Learning a bag of features based nonlinear metric for facial similarity

Grégoire Lefebvre, Christophe Garcia

Learning a Bag of Features based Nonlinear Metric for Facial Similarity

Grégoire Lefebvre
Orange Labs, R&D
Meylan, France
gregoire.lefebvre@orange.com

Christophe Garcia
LIRIS, UMR 5205 CNRS, INSA-Lyon
F-69621, France
christophe.garcia@liris.cnrs.fr

Abstract

This article presents a new method aiming at automatically learning a visual similarity between two images from a class model. This kind of problem is present in many research domains such as object tracking, image classification, signing identification, etc. We propose a new method for facial recognition with a system based on non-linear projection and metric learning. To achieve this objective, we feed a "Bag of Features" representation of the face images into a specific neural network that learns a mapping to a more compact and discriminant representation. This learning process aims at non-linearly projecting the facial features into a reduced space where two images belonging to the same category (i.e. a person) are "close" according to a given similarity metric and "distant" otherwise. The proposed method gives very promising results for face identification in adverse conditions like expression, illumination and facial pose variations. Experimental results give 97% *correct recognition rate on the CMU PIE database containing* 68 *individuals, under vary variable pose and illumination conditions.*

1. Introduction

Due to a large number of possible applications like biometrics, video-surveillance or advanced human-computer interaction, face recognition systems received an increasing interest during the last decade. Lots of approaches have been proposed in the literature [27], but identifying human faces remains a challenging problem. The main difficulties are due to unconstraint illumination conditions, variable facial expressions and face poses. We can discern three categories of face recognition method: holistic matching methods using the whole face information, facial region-based methods (*e.g.* eyes, nose and mouth recognition) and "Bag of Facial Features" representation [16]. In this article, we focus on methods from the last category by assuming that the local biometric description is salient whatever view is considered to cluster the individual faces. For each salient description, we compute the local signal singularities. Classically, a clustering method is then applied before performing a dimension reduction of the feature vectors using a Self Organizing Map (SOM [14]). A facial feature similarity is then learned with a customized Siamese neural network that performs a non-linear projection of the facial features.

This paper is organized as follows. A brief state-of-the-art about face recognition is proposed in section 2. In section 3, we describe our face recognition strategy. Section 4 presents the experimental results with a comparison with standard state-of-the-art face recognition algorithms. Finally, conclusions and perspectives are drawn.

2. State-of-the-art

Facial similarity is a pattern recognition problem that has to be solved in a high-dimensional non-linear space. Classical approaches propose a dimensional reduction technique to solve the recognition problem in a lower dimensional feature space. In particular, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) have been deeply studied for face recognition over the last decade. The features encoded by Eigenfaces (using PCA [25]) and Fisherfaces (using LDA [5]) are based on the second order dependencies and these methods are linear projection methods.

In [25], the authors compute after a linearization step the mean vector, among all images, and subtracted it from all the vectors, corresponding to the original faces. The covariance matrix is then computed, in order to extract a limited number of its eigenvectors, corresponding to the greatest eigenvalues. These eigenvectors, referenced as Eigenfaces, describe a low-dimensionality space base. Face images are then projected in this EigenSpace and classically compared using Euclidian or L1 distances. The LDA approach [5] offers an alternative taking into account the individual classes.

PCA is still a preliminary step reducing the input space, and then LDA is applied in order to maximize the between-class covariances, minimizing the within-class ones.

Other techniques relying on LBPH (Local binary Pattern Histograms [1]) or SIFT (Scale Invariant Feature Transform [20]) use a local texture-based classification. The facial image is divided into local regions and descriptors are extracted from each region independently. The global classification is then performed by a combination of local and global similarity.

