An Improved Face Verification Approach based on Speedup Robust Features and Pairwise Matching

Eduardo Santiago Moura, Herman Martins Gomes and João Marques de Carvalho
Center for Electrical Engineering and Informatics (CEEI)
Federal University of Campina Grande (UFCG)
Campina Grande, Paraíba, Brazil
Email: {edumoura,hmg}@dsc.ufcg.edu.br, carvalho@dee.ufcg.edu.br

***Abstract*—Human faces are known to present large variability due to factors like pose and facial expression variations, changes in illumination and occlusion, among others, thus making face verification a very challenging problem. In this paper we address the problem of face verification with special interest on how to reduce degradation usually associated with face images acquired under uncontrolled environments. The approach we propose in this paper starts with a preprocessing step to correct in-plane face orientation and to compensate for illumination changes. SURF features are then extracted, which adds to the method a certain degree of invariance to pose, facial expression and other sources of variation. Taking the SURF features as input, an original pairwise face matching procedure is performed. The resulting matching scores are stored in a similarity matrix, which is then evaluated. An experimental study has revealed that the proposed approach produced the best ROC curve when compared to published work regarding the unsupervised setup of the Labeled Faces in the Wild (LFW) [1] face database.**

***Keywords*-speedup robust features; pairwise matching; face verification; Labeled Faces in the Wild, unsupervised protocol**

I. Introduction

As a result of the popularization of the Internet (motivated in the recent years by the social networks like Facebook, Orkut, MySpace) and the proliferation of digital cameras and mobile devices, the automatic organization of large digital image collections has become an extremely relevant resource [2], [3], [4].

Traditional systems use only simple information (such as date, file and folder name) to help with the organization task [5]. However, for large collections, typically formed by millions of images, this information is insufficient to achieve good levels of organization and user satisfaction. Most advanced techniques in this area aim to analyze image content and to extract high level information, e.g., faces. In this sense, faces occupy a preponderant role, given their importance to human relations.

Therefore, within the scope of digital images containing faces, face analysis is a very relevant topic. Typical face analysis problems include detection and classification. This paper is not concerned with face detection. Face classification normally falls into three main categories: recognition, verification and clustering. In a recognition (or identification) task, the identity of a test face is inferred from a set of known identities. Facial expression recognition, gender or age classification among others, are also related recognition tasks. Verification, on the other hand, is a binary classification task that assigns *true* or *false* to a comparison (or matching) between a pair of faces. That comparison is frequently based on some similarity measure. The task is to accept or deny the identity claimed by a person. Finally, face clustering aims to divide a collection of faces into a finite set of groups or clusters. Some clustering methods can operate on the similarity measures returned by face verification checks on all pairs of faces extracted from the collection.

In this paper we address the problem of face verification with special interest to the following aspects: (i) to propose an approach to reduce degradation usually associated with face images originated from lighting, facial expressions and pose variations, among other possible sources; and and (ii) to compare the proposed approach with state-of-the-art techniques on a reference face image database.

Facial features have been used by most techniques for identifying, verifying and grouping together all images of a same person in large collections. Facial feature extraction usually is preceded by face detection and consists of applying extractors on the detected face region in order to obtain a unique representation.

Cao et al. [5] have argued that the most critical decisions to be taken when designing an automatic system for grouping similar faces, consists of choosing the facial representation and defining a metric for face comparison. Unsupervised recognition is adequate to this problem, as usually no previous knowledge about people in the photos is available to the system.

Human faces are known to present large variability due to factors like pose and facial expression variations, changes in illumination and occlusion, among others, thus making face verification a very challenging task [6]. Since, as discussed above, this task can be seen as a binary classification problem on pairs of input face images, a global decision boundary must be found, which makes face verification usually harder to perform than face identification. Also, both extracted features and the employed correspondence scheme must be robust against all sorts of variation.

Many researchers have been attracted to the problem of facial verification in uncontrolled settings [7], [8], [9], [10], [5]. The existence of reference image databases, such as the Labeled Faces in the Wild (LFW) [1] database provides support to the investigation of this problem, and makes it possible to evaluate and compare the performance of different methods. LFW is a photo collection that contains over 13,000 face images from 5,749 people, extracted from electronic and printed news media (Yahoo!, News, etc.) presenting a large natural variation of pose, lighting, focus, resolution, expression, age, sex, ethnic group, accessories, occlusion and image quality. Three evaluation protocols are available for LFW: 1) image restricted training setting, 2) image unrestricted training setting, and 3) unsupervised setting.

Facial representation schemes based on local descriptors have proven to be adequate for the face verification problem. Among those are the Local Binary Patterns - LBP [11] and its variants Three-patch LBP - TPLBP [7] and Four-patch LBP - FPLBP [7], as well as Scale-Invariant Feature Transform SIFT [12] and Histogram of Oriented Gradient HOG [13], among other techniques based on descriptor histograms. Those schemes typically employ quantized local patterns as well as quantized image gradient codes to describe local geometric structures.

In the present work we focus on face verification in a unsupervised setting, with the LFW base. Our method consists of a preprocessing stage followed by feature extraction stages, using Speeded Up Robust Features SURF [14] as visual descriptor. An improved strategy for feature comparison is also proposed.

