the LFWa version [25] of the dataset, which consists of the LFW dataset images that are pre-aligned using a commercial alignment software. A few such images can be seen in Fig. 2a. Despite the pre-alignment, these images are not well suited for RSC [9] based recognition. As we note later in Table 1, our alignment proves to be more effective for improving recognition rates.

In order to prove the effectiveness of our method, for the 158 classes we randomly select 7 images for training and 3 for testing we run the following three recognition experiments:

1. Applying RSC algorithm on the original LFWa dataset.

(a) Original image (b) Triangulation (c) Warped

Fig. 2: *Mesh warping examples on pictures of David Beckham, Tony Blair, Gordon Brown, Angelina Jolie and Hu Jintao. Despite the distortions introduced by our warping, the resulting aligned images of column (c) are better suited for recognition than the original images of column (a). The results in Table 1 prove this.*

2. Applying RSC algorithm on the faces detected on the LFWa images.
3. Applying our modified RSC algorithm, after performing our alignment step on the LFWa images.

The images from the LFWa dataset are 250×250 pixels in size. For the first experiment, we resize them to 50×50 without any other modifications. For the second experiment we detect faces in the LFWa images and we resize the face region of the images to size 50×50. In the third experiment, which corresponds to our algorithm, the final aligned images are of size 30×30. The results are summarized in the Table 1.

	Exp. 1	Exp. 2	**Exp. 3**
Recognition rate (%)	19.6	28.8	**76.4**
Time for one image (s)	3.2	3.0	**1.6**

Table 1: *Results on the LFWa database with 7 training images and 3 test images.*

The results presented in the Table 1 show clearly that the alignment phase performed with our method is essential to obtain good recognition rates. We obtain scores that are nearly 4 times better than those with the use of raw LFWa images while simultaneously halving the runtime. This demonstrates the effectiveness our alignment before running face recognition algorithms on real-world images.

We also compare our method with recent state-of-the-art [7, 12, 13] that present benchmark results on the LFWa dataset using only 2 and 5 training samples. Table 2 shows this comparison.

Method	2	5
NN	$9.3 \pm 1.7\%$	$14.3 \pm 1.9\%$
SRC [7]	$24.4 \pm 2.4\%$	$44.1 \pm 2.6\%$
CRC [12]	$27.4 \pm 2.1\%$	$42.0 \pm 3.2\%$
MSPCRC [13]	$35.0 \pm 1.6\%$	$41.1 \pm 2.8\%$
Ours	$\mathbf{51.1 \pm 2.9}\%$	$\mathbf{74.2 \pm 2.5}\%$
Our time	*0.15 s*	*0.85 s*

Table 2: *Recognition rates on the LFWa dataset for different methods and with 2 and 5 training samples per person using settings as proposed in [13].*

We finally run experiments with one unique training image per person. This setting is quite extreme because the identity of a person is only determined by a single image and affects the robustnesss of the algorithms. However, using such a small number of training images may be inevitable for many real life applications. The results of this comparison are presented in Table 3. Note that the recognition using single training samples is very fast since the size of the dictionary reduces significantly.

Method	Rate
NN	10.6%
SRC [7]	22.3%
ESRC [11]	26.7%
PCRC [13]	$25.0 \pm 1.8\%$
SVDL [10]	30.2%
Ours	$\mathbf{33.3 \pm 3.4}\%$
Our time	*0.02 s*

Table 3: *Recognition rates on the LFWa dataset for the extreme case of using a single training sample per person.*

For the Nearest Neighbor, SRC [7], ESRC [11] and SVDL [10] methods we took the results from Yang *et al.* [10], which correspond to the algorithms run with 2000 dimensions, which gives the best performance in their case. The PCRC values are taken from Zhu *et al.* article [13] which uses images resized to 80×80. For our simulations, we run our algorithm 10 times with different training and testing images in order to compute the means and standard deviations.