The main disadvantage of these projection techniques is their linearity. Bartlett et al. show in [3] that first- and second order statistics capture information only the amplitude spectrum of an image, discarding the phase-spectrum. And yet the human capability seems driven by the phase-spectrum to recognize faces. A further nonlinear solution to the face recognition problem is provided by neural network based approaches. Independent Component Analysis (ICA [3]) is derived from the principle of optimal information transfer through sigmoidal neurons. The main idea is to build a neural network with a neuron for every pixel in the image that captures discriminant features not necessarily orthogonal, exploiting the covariance matrix and considering the high-order statistics. The advantage of neural classifiers over linear ones is that they can reduce misclassifications among the neighborhood classes. For example, Self Organizing Map (SOM) is invariant to minor changes, while Convolutional Neural Networks (CovNets) provide a partial rotation, translation and scaling invariance.

Hereafter, our proposal combines the advantages of local description clustering and neural network to build a very efficient system.

3. Our Face Recognition System

3.1. Facial Identification

Our face recognition system is mainly divided into three steps: facial feature extraction, facial pattern clustering, and facial feature vector classification. The first step detects perceptually relevant points in the face image. The second step computes local feature vectors and clusters each vector into a neural activity histogram built from Self-Organizing Map winning neurons, as described by Lefebvre *et al.* in [16]. For the third step, we propose a non-linear mapping strategy based on Siamese network [6, 7] to classify facial feature vectors. These methods are described in the next sections.

3.2. Feature Extraction, Description and Clustering

In the literature, feature extraction, image description and data clustering may be performed with numerous methods. Classically, point detectors aim at extracting local image features (*e.g.* corner detection [11], blob detection[17], edge and ridge detection [18], *etc.*) while local image descriptors computes invariant features to geometric transforms [19, 15, 4, 23]. As described in Lefebvre *et al.* [16], our system focuses on wavelet analysis to detect salient point and compute singularity descriptors. Using wavelets is justified by the consideration of the human visual system for which multi-resolution, orientation and frequency analysis are of prime importance according to Hoffman *et al.* [12]. For clustering all facial feature vectors from all individuals, a global Self Organized Map (SOM) is learned. Even many clustering approaches exists (*e.g.* centroid models [21], connectivity models [8], fuzzy clustering [22], *etc.*), SOMs allow a relevant selection and learning process in order to capture facial information. The Kohonen model [14] constructs a neuron lattice in which the topology of the input space is preserved and each neuron is specialized in a stimuli set (*i.e.* a neighborhood function limits the neurons to respond to a given stimulus).

Let M be the input space and $X = x(t),\ t \in \{1, 2, \ldots, T\}$ be a facial feature vector set with $x(t) \in M \subset \Re^d$, where t is the time index. Lets assume that $M = m_i(t),\ i \in \{1, 2, \ldots, N\}$ is represented by the set of the reference SOM vectors with $m_i(t) \in \Re^d$, randomly initialized. For each input vector, the best matching unit (BMU) is defined $m_c(t)$ where:

$$c = \arg\min_i \|x(t) - m_i(t)\|, \forall i = 1, \ldots, N \qquad (1)$$

The topologic neuron map is then updated with the following equation:

$$m_i(t+1) = m_i(t) + \lambda(t)\phi_c^{(i)}(t)\left(x(t) - m_i(t)\right) \qquad (2)$$

where $\lambda(t)$ is the learning rate and $\phi_c^{(i)}(t)$ a neighborhood function.

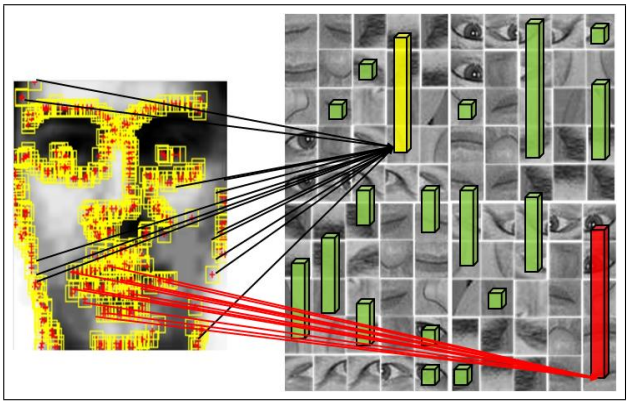

Figure 1. SOM activation histogram.