In Section 2 we present an overview of the state of the art on facial verification techniques. Section 3 details the main techniques used in the proposed approach. In Section 4 our proposed method is described in detail. In Section 5, we present some implementation details and, in Section 6, the experimental evaluation performed and the results obtained using the LFW base are described. We present our final conclusions in Section 7, along with the main contributions of the work and some ideas for future work.

II. Related work

Although this paper is focused on face verification, the review of related work presented in this section also includes work on face recognition and clustering, since they posses many representations in common.

Humans automatically and naturally employ features like eyes, mouth, nose and hair to recognize, verify the identity or to group other people. Shape, position, size and distances of and between these features can also be used to to describe and characterize a person. Face analysis systems usually work by extracting these and other information from a face image and relating them to a specific person.

Face representation schemes can be classified as either global or local. In a global scheme each component of the feature vector is related to the face as a whole, while, in local schemes, only a specific region is represented in each component. Although many successful methods for face representation have been proposed, determining the best representation for face verification is still an open problem [15].

Global face features have traditionally been very popular for face recognition. However, more recently, considerable effort has been made to develop face analysis systems based on local features, due to the robustness those features are believed to posses against illumination, occlusion and facial expression variations [15]. All the research developed so far has made evident that global and local features play different roles regarding face representation, which leads to the conclusion that combining them may be the best way to improve performance of systems for face recognition and verification.

Active Shape Models - ASM [16] have been used by Zhang et al. [17] to extract 68 fiducial points from faces and its features (eyes, nose and mouth) contours in frontal face images aiming to represent the corresponding shapes for face verification. Hausdorff distances to reference feature vectors are employed to measure similarity between face models. The lack of an objective evaluation in that work compromises the authors claims on the potential of the method.

The SIFT (Scale Invariant Feature Transform) descriptor, proposed by Lowe [12], extracts image features which are tolerant to shape, scale, point of view and illumination variations. SIFT works by extracting feature vectors from a neighborhood around keypoints, representing region orientation, scale and location. Both, Zhu et al. [18], and Wright and Hua [19] have used SIFT features for face tagging and face recognition, respectively. In the former work, L1 distance (which is translation and rotation invariant according to the authors), was used to measure similarity, while in the latter, a similarity metric called Inverse Document Frequency - IDF was utilized. Wright and Hua [19] report precision and recall rates of 97% and 86% respectively, obtained on the Gallagher images base [20].

Local features based on a 3x3 region neighborhood were used by Jayech and Mahjoub [21] to classify face images. For each region the features calculated were: average, standard deviation, energy, entropy contrast and homogeneity. The distance between images was calculated by the Tangent distance [22]. However, the experimental validation presented in the work lacks an objective evaluation metric, which compromises the authors claims about the merits of the method.

An objective evaluation was performed by Seo and Milanfar [6] for a system that utilizes a PCA based Locally Adaptive Regression Kernel (LARK) descriptor to represent faces. The LARK descriptor measures self-similarity from the geodesical distance between a region central point and its adjacent neighbors. Experiments performed on the unsupervised setting protocol of the LFW database reached an average accuracy of 72.23%. Other experiments performed on the image restricted training setting protocol of LFW reached a 78.90% classification rate.

Color and texture features, extracted by a set of filters were used for face clustering by Zhang et al. [23]. That set was originally proposed by Winn et al. [24] and is composed of three Gaussians, four first order Gaussian derivatives and four Laplacian of Gaussian (LoG) filters. Gaussian Mixture Models (GMM) were used for grouping features and to reduce noise.

Distance between sets of features was measured by the Earth Mover's Distance (EMD) [25] and experiments performed produced 99% precision (P) and 80% recall (R) rates. For comparison the authors provide the rates of commercial tools PICASA (100% P and 15% R) [26] and EasyAlbum (49% P and 9% R) [27] on the CMU image database [20].

Pose variation is one factor limiting the performance of many face recognition systems. Trying to overcome this factor, Prince and Elder [28] proposed a method called PLDA (Probabilistic Linear Discriminant Analysis) based on Fisherfaces [29]. Pose variations are modeled as noise added to the main frontal face representation (identity). Based on experiments from a previous work, the authors claim to have obtained good results with this method with error rate of 0.3%. By comparison, the same error rate was 33.9% and 11.9% for Principal Component Analysis - PCA [30] and Linear Discriminant Analysis - LDA, respectively, on the XM2VTS image database [31].

Unsupervised learning techniques like k-means, kd-tree and random projection tree [32] have been used by Cao et al. [5] for face recognition. The authors focused on the learning of uniform descriptors from a collection of low level histogram based descriptors alleging that uniformity is important when L1 and L2 distances are used as similarity metric. When using several learning descriptors on nine fiducial areas and a pose adaptive correspondence, a verification rate of 84.45% on the LFW base was obtained.