The results of our experiments (Tables 1, 2, 3) using the LFW dataset with a small number of training samples show that our method provides better recognition rates than other algorithms in near real-time. It is worth mentioning that deep learning methods perform considerably better, achieving recognition rates around 96% on the LFW dataset as explained in [17]. However, such methods need a large amount of training data (300000 images [17]) and powerful hardware to handle the computation involved. Our method, on the other hand, can perform fairly accurate recognition with less than 10 training samples in a very efficient manner.

5. CONCLUSION

We presented a computationally efficient face recognition technique for real-world images that requires less than 10 training examples and ordinary hardware to deliver near real-time recognition. Compared to existing state-of-the-art we nearly double the recognition rate while halving the computational runtime. We presented results on the LFW dataset that shows that our method significantly outperforms existing non-deep learning algorithms.

6. REFERENCES

[1] Michel Owayjan, Amer Dergham, Gerges Haber, Nidal Fakih, Ahmad Hamoush, and Elie Abdo, "Face recognition security system," in *New Trends in Networking, Computing, E-learning, Systems Sciences, and Engineering*, pp. 343–348. Springer, 2015.

[2] Woodrow Wilson Bledsoe, "Man-machine facial recognition," *Panoramic Research Inc.*, 1966.

[3] Matthew Turk and Alex Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, 1991.

[4] Peter N. Belhumeur, João P. Hespanha, and David J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[5] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang, "Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 328–340, 2005.

[6] Baback Moghaddam, Tony Jebara, and Alex Pentland, "Bayesian face recognition," *Pattern Recognition*, vol. 33, pp. 1771–1782, 2000.

[7] John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma, "Robust Face Recognition via Sparse Representation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 31, no. 2, pp. 210–227, February 2009.

[8] Andrew Wagner, John Wright, Arvind Ganesh, Zhou Zihan, Hossein Mobahi, and Yi Ma, "Towards a Practical Face Recognition System: Robust Alignment and Illumination by Sarse Representation," .

[9] Meng Yang, Jian Yang, and David Zhang, "Robust Sparse Coding for Face Recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 625–632, June 2011.

[10] Meng Yang, Luc Van Gool, and Lei Zhang, "Sparse variation dictionary learning for face recognition with a single training sample per person," in *The IEEE International Conference on Computer Vision*, December 2013.

[11] Weihong Deng, Jiani Hu, and Jun Guo, "Extended SRC: Undersampled Face Recognition via Intraclass Variant Dictionary," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1864–1870, September 2012.

[12] Lei Zhang, Meng Yang, and Xiangchu Feng, "Sparse representation or collaborative representation: Which helps face recognition?," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 471–478.

[13] Pengfei Zhu, Lei Zhang, Qinghua Hu, and Simon C. K. Shiu, "Multi-scale patch based collaborative representation for face recognition with margin distribution optimization," *ECCV*, 2012.

[14] Volker Blanz and Thomas Vetter, "Face Recognition Based on Fitting a 3D Morphable Model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, September 2003.

[15] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 787–796.

[16] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," 2015.

[17] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang, "DeepID3: Face Recognition with Very Deep Neural Networks," February 2015.

[18] Paul Viola and Michael Jones, "Robust real-time object detection," in *IJCV*, 2001.

[19] Vahid Kazemi and Josephine Sullivan, "One Millisecond Face Alignment with an Ensemble of Regression Trees," in *CVPR*. 2014, IEEE Computer Society.

[20] Davis E. King, "Dlib-ml: A Machine Learning Toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[21] Gary B. Huang and Vidit Jain, "Unsupervised joint alignment of complex images," in *ICCV*, 2007.

[22] A.M Martinez and R. Benavente, "The AR Face Database," Tech. Rep. 24, CVC, June 1998.

[23] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[24] Gary B. Huang Erik Learned-Miller, "Labeled Faces in the Wild: Updates and New Reporting Procedures," Tech. Rep. UM-CS-2014-003, University of Massachusetts, Amherst, May 2014.

[25] Lior Wolf, Tal Hassner, and Yaniv Taigman, "Similarity scores based on background samples," *Asian Conference on Computer Vision*, September 2009.