When the SOM is learned, a "Bag of Features" is composed of the best representative vectors from the learning database. Each identity vector is then described by the stimulation of the SOM winning neurons with the individual

facial feature vectors. This SOM activation histogram H is defined by Equation 3 and illustrated in Figure 1.

$$H[c] = \sum_{t=1}^{T} \|m_c(t) - x(t)\|, \forall c = 1, \ldots, N \qquad (3)$$

3.3. Siamese Neural Networks

In order to overcome the issues of classification in high dimensional spaces (*i.e.* curse of dimensionality: *e.g.* the facial feature vector dimension is here the SOM size), a large number of dimensionality reduction techniques have appeared over the last decade. As previously mentioned, one can cite the classical PCA and LDA methods. The first method performs a linear projection in a space of reduced dimensions, where the variances of the original data are maximized. The second method performs also a linear projection and aims at maximizing inter-class variances while minimizing intra-class variances. In order to enhance the robustness of such linear approaches versus non-linear variations of the data, kernel versions of these algorithms have been investigated (*i.e.* KPCA, KLDA [26]). Other non-linear manifold learning methods have recently appeared, like Locally Linear Embedding (LLE) [24], Isomap [10], and Laplacian Eigenmaps [13]. These non-linear dimensionality reduction techniques aim at optimally preserving the local geometry around each data as well as the global structure of the data but we need here to define and learn a visual similarity metric from the data.

Siamese Neural Networks have first been presented by Bromley *et al.* [6] using Time Delay Neural Networks (TDNN) and applied to the problem of signature verification. Siamese neural networks learn a non-linear similarity metric by repeatedly presenting pairs of positive and negative examples, *i.e.* pairs of examples belonging to the same class or not. The principal idea is to train the neural network to non-linearly map the input vectors into a subspace such that a specific metric in this subspace approximates, not the local geometry like in the methods cited above, but the "semantic" distance in the input space. Two examples of the same category are supposed to yield a small distance in this subspace and two examples of different categories a large distance.

Let us call this mapping $G_W(X)$ and its parameters W(*i.e.* neural weights). Thus, the goal is to learn the parameters W of the function $G_W(X)$ such that the similarity metric (see Equation 4) is small if X_1 and X_2 belong to the same class and large otherwise.

$$E_W(X_1, X_2) = \|G_W(X_1) - G_W(X_2)\|_2 \qquad (4)$$

The choice of $G_W(X)$ is arbitrary and, in our approach, is given by a Multilayer Perceptron (MLP). Note that the parameters W are shared by the neural networks(hence the name "Siamese" neural network) and therefore the distance metric is generally symmetric. Figure 2 illustrates the functional scheme of this learning machine.

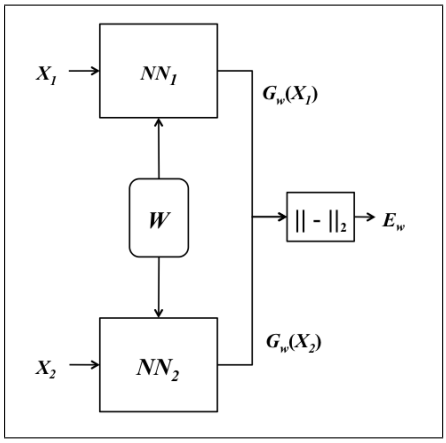

Figure 2. A Siamese Neural Network.

3.4. Similarity Distance

This idea was then adopted by Chopra *et al.* [7] who used Siamese ConvNets in the context of face verification. More precisely, the system receives two face images and has to decide if they belong to the same person or not. The ConvNet allows learning to extract features using a cascade of convolution and subsampling filters. They use the similarity (Euclidean distance) E_W that they try to minimize for genuine pairs of facial features and maximize for impostor pairs. One of the problems of this energy function is that, if minimizing E_W to zero for genuine pairs is tractable by error backpropagation, on the other hand, one cannot decide which distance should be the target for impostor pairs. The problem could be overcome by minimizing another energy function designed for impostor pairs, $L(E_W)$ where L is a monotonically decreasing function, which is difficult to choose.