Freund et al. [32] tested several facial features that were employed by Wolf et al. [7]. Freund et al. [32] also added the SIFT descriptor, computed in fixed places of the face (corners of the mouth, eyes and nose), as the basis for a facial feature detector. The authors concluded that the SIFT based features produced better results (about 1% increase in the recognition rate) when compared to the descriptors considered by Wolf et al. [7]. In that work, LDML (Logistic Discriminant based Metric Learning) was employed, which is based on the learning of Mahalanobis metrics over a given spatial representation. The method considers a set of pairs of labeled images as training set.

Considering face recognition applications, Hua and Akbarzadeh [33] pointed out that Difference of Gaussians - DoG outperforms other more utilized methods for situations involving illumination changes. Their proposed method presented a 1.6% error rate on the ORL images base, better than other methods on the same base, like LDA (7.2%), Laplacian-Face [34] (6.8%), Spatially Smooth Fisherface - SLDA [35] (2.3%) and Regularized Fisherface - RLDA [36] (3.6%). On the LFW base a 78.64% recognition rate was reached by the method. The authors also proposed a metric called Robust Elastic and Partial Matching defined on the features space (i.e. on the local image descriptors space), like the Hausdorff distance.

Wright et al. [37] consider the face recognition problem as one of classifying linear regression models of multiple frontal faces with varying expression and illumination, as well as occlusion. Based on a sparse representation computed by 1-minimization, the authors propose a general classification algorithm for (image-based) object recognition. Based on experiments on the AR database [38] (recognition rate between 92.0% and 94.7%) and Extended Yale B database [39] (recognition rates between 92.1% and 95.6%) the authors claim that unconventional features such as downsampled images and random projections perform just as well as conventional features such as Eigenfaces and Laplacianfaces and the theory of sparse representation helps predict how much occlusion the recognition algorithm can handle and how to choose the training images to maximize robustness to occlusion.

A comprehensive study on image descriptors like PCA, LBP histograms, Gabor Jets [40], SIFT and Extremely Randomized Clustering Forest - ERCF [41] has been developed by del Solar et al. [42] for an unsupervised face verification setting (with no training). An comparison between the analyzed methods was performed by varying image cropping size, descriptors parameters and image block size. Average accuracy obtained with the LFW unsupervised setting protocol were 64.10% for SIFT, 69.45% for LBP histograms, 68.47% for Gabor Jets descriptors and 73.33% for ERCF. Although Gabor jets and ERCF presented the best accuracy they require a processing time that is too high for real-time applications.

From the above review, it is clear that there is a great variety of methods, experimental image databases and tasks regarding face classification. Thus, for comparison purposes, we selected the works of del Solar et al. [42] and Seo e Milanfar [6], which share the same unsupervised classification principle of our method and present experimental results adopting the LFW image database under the unsupervised protocol.

In the following section we provide some technical background regarding the techniques employed in our proposed solution.

III. Technical Background

In this section, we detail the main techniques used in our approach for preprocessing, feature extraction and classification, as described next.

A. *Image illumination compensation and equalization*

The illumination-reflectance model has been largely used by image enhancing algorithms for processing images acquired under poor lighting conditions. This model describes the image as been formed by two components: (i) the amount of illumination falling onto the scene and (ii) the amount of illumination reflected by the scene components. Therefore, under this model an image $f(x, y)$ is expressed as the product

$$f(x, y) = i(x, y) \cdot r(x, y) \tag{1}$$

where $i(x, y)$ and $r(x, y)$ are the illumination and reflectance components, respectively [43].

The illumination component of an image is generally characterized by smooth spatial variations, which are associated to the low-frequency spectral components. The image reflectance component tends to vary abruptly, especially on the connections between different objects, being therefore associated to the image spectral high frequency components. The goal of the illumination compensation procedure used in this work is to reduce the illumination component of a face image so that the final image approximates the face reflectance, which is independent of lighting conditions.

Homomorphic filtering was used for this purpose. Homomorphic filtering is a well known generalized technique for signal and image processing, which essentially utilizes a $H(u, v)$ filter and the convolution separability property of the Fourier Transform to map a non-linear combination problem into a linear combination one, followed by mapping back to the original domain [43].

Histogram equalization is a technique which seeks to enhance contrast by redistributing the gray values of an image pixels to obtain a histogram that approximates a uniform distribution, i.e., a histogram with ideally the same number (percentage) of pixels for any gray level [43]. In our approach the objective of using histogram equalization in the first preprocessing stage is to highlight the details of facial features present in the detected faces, in order to facilitate the task of the subsequent feature extraction method. We employed local equalization (i.e., only the cropped face image is processed) in order to prevent eventual over-equalization issues normally associated with the global approach.

B. Speeded Up Robust Features (SURF) Algorithm

The SURF algorithm has been proposed by Bay et al. [14], as a scale- and rotation-invariant interest point detector and descriptor which approximates or even outperforms previously proposed schemes, like the Scale Invariant Feature Transform SIFT [12], with respect to repeatability, distinctiveness, and robustness, yet it can be computed much faster. SURF finds keypoints using a so called Fast-Hessian Detector, based on an approximation of the Hessian matrix for a given image point. Assuming a bivariate continuous function $f(x, y)$, according with Laganière [44], the Hessian matrix is defined as the matrix of f partial derivatives, expressed as:

$$H(f(x,y)) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} \tag{2}$$

The determinant of the Hessian matrix, known as discriminant, is calculated as:

$$det(H) = \frac{\partial^2 f}{\partial x^2} \cdot \frac{\partial^2 f}{\partial y^2} - \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 \tag{3}$$

A negative discriminant indicates eigenvalues with different signs, therefore the analyzed point is neither a local maximum or a minimum. If the discriminant is positive, indicating that both eigenvalues are either positive or negative, then an extreme point (maximum or minimum) is detected. SURF utilizes a second order Gaussian filter to approximate the image partial derivatives, since this filter allows both scale and space analysis.

