

Biometric Identification from Facial Sketches of Poor Reliability: Comparison of Human and Machine Performance

H. Proença, J. C. Neves, J. Sequeiros, N. Carapito and N.C. Garcia

Abstract Facial sketch recognition refers to the establishment of a link between a drew representation of a human face and an identity, based on information given by a eyewitness of some illegal act. It is a topic of growing interest, and various software frameworks to synthesize sketches are available nowadays. When compared to the traditional hand-made sketches, such sketches resemble more closely the appearance of real mugshots, and led to the possibility of using automated face recognition methods in the identification task. However, there are often deficiencies of witnesses in describing the subjects' appearance, which might bias the main features of sketches with respect to the corresponding identity. This chapter compares the human and machine performance in the task of sketch identification (rank-1 identification). One hundred subjects were considered as gallery data, and five images from each stored in a database. Also, one hundred sketches were drew by non-professionals and used as probe data, each of these resembling an identity in the gallery set. Next, a set of volunteers was asked to identify each sketch, and their answers compared to the rank-1 identification responses given by automated face recognition techniques. Three appearance-based face recognition algorithms were used: 1) Gabor-based description, with ℓ_2 norm distance ; 2) sparse representation for classification; and 3) eigenfaces. The sparse representation for classification algorithm yielded the best results, whereas the responses given by the Gabor-based description algorithm were the most correlated to human responses.

1 Introduction

The pursuit of criminals based on sketches drew by eyewitnesses of some illegal act has been used by law enforcement agencies for many years. This process has

University of Beira Interior, Department of Computer Science, IT- Instituto de Telecomunicações, 6201-001 Covilhã, Portugal. e-mail: {hugomcp, jcneves, a29192, a29195, a26931}@ubi.pt

two major phases: 1) the eyewitness is asked to draw or coordinate the drawing of a representation of a face; 2) once the eyewitness considers the sketch close enough to the desired appearance, a range search over the enrolled identities is carried out, in order to find the sub-set of identities that maximally resemble the appearance of the sketch.

Sketches were traditionally hand-made, either by the witness or law enforcement agents, which conditioned their appearance and pushed them away from real mugshots. Due to this, the earliest attempts to perform facial sketch recognition translated the gallery data into a simplified *sketch domain*, performing recognition subsequently. However, unlike hand-drew sketches, *composite sketches* are synthesized using facial composite software systems and more closely resemble real faces, which raises the possibility of using automated face recognition algorithms in the identification task.

Most of the judicial processes of modern societies rely on the honesty of eyewitnesses [7]. Their testimony is known to impress deeply elements of a jury and, as such, their reliability strongly matters. In this scope, perjury is usually considered a crime, but it is not simple to distinguish it from mere misremembering, which is always a strong possibility due to the emotional condition of witnesses both at the moment of the crime and of the trial. Several studies about the reliability of eyewitnesses corroborate the vulnerability of human memory to bias [25], which increases the skepticism about the *reliability* of the sketches. i.e., how closely the appearance of a sketch actually resembles the appearance of the human identity that it should? This topic motivated several studies over the last years and concerns a significant amount of researchers (e.g., [22], [11], [16] and [8]).

In this chapter we address the performance of automated face recognition methods on composite sketches and compare it to the attained by humans, considering that humans are extremely efficient in identifying a previously seen face [2]. In the scope of the work reported in this chapter, Hancock, Bruce and Burton [10] compared the performance of two sketch recognition algorithms to the ability of humans, having concluded that automated methods provided reasonable correlations with human ratings and memory data. Also, they studied the ability of each algorithm to perform recognition with and without the hair visible, and with and without alteration in facial expressions.

According to the above, the main singularities of this work are two-fold: 1) we provide the performance of algorithms devised for recognizing real mugshots on composite-sketch data; 2) we analyze the results with respect to the level of *reliability* of the sketches, measured by the agreement of human responses on each sketch, i.e., if most people agree in the identity established for a given sketch, then we consider it a reliable sketch.

The main motivations for the study reported in this chapter can be summarized as follows:

1. there is an increasingly high realism of sketches drew by software systems. These frameworks produce such high realistic sketches that are often hard to distinguish from real faces;

2. the fact that eyewitnesses often don't remember with detail some of the subjects features, which results in sketches that might not reproduce closely the identity of interest. In the scope of this work, this concept is referred as *reliability* of the sketch.
3. the relative effectiveness attained by humans and automated recognition algorithms remains to be perceived. This enables to assess the degree of certainty of humans in the identification task, correlating it to the performance of automated recognition methods.

The remainder of this chapter is organized as follows: Section 2 gives an overview of the most relevant methods to perform facial sketch recognition. The description of our experiments and the corresponding discussion is given in Section 3. Finally, Section 4 concludes the chapter.