In our proposal, instead of extracting features with ConvNets, we represent the facial feature vector H extracted from SOM activations (cf. Equation 3). Then, we apply a Siamese one-hidden-layer MLP which projects the input feature vectors H into a vector of smaller dimension $G_W(H)$. More formally, given a triplet (H, H_+, H_-) such that X is a facial feature vector, H_+ belongs to the same class as H and H_- belongs to another class, we would like the scalar product of the similar ones to be higher than that of the dissimilar ones: $G_W(H).G_W(H_+) > G_W(H).G_W(H_-)$. We therefore define the following objective functions to minimize in order to be able to apply supervised learning via classical backpropagation:

$$E_W(H) = (1 - cos(G_W(H), G_W(H_+)))^2 + (0 - cos(G_W(H), G_W(H_-)))^2 \quad (5)$$

Minimizing E_W can be interpreted as searching for a projection G_W (estimating W in the neural network) such that facial feature vectors H and H_+ are collinear and facial feature vector H and H_- are orthogonal. This choice is also motivated by the better performance of the cosine distance in PCA and LDA space compared to Euclidean distance.

4. Experimental Results

We evaluated the proposed method on the CMU Pose, Illumination, and Expression (PIE) face database [2]. It contains 41,368 images of 68 people, each person being pictured under 13 different poses, 43 different illumination conditions, and with 4 different expressions (*c.f.* Figure 3). All faces are automatically extracted using the face detector proposed in [9] and resized to 200×200 pixels. In order to assess the system performances, we use a 3-fold cross validation method in the following experiments. For each fold, the database is divided into three datasets: one third of the images for learning, one third for validation and one third for test. Among the different datasets, we respect the proportion per individual of illumination, expression and pose images.

Figure 3. Some face samples from the CMU PIE database with different poses, illumination conditions and expressions.

For each face image of size 200×200 pixels, the facial feature vectors are extracted from 1000 local signatures, clustered in a squared SOM composed of N_s x N_s neural units. Each facial feature H is therefore the activation results of the SOM and is of size N_s x N_s. The best configuration, i.e. the choice of N_s, is determined by a 3 cross-validation on the training and validation databases, as shown in Figure 4. As shown by this figure, the best result of 89.85% is obtained with $N_s = 30$, but a good compromise between correct recognition rate and feature vector size is for N_s = 20 giving a SOM size of 20x20 $= 400$ with 88.12% and a standard deviation of 0.47.

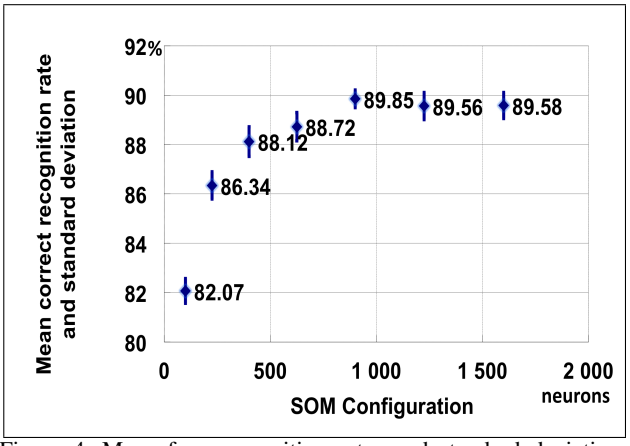

Figure 4. Mean face recognition rates and standard deviations (bars) for different SOM size configurations.

Concerning the non-linear projection with Siamese MLP, experiments have been performed to choose a correct configuration with the facial feature vectors H of size 400. We consider a two-layer MLP architecture with a varying number of neurons in each layer (N_h sigmoid neurons in the hidden layer and N_o linear neurons in the output layer). For each configuration, the classical backpropagation algorithm is applied on the training set, and stopped when the mean square error between the actual output and the target outputs starts to grow on the validation set (with early stopping strategy), so that overlearning is avoided.