Hessian-based detectors such as SURF are stable, repeatable, and fire less on elongated, ill-localized feature structures. The responses to Haar wavelets are used for orientation assignment, before the keypoint descriptor is formed from the wavelet responses in a given surrounding keypoint neighborhood. Therefore, the Hessian matrix determinant provides a metric to select the location and the scale points. For the blurring step of the calculations, SURF utilizes an approximate second order Gaussian derivative using box filters which increases its performance.

In the following section, more details of the proposed method are given.

IV. Proposed Method

The method we propose includes a preprocessing stage which is composed of two steps: 1) detecting and correcting face orientation, and 2) illumination compensation and equalization. Following that, the main processing stage is also composed of two additional steps: 1) extracting features for SURF based facial representation; and 2) comparing facial representations for determining a similarity matrix. These stages are described next. The complete pipeline of our method can be visualized by the sequence of images in Figures 1, 2 and 3.

A. Preprocessing

As previously mentioned, preprocessing starts by detecting and correcting face orientation, which requires locating, cropping and normalizing human faces, both in size and orientation. Orientation normalization or correction consists of rotating the face image around its central point by the angle between the line segment joining the eyes mid point and the horizontal axis.

Given the position of left and right pupils, eye alignment is an affine transformation of the original detected face image. Suppose that we have acquired two pupils with coordinates (x_1, y_1), and (x_2, y_2), the expected distance between the two eyes on the aligned face is *d*, the rotation center is calculated as $(c_x, c_y) = (x_1 - x_2, y_1 - y_2)$ and the rotation angle $\theta = arctan((y_2 - y_1)/(x_2 - x_1))$. Mapping matrices *A* and *B* for rotation can be created as follows.

$$A = \begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix} \tag{4}$$

$$B = \begin{bmatrix} (1-\alpha) \cdot c_x - \beta \cdot c_y \\ (1-\alpha) \cdot c_y - \beta \cdot c_x \end{bmatrix}, \tag{5}$$

where $\alpha = d \cdot \cos(\theta)$ and $\beta = d \cdot \sin(\theta)$.

The aligned location (x', y') for the original point in (x, y) can be obtained by:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = A \cdot \begin{bmatrix} x \\ y \end{bmatrix} + B \tag{6}$$

This operation is illustrated in Figure 1.

After orientation is corrected, the input image is submitted to an enhancing step, consisting of: i) homomorphic filtering for improving image quality by dynamic range compression, and ii) local histogram equalization for contrast improvement (see Figure 2).

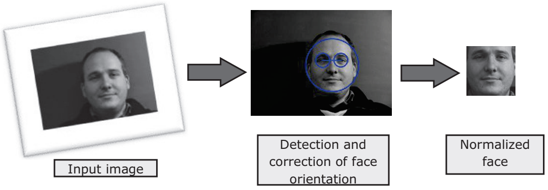


Fig. 1. Detection and correction of face orientation.

Once preprocessed and normalized, face images are submitted to the main processing module, illustrated in Figure 3. The description of the components of this module is presented in the following subsection.

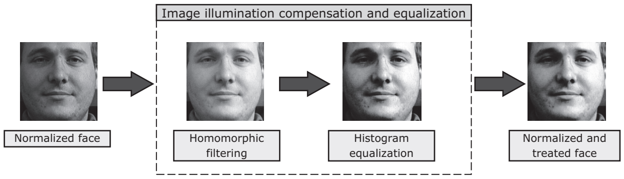


Image illumination compensation and equalization
Normalized face
Homomorphic filtering
Histogram equalization
Normalized and treated face

Fig. 2. Image illumination compensation and equalization.

B. Main Processing

The core of the main processing module consists of facial features extraction with SURF descriptors and facial representations comparison for building a similarity matrix.

When comparing image features, scale changes may be a problem if object images are captured from different distances, thus resulting in different size representations. Thus, if a fixed size neighborhood is used for feature extraction, pixels intensity values will not match [44]. To overcome this problem, scale invariant features have been utilized. Two techniques in particular have received considerable attention, Scale Invariant Feature Transform SIFT [12] and Speeded Up Robust Features SURF [14], described in the previous section. Both detect image keypoints and generate from those a scale invariant descriptor. The resulting descriptor is also robust to changes in rotation, illumination and pose [45], factors which normally difficult recognition.

A comparative study between SIFT and SURF by Juan and Gwon (2009) showed a better performance for the latter, with a recognition rate of 85,7% against 78,1%. This result led us to choose SURF as facial features extractor for the present work. Another aspect considered for this choice was the innovation involved, as no other work was found that utilizes SURF for facial verification.