2 Facial Sketch Recognition

Figure 1 gives a cohesive perspective of facial sketch recognition algorithms. Having a set of gallery images of known identity (usually designated as *mugshots*), the problem is to assign the probable identity to a sketch drew by a witness of some crime. There are two general families of algorithms for this task: 1) one that converts the available gallery data into the sketches domain, which is believed to most probably resemble the appearance of the query sketch; and 2) algorithms that match the query sketch directly to the available gallery data, provided that the feature description and matching algorithms are robust enough to handle such obvious differences in appearance of real data and sketches.

Most representative works on this subject tackled the problem in the *sketch domain*, i.e., by transforming gallery mugshots into sketches (e.g., using eigenspace techniques) and performing recognition subsequently (using methods such as Bayesian classifiers, nearest neighbor techniques and analysis of principal components). Among these works, the proposal of Wang and Tang [27] consists of a photo-sketch synthesis and a recognition method based on Markov Random Fields. Their proposal has three major phases: 1) given a mugshot, synthesize a sketch; 2) given a sketch, synthesizing a photo; and 3) searching for face photos in the database based on a query sketch drew by an artist. Authors used a training set containing photo-sketch pairs of frontal poses, with normal lighting, neutral expression and no significant occlusions. In the recognition phase, two variants were tested: 1) firstly transforming the photos into sketches and match on that domain; and 2) transforming sketches to photos and match in the image domain. Li, Savvides and Bhagavatula [17] proposed a system for face sketch synthesis and recognition based in two main phases: pseudo-sketch generation and sketch recognition. The pseudo-sketch generation method is based on local linear preserving of geometry between photo and sketch images, and the recognition phase relies in nonlinear discriminate analysis, in order to match the identity of interest from the previously synthesized

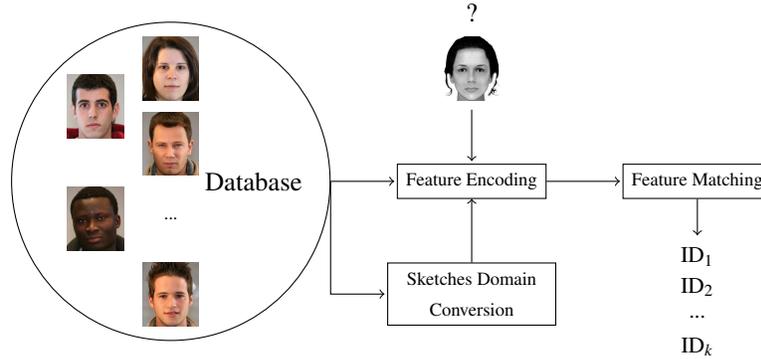


Fig. 1 Block diagram of a typical facial sketch recognition algorithm. Given a set of mugshots of known identities enrolled in a database, the problem is to establish the identity of a sketch. While some of the existing methods translate the gallery data into the sketch domain, the problem has two major phases: feature encoding and matching.

pseudo-sketches. Klare, Li and Jain [15] were particularly concerned with the forensic domain, based on the fact that most times sketches are drawn while not looking for the subject, which decreases the quality of the result. Hence, they presented a recognition method based in local feature-based discriminant analysis (LFDA). In LFDA, both sketches and photos are represented by Scale Invariant Feature Descriptors (SIFT) and Local Binary Patterns (LPB), from where multiple discriminant projections are taken for feature representation. Nearest neighbor is used in classification, resulting in a system that offers substantial improvements in matching forensic sketches to the corresponding face images, when compared to state-of-the-art face recognition methods. An overview of this approach is also given in [12]. The approach of Liu, Tang, Jin, Lu and Ma [18] has a main distinguishing feature: instead of transforming mugshots into the sketch domain, they inverted the problem and started by generating a realistic face image from the composite sketch using a Hybrid subspace method and then built an illumination tolerant correlation filter able to recognize the person under different illumination variations from a surveillance video footage. Najati and Sim [19] approached the problem from the perspective of the quality of the sketch. Considering that these are typically unreliable, they asked the eyewitness to draw a sketch of the target face, providing some ancillary information. Then, a drawing profile from the witness was obtained, by asking him/her to draw a set of face photos. This profile is used to correct the sketch for the witness bias, and feeds the recognition module, having authors observed consistent improvements in performance, when compared to starting from the raw sketches. The approach of Pramanik and Bhattacharjee [21] relies in statistics drew from facial landmarks, such as the eye corners, nose, eyebrows and lips. They extracted a set of lengths, areas and ratios, from both gallery mugshots and sketches. Next, the feature vectors were normalized by their energy removed. Nearest neighbor techniques were used to find the most probably identity for a sketch. Authors concluded

that even this simple approach is effective when faces are in frontal pose, under homogenous lighting conditions and without significant occlusions.

To the best of our knowledge, only Han, Klare, Bonnen and Jain [9] considered the specific features of sketches synthesized using facial composite software and proposed a component-based representation model to measure the similarity between the sketch and a mugshot. Facial landmarks in composite sketches and mugshots were detected by means of an active shape model. Features were extracted for each facial component using multi-scale local binary patterns, and similarity per component was calculated. Finally, the similarity scores obtained from individual components were fused, yielding the final similarity score.