Figure 5 reports the recognition rates obtained for different configurations which have been chosen to perform a non-linear projection into a smaller target space: (N_h,N_o)=(300,200), (200,100), (100, 50), (50,25). One can notice that the best recognition result is obtained for (N_h,N_o)=(100,50), with corresponds to a MLP Siamese Network with one hidden layer of 100 sigmoid neurons and an output layer of 50 linear neurons. Therefore, in the final scheme, each facial feature vector of size 400 is projected into a 50-dimensional space, where the similarity metric (i.e. the cosine distance) is applied.

For each of the three folds, the trained system is then evaluated on the test dataset. Table 1 presents our results for each fold with a k nearest neighbor classification (here $k = 1$) directly from the SOM activation histograms H (*i.e.* SOM-1NN) and after non-linear projection with the Siamese MLP and the application of the cosine similarity metrics (*i.e.* SOM-SIAM).

For comparison, we have also implemented three standard face recognition methods: Eigenfaces [25], Fisherfaces [5] and Local Binary Pattern Histograms (LBPH) [1] (see section 2).

These results show first that SOM-1NN is efficient gi-

Methods	Fold 1	Fold 2	Fold 3	Mean	Standard Deviation
Eigenfaces	54.65%	55.97%	59.22%	57.60%	(2.30)
Fisherfaces	87.26%	81.19%	83.63%	84.03%	(3.05)
LBPH	89.30%	89.97%	89.97%	89.75%	(0.39)
SOM-1NN	90.35%	89.03%	90.17%	89.85%	(0.72)
SOM-SIAM	**97.27%**	**96.32%**	**97.07%**	**96.89%**	**(0.50)**

Table 1. Recognition rates, means and Standard Deviations (SD) on the CMU PIE database.

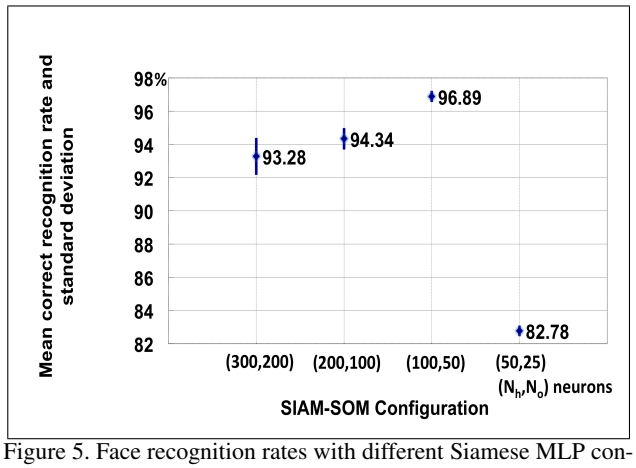

Figure 5. Face recognition rates with different Siamese MLP configurations.

Figure 6. Some misclassified face samples from the CMU PIE database using the method SOM-SIAM.

ving a correct recognition rate of around 90%. This first result overcomes the classical Eigenfaces, Fisherfaces and LBPH methods. This observation proves that the database is composed of many variations in illumination, expression and viewpoints that linear projection methods can not deal with. The LBPH method also outperforms the Eigenfaces and Fisherfaces approaches. This texture classification is built from predefined facial regions, consequently different viewpoints are difficult to handle. Within the SOM feature extraction, point detection allows us to cluster salient facial information among all views. Moreover, the application of the Siamese MLP in our method SOM-SIAM that learns how to project the facial feature vectors in order to maximize the similarity cosine distance gives a gain of around 7% to reach around 97% of correct recognition rate. These results show that the proposed method is able to handle strong variations in pose, illumination and expression. Nevertheless, the Figure 6 shows some misclassified examples from the CMU PIE database using SOM-SIAM.