Comparing all possible pairs of face images and determining similarity between them is the next processing step in our system. For that, we propose a procedure based on the Fast Approximate Nearest Neighbors FANN algorithm [46] and on the algorithm proposed by Antonopoulos et al. [47], which maps the results into a similarity matrix. The FANN algorithm [46] explores the K-Means hierarchical tree in a best-bin-first fashion (based on kd-trees), i.e., returns the nearest neighbor for a large fraction of queries and a very close neighbor in the remaining cases. Operation of this module is illustrated in Figure 3.

The FANN algorithm provides the amount of correspondences between the SURF keypoints for a pair of descriptors under comparison. Grey level histograms are calculated for those keypoints neighborhoods and weighted by their intersection, according to Swain and Ballard [48]:

$$d(H_p, H_q) = \frac{\sum_i min(H_p^i, H_q^i)}{\sum_i H_p^i} \tag{7}$$

In Equation 7, H_p and H_q are the correspondent keypoints neighborhood histograms. This equation assumes values from 0 (for totally distinct histograms) to 1 (when identical histograms are compared).

For each descriptors pair, correspondence has to be calculated twice, one for (A, B) and another for (B, A), as those values are not identical. The final correspondence for the pair will be the maximum value between (A, B) and (B, A). The resulting symmetrical similarity matrix, is defined by the similarity function in Equation 8 [47]. Note that not all pairwise comparisons of the four input SURF descriptors in Figure 3 are shown. Moreover, the matrix seen in that same figure is input to the transformation described in Equation 8.

$$S(A, B) = S(B, A) = 100\left(1 - \frac{M_{AB}}{min(K_A.K_B)}\right), \tag{8}$$

where $M_{AB} = max(d(H_A, H_B), d(H_B, H_A))$.

In Equation 8, M_{AB} is the maximum weighted histogram intersection value between either (A, B) or (B, A), and K_A and K_B are the amount of keypoints for those descriptors, respectively. This function takes values in the interval $[0, 100]$, the lower the value the more similar are the descriptors.

The face verification classifier C for comparing faces A and B is based on the output of Equation 8, and can be defined as follows:

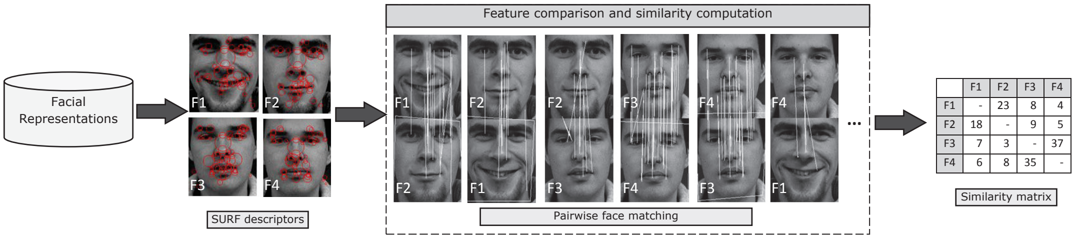
$$C(A, B) = \begin{cases} \text{Match} & \text{if } S(A, B) \geq T; \\ \text{NoMatch} & \text{if } S(A, B) < T. \end{cases} \tag{9}$$

where T is the threshold parameter, Match and NoMatch indicate whether there is a match ou not between faces A and B, respectively.

The parameter T is empirically determined in order to produce the desired performance, expressed by the acceptance and rejection rates, as will be shown in the next section.

V. Implementation Details

The proposed approach was implemented using the C++ language, from the Integrated Development Environment (IDE) Microsoft Visual Studio 2005, which allows easy code creation and visual project organization. Moreover, we used functions from the OpenCV [49] library, an open source computer vision library written in C and C++. OpenCV integrates well with the Visual Studio IDE compiler and is



	F1	F2	F3	F4
F1	-	23	8	4
F2	18	-	9	5
F3	7	3	-	37
F4	6	8	35	-

Fig. 3. Features comparison and similarity determination.

fully supported. The results were obtained using a standard 2 GHz dual core personal computer with 2 GB of RAM, running Windows XP operating system. OpenCV functions were used for image i/o operations, histogram equalization, SURF features extraction, as well as eyes and face detection (which is based on the work of Viola and Jones [50]). The homomorphic filtering and face comparing/matching procedure were implemented from scratch based on standard OpenCV filters. Considering a comparison of a pair of images from the LFW database, the computational time (T, in ms) and memory (M, in KB) average costs of the proposed method (over 30 runs) were the following: (i) detection step: T=489.6, M=14; (ii) illumination correction step: T=124.5, M=38; and (iii) main processing step: T=331.5, M=72.

VI. Experimental Evaluation

In this section we present a performance evaluation of our method for facial verification. For that we employed the LFW image database. No annotations are present in those images, although the faces have been aligned by the Huang et al. [51] algorithm.

Our method receives pairs of face images from the LFW database and determines whether the two images correspond or not to the same person, instead of looking for the most similar face to a given input face image. We previously mentioned the existence of three evaluation protocols for the LFW database, from which we choose the unsupervised setting protocol.

For our test experiments, the View 2 test set was utilized, composed of 6,000 image pairs, which was split in ten 600 pairs partitions. Half of those pairs belong to the same person and the other half belong to different people. All face images were reduced to a standard 186x94 pixels size, in order to maintain the face region only, as in the del Solar et al. [42] experiments. The final performance is expressed as the area under the ROC curve (AUC) and the standard error [52]. We also present a visual comparison of the ROC curves. The rates TPR (true positive), FPR (false positive) and accuracy (ACC) are, respectively, calculated [53] as:

$$TPR = \frac{TP}{P} \tag{10}$$

$$FPR = \frac{FP}{P} \tag{11}$$

$$ACC = \frac{TP+TN}{P+N} \tag{12}$$

where TP is number of face pairs classified as belonging to a sole person; P is the number of face pairs containing images of the same person; FP is number of misclassified face pairs with images of different persons; N is the number of face pairs with images of different persons; TP + TN is number of correctly classified face pairs; and P + N is the total number of face pairs.

The ROC curves were generated by varying classification threshold values (Equation 8). Each value of TPR and FPR correspond to the best accuracy point along the ROC curve for a particular classification threshold. For each of the 10 partitions, the best accuracies obtained by the proposed method are shown in Table I.

An extensive set of experiments was conducted by del Solar et al. [42] with three methods, H-XS-40, GJD-BC-100 and SD-MATCHES, using the unsupervised protocol of the LFW database to find the best combination of image descriptor and similarity measure. Those methods are provided by the LFW web page as examples of classifier performance that can be achieved. The H-XS-40 method utilizes LBP histograms as descriptors and the Qui-square distance to measure similarity for a 81x150 pixels face region. The GJD-BC-100 method considers a 100x150 pixels face region from which Gabor Jets descriptors are extracted and the Borda Count distance used as similarity measure. Finally, for the SD-MATCHES method the face region is also of 100x150 size, with SIFT descriptors and the number of matches between keypoints measuring similarity. According to the authors, the best performance among the three methods was achieved by the H-XS-40 method.

Seo and Milanfar [6] proposed the LARK descriptor using principal components analysis (PCA). Similarity is computed by the geodesical distance between the central and adjacent pixels in a given neighborhood. The method is reported to achieve 72,23% of accuracy for the LFW base and the unsupervised protocol. The results for the four methods presented above and for our method, are summarized in Figure 4.

It can be observed that our method (Proposed) shows the best performance. This superior performance is confirmed in Table II, which shows the AUC and the standard error values for the methods as provided by the cited papers [42], [6].

	Partition									
	1	2	3	4	5	6	7	8	9	10
ACC	0.7383	0.7367	0.7267	0.7350	0.7317	0.7417	0.7283	0.7233	0.7367	0.7367

TABLE I
BEST ACCURACIES OF THE PROPOSED METHOD FOR THE LFW VIEW 2 TEST SET.

	Method				
	Proposed	SD-MATCHES	GJD-BC-100	H-XS-40	LARK
AUC	0.82194 ± 0.00383	0.67562 ± 0.00486	0.73917 ± 0.00450	0.75468 ± 0.00439	0.78304 ± 0.00418

TABLE II
AREA UNDER ROC AND STANDARD ERROR FOR THE COMPARED METHODS.

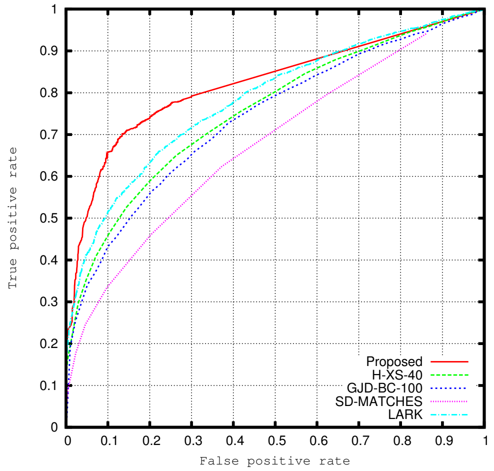


Fig. 4. ROC curves for LFW unsupervised evaluation. The method proposed in this paper has the best performance among existing methods.

VII. CONCLUSIONS AND CONTRIBUTIONS

This paper presents a novel approach to the face verification problem with state-of-the art performance. The proposed method is characterized by an initial preprocessing stage, the use of SURF features extracted from keypoints on the face image and a new features comparison strategy, which is based on an existing similarity function [47]. Similarity is calculated between gray levels histograms of the keypoints and weighted by the histograms intersections [48]. Tested with the LFW images database under the unsupervised protocol, our method performed better than other state of the art methods reported at the LFW website [42], [6].

Our contributions to the face verification area are three-fold. First, one innovative aspect of our method is its robustness, achieved by aggregating modules to prevent degradation usually caused by illumination changes as well as facial expression and pose variations. Second, we propose a novel similarity function for face verification. Lastly, we show that the proposed method achieves the best performance among published work on the LFW unsupervised setting protocol.