5. Conclusion and Perspectives

In this paper, we have proposed a novel face recognition method using a neural non-linear projection scheme. Based on the two main properties of SOM, which are dimension reduction and topology preservation, this architecture features all facial identities by neural activity counts. In order to quantify the visual similarity between two face images, a non-linear metric is directly learned, via a Siamese neural network that searches for an optimal projection that clusters facial feature according to "semantic" order instead of geometric distances. The proposed solution gives very promising results on a difficult face dataset. As an extension of this work, we plan to develop a strategy to optimally learn the hyperparameters of the proposed system related to the architectures of the SOM and the Siamese MLP.

References

[1] Ahonen, T., Hadid, A., and Pietikainen, M. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):2037 –2041, dec. 2006.

[2] Baker S. and Bsat M. The cmu pose, illumination, and expression (pie) database. *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–51, 2002.

[3] Bartlett M.S., Movellan J.R., and Sejnowski T.S. Face recognition by independent component analysis. *IEEE*

Transactions on Neural Networks, 13(6):1450–1464, 2002.

[4] Bay H., Tuytelaars T., and Van Gool L.J. Surf: Speeded up robust features. In *ECCV*, pages 404–417, 2006.

[5] Belhumeur P.N., Hespanha J., and Kriegman D. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, 1997.

[6] Bromley J., Guyon I., LeCun Y., Sackinger E., and Shah R. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(4):669–688, 1993.

[7] Chopra S., Hadsell R., and LeCun Y. Learning a similarity metric discriminatively, with application to face verification. *ICCVPR. IEEE Press*, 2005.

[8] Fernández A. and Gómez S. Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. *Journal of Classification*, (25):43–65.

[9] Garcia C. and Delakis M. Convolutional face finder: A neural architecture for fast and robust face detection. *IEEE TPAMI*, 26(11):1408–1423, 2004.

[10] Geng X., Zhan D., and Zhou Z. Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Transactions on Systems, Man and Cybernetics*, (35):1098–1107, 2005.

[11] Harris C. and Stephens M. A Combined Corner and Edge Detector. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[12] Hoffman J.E. and Subramaniam B. The Role of Visual Attention in Saccadic Eye Movements. *Perception & Psychophysics*, pages 787–795, 1995.

[13] Jia P., Yin J., Huang X., and Hu D. Incremental laplacian eigenmaps by preserving adjacent information between data points. *PRL*, 30(16):1457–1463, 2009.

[14] Kohonen T. *Self-Organizing Maps*. Springer, 2001.

[15] Lazebnik S., Schmid C., and Ponce J. Semi-local affine parts for object recognition. In *British Machine Vision Conference*, pages 959–968, 2004.

[16] Lefebvre, G. and Garcia, C. Facial biometry by stimulating salient singularity masks. In *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 511–516, Sept.

[17] Lindeberg T. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *International Journal of Computer Vision*, 11:283–318, 1993. 10.1007/BF01469346.

[18] Lindeberg T. Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision*, 30:117–154, 1998.

[19] Lowe D.G. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60(2):91–110, 2004.

[20] Luo J., Ma Y., Takikawa E., Lao S., Kawade M., and Lu B.-L. Person-specific sift features for face recognition. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 2, pages II–593 –II–596, april 2007.

[21] Moody J.E. and Darken C. Fast Learning in Networks of Locally-Tuned Processing Units. *Neural Computation*, 1:281–294, 1989.

[22] Nock, R. and Nielsen F. On weighting clustering. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(8):1–13.

[23] Ros J., Laurent C., and Lefebvre G. A cascade of unsupervised and supervised neural networks for natural image classification. In *CIVR*, pages 92–101, 2006.

[24] Roweis S. and Saul L. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[25] Turk M. and Pentland A. Eigenfaces for recognition, 1991.

[26] M.-H. Yang. Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 215 – 220, may 2002.

[27] Zhao W., Chellappa R., Phillips J.R., and Rosenfeld A. Face recognition: A literature survey. *ACM Computing Surveys (CSUR)*, 35(4):399–458, 2003.