Comparison with additional methods is not straightforward since it either requires an implementation of the methods or the existence of a publication with experiments in a public database using the unsupervised protocol. The methods in the experimental evaluation satisfied the second requirement. Until now, we could not find related work with available code or sufficient details for a faithful implementation. Future work will consider extending the proposed method for digital video applications, like security summarization [54], people identification [55], people based video archiving [56] and extraction of cast lists [57].

ACKNOWLEDGMENT

The authors would like to thank CAPES for the financial support of the PhD work of the first author.

REFERENCES

[1] G. Huang, M. Ramesh, T. Berg, and R. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, 2007.

[2] J. Choi, S. Yang, Y. Ro, and K. Plataniotis, "Face annotation for personal photos using context-assisted face recognitionace recognition," in *Proceedings of the 1st ACM International Conference on Multimedia information retrieval*, vol. 1, 2008, pp. 44–51.

[3] A. Kapoor, G. Hua, A. Akbarzadeh, and S. Baker, "Which faces to tag: Adding prior constraints into active learning," in *Proceedings of the 13th International Conference on Computer Vision*, 2009, pp. 1058–1065.

[4] D. Lin, A. Kapoor, G. Hua, and S. Baker, "Joint people, event, and location recognition in personal photo collections using cross-domain context," in *Proceedings of the 11th European Conference on Computer Vision*, 2010, pp. 243–256.

[5] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *Proceedings of the 23th International Conference on Computer Vision and Pattern Recognition*, vol. 10, 2010, pp. 2707–2714.

[6] H. J. Seo and P. Milanfar, "Training-free, generic object detection using locally adaptive regression kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1688–1704, 2010.

[7] L. Wolf, T. Hassner, and Y. Taigman, "Descriptor based methods in the wild," in *Faces in Real-Life Images Workshop in European Conference on Computer Vision*, 2008, pp. 1–14.

[8] N. Pinto, J. Dicarlo, and D. Cox, "How far can you get with a modern face recognition test set using only simple features?" in *Proceedings of the 22th International Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2591–2598.

[9] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2009, pp. 498–505.

[10] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayer, "Attribute and simile and classifiers for face verification," in *IEEE International Conference on Computer Vision*, 2009, pp. 365–372.

[11] T. Ojala, M. Pietikäinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions,” *Pattern Recognition*, vol. 29, pp. 51–59, 1996.

[12] G. D. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the International Conference on Computer Vision*, 1999, pp. 1150–1157.

[13] M. Dalal and B. Triggs, “Histogram of oriented gradietns for human detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.

[14] H. Bay, T. Tuytelaars, and L. V. Gool, “Surf: Speeded up robust features,” in *European Conference on Computer Vision*, 2006, pp. 404–417.

[15] Y. Su, S. Shan, X. Chen, and W. Gao, “Hierarchical ensemble of global and local classifiers for face recognition,” *IEEE Transactions on Image Processing*, vol. 18, no. 8, pp. 1885–1896, 2009.

[16] T. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, “Active shape models - their training and application,” *Computer Vision and Image Understanding*, vol. 61, pp. 38–59, 1995.

[17] S.-C. Zhang, B. Fang, Y.-Z. Liang, J. Wen, and L. Wu, “A face clustering method based on facial shape information,” in *International Conference on Wavelet Analysis and Pattern Recognition*, 2011, pp. 44–49.

[18] C. Zhu, F. Wen, and J. Sun, “A rank-order distance based clustering algorithm for face tagging,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 481–488.

[19] J. Wright and G. Hua, “Implicit elastic matching with randomized projections for pose-variant face recognition,” in *Proceedings of the 22th International Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1502–1509.

[20] A. Gallagher and T. Chen, “Clothing cosegmentation for recognizing people,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[21] K. Jayech and M. Mahjoub, “Clustering and bayesian network for image of faces classification,” *International Journal of Advanced Computer Science and Applications, Special Issue on Image Processing and Analysis*, vol. Special Issue: Image Processing, pp. 35–44, 2011.

[22] P. Simard, Y. L. Cun, J. Denker, and B. Victorri, “Transformation invariance in pattern recognition - tangent distance and tangent propagation,” *Lecture Notes in Computer Science - Neural Networks: Tricks of the Trade*, vol. 1524, pp. 239–274, 1998.

[23] W. Zhang, T. Zhang, and D. Tretter, “Beyond face: Improving person clustering in consumer photos by exploring contextual information,” in *IEEE International Conference on Multimedia and Expo*, 2010, pp. 1540–1545.

[24] J. Winn, A. Criminisi, and T. Minka, “Object categorization by learned universal visual dictionary,” in *Proceedings of the 10th International Conference on Computer Vision*, vol. 2, 2005, pp. 1800–1807.

[25] Y. Rubner, C. Tomasi, and L. Guibas, “The earth mover’s distance as a metric for image retrieval,” *International Journal of Computer Vision*, vol. 40, pp. 99–121, 2000.

[26] Picasa, “Google - Picasa,” 2010, last accessed: 12/July/2013. [Online]. Available: http://picasa.google.com/features.html

[27] J. Cui, F. Wen, R. Xiao, O. Tian, and X. Tang, “Easyalbum: an interactive photo annotation system based on face clustering and re-ranking,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2007, pp. 1222–1228.

[28] S. J. D. Prince and J. H. Elder, “Bayesian identity clustering,” in *Proceedings of the Canadian Conference on Computer and Robot Vision*, 2010, pp. 32–39.

[29] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” *IEEE Transactions on Patter*, vol. 19, pp. 711–720, 1997.

[30] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer-Verlag, 2002.

[31] K. Messer, J. Matas, J. Kittler, J. Lüttin, and G. Maitre, “Xm2vtsdb: The extended m2vts database,” in *Proceedings of the 2nd International Conference on Audio and Videobased Biometric Person Authentication*, 1999, pp. 72–77.

[32] Y. Freund, S. Dasgupta, M. Kabra, and N. Nerma, “Learning the structure of manifolds using random projections,” in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2007, pp. 1–8.

[33] G. Hua and A. Akbarzadeh, “A robust elastic and partial matching metric for face recognition,” in *Proceedings of the IEEE 12th International Conference on Computer Vision*, 2009, pp. 2082–2089.

[34] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, “Face recognition using laplacian faces,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 328–340, 2005.

[35] D. Cai, X. He, and J. Han, “Spectral regression for eficient regularized subspace learning,” in *Proceedings of the 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[36] D. Cai, X. He, Y. Hu, T. Huang, and J. Han, “Learning a spatially smooth subspace for face recognition,” in *Proceedings of the 20th International Conference on Computer Vision and Pattern Recognition*, vol. 7, 2007, pp. 1–7.

[37] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[38] A. Martínez and R. Benavente, “The ar face database,” Computer Vision Center, Tech. Rep. 24, Jun 1998.

[39] A. Georghiades, P. Belhumeur, and D. Kriegman, “From few to many: Illumination cone models for face recognition under variable lighting and pose,” *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[40] J. Zou, Q. Ji, and G. Nagy, “A comparative study of local matching approach for face recognition,” *IEEE Transactions on Image Processing*, vol. 16, no. 10, pp. 2617–2628, 2007.

[41] F. Moosmann, E. Noak, and F. Jurie, “Randomized clustering forests for image classification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 9, pp. 1632–1646, 2008.

[42] J. R. del Solar, R. Verscheae, and M. Correa, “Recognition of faces in unconstrained enviornments: A comparative study,” *EURASIP Journal on Advances in Signal Processing (Recent Advances in Biometric Systems: A Signal Processing Perspective)*, vol. 2009, pp. 1–19, 2009, article ID 184617.

[43] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Addison-Wesley Pub, 2010.

[44] R. Laganiére, *OpenCV 2 Computer Vision Application Programming Cookbook*. Packt Publishing, UK, 2011.

[45] J. Bauer, N. Sünderhauf, and P. Protzel, “Comparing several implementations of two recently published feature detectors,” in *International Conference on Intelligent and Autonomous Systems*, vol. 22, 2007, pp. 481–494.

[46] M. Muja and D. G. Lowe, “Fast approximate nearest neighbors with automatic algorithm configuration,” in *International Conference on Computer Vision Theory and Applications*, 2009, pp. 331–340.

[47] P. Antonopoulos, N. Nikolaidis, and I. Pitas, “Hierarchical face clustering using sift image features,” in *IEEE Symposium on Computational Intelligence in Image and SignalProcessing*, 2007, pp. 325–329.

[48] M. J. Swain and D. H. Ballard, “Color indexing,” *International Journal of Computer Vision*, vol. 7, pp. 11–32, 1991.

[49] OpenCV. (2013) Open source computer vision library. [Online]. Available: http://sourceforge.net/projects/opencvlibrary/

[50] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, pp. 511–518.

[51] G. Huang, V. Jain, and E. Learned-Miller, “Unsupervised joint alignment of complex images,” in *Proceedings of the 11th IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[52] J. A. Hanley and B. J. Mcneil, “The meaning and use of the area under a receiver operating characteristic (roc) curve,” *Radiology*, vol. 143, pp. 29–36, 1982.

[53] T. Fawcett, “Roc graphs: Notes and practical considerations for researchers,” HP Laboratories, Tech. Rep. HPL-2003-4, 2003.

[54] A. Sony, K. Ajuth, K. Thomas, T. Thomas, and P. L. Oeepa, “Video summarization by clustering using euclidean distance,” in *Proceedings of the 2011 International Conference on Signal Processing, Communication, Computing and Networking Technologies*, 2011, pp. 642–646.

[55] H. Gao, H. K. Ekenel, and R. Stiefelhagen, “Identifying important people in broadcast news videos,” in *Proceedings of the Conference on Machine Vision Applications*, 2011, pp. 127–136.

[56] K. Yamamoto, O. Yamacuchi, and H. Aoki, “Fast face clustering based on shot similarity for browsing video,” in *Proceedings of the Progress in Informatics, Special issue: 3D image and video technology*, vol. 7, 2010, pp. 53–62.

[57] Y.-F. Zhang, C. Xu, H. Lu, and Y.-M. Huang, “Character identification in feature-length films using global face-name matching,” *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1276–1288, 2009.