3 Experiments and Discussion

In our experiments the responses of human observers and of automated facial sketch recognition algorithms were compared. Regarding human, we asked to a set of volunteers to look for a sketch and match it to one of the enrolled identities (available as mugshots). The used protocol is summarized below:

1. Ask to each human volunteer to observe each sketch on a computer screen.
2. After fifteen seconds, remove the sketch from the desktop. In case of any doubt, humans were allowed to observe the sketch again.
3. Based on a single mugshot per enrolled identity, ask to the human observer to match the sketch to one identity. The human observer may go over the database any number of times and gradually eliminate the less likely foils until they are left with just one.
4. Run three automated face recognition algorithms, using the set of sketches as probes and five gallery mugshots per subject for learning purposes.
5. Evaluate the levels of correlation between the responses given by humans and rank-1 identification results of the automated recognition algorithms.

3.1 Recognition using Sparse Representations

Sparse representations have been largely reported in the literature (e.g. [20]). The main idea is to obtain a sparse linear representation of a probe with respect to a set of samples that constitute an overcomplete dictionary of base elements. In this chapter, the results yielded by the approach of Wright, Yang, Ganesh, Sastry and Ma [28] are reported. The singularity of this proposal lies in the fact that, instead of using sparsity to identify a relevant model, it uses the sparse representation of each probe directly for classification, obtaining a linear combination of the base elements that give a compact representation of the probe, assuming that a sufficient number of training samples per class i exists, i.e., $\mathbf{A}_i = [v_{i,1}, \dots, v_{i,n}] \in \mathbb{R}^{m \times n_i}$. Formally,

each probe is regarded as a column vector \mathbf{y} that is possible to represent by a linear combination of samples of the corresponding i^{th} class:

$$\mathbf{y} = \alpha_{i,1}v_{i,1} + \dots + \alpha_{i,n}v_{i,n}, \quad (1)$$

being $v_{i,\cdot}$ the column representation of the gallery data. For classification, a matrix \mathbf{A} yields from the concatenation of all \mathbf{A}_i elements:

$$\mathbf{A} \doteq [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_k] = [v_{1,1}, v_{1,2}, \dots, v_{k,n_k}]. \quad (2)$$

The linear representation of a probe \mathbf{y} belonging to the i^{th} class can be written in terms of \mathbf{A} :

$$\mathbf{y} = \mathbf{A}\mathbf{x}_i \in \mathbb{R}^m, \quad (3)$$

where $\mathbf{x}_i = [0, \dots, 0, \alpha_{i,1}, \dots, \alpha_{i,n_i}, 0, \dots, 0]^T \in \mathbb{R}^n$ is considered a *sparse* vector, as all its entries are zero, except those associated to the corresponding i^{th} class.

Considering that the system $\mathbf{y} = \mathbf{A}\mathbf{x}$ is typically underdetermined, authors seek the sparsest solution to it, by choosing the minimum ℓ_0 -norm solution:

$$(l^0) : \hat{\mathbf{x}}_0 = \arg \min \|\mathbf{x}\|_0, \text{ subject to } \mathbf{A}\mathbf{x} = \mathbf{y}. \quad (4)$$

Based on recent developments in the theory of sparse representation, authors noted that if the solution \mathbf{x}_0 is sparse *enough*, the solution to the ℓ_0 minimization problem is equal to the corresponding (ℓ_1) minimization problem:

$$(l^1) : \hat{\mathbf{x}}_1 = \arg \min \|\mathbf{x}\|_1, \text{ subject to } \mathbf{A}\mathbf{x} = \mathbf{y}. \quad (5)$$

This problem can be solved in polynomial time by linear programming methods [1]. Considering the noisy environment of biometric recognition, the model in (5) might not hold exactly, and a modification to account for small dense noise was proposed, $\mathbf{y} = \mathbf{A}\mathbf{x}_i + \mathbf{z}$, being $\mathbf{z} \in \mathbb{R}^m$ the noise term with bounded energy $\|\mathbf{z}\|_2 < \varepsilon$. For this model, the sparsest solution can be found by following l^1 minimization problem[5], possible to obtain via second order cone programming techniques:

$$(l_s^1) : \hat{\mathbf{x}}_1 = \arg \min \|\mathbf{x}\|_1, \text{ subject to } \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \varepsilon. \quad (6)$$

Having a probe \mathbf{y} of unknown class, the label is found based on how well the coefficients associated with all training samples of each class reproduce that probe. A characteristic function δ_i for each class is defined such that $\delta_i(\mathbf{x})$ is a vector whose only non-zero entries are associated with class i . Hence, $\hat{\mathbf{y}}_i = \mathbf{A}\delta_i(\hat{\mathbf{x}}_1)$ is an approximation of \mathbf{y} , based on a linear combination of gallery data of the i^{th} class. The representation $\hat{\mathbf{y}}_i$ that more closely fits \mathbf{y} is deemed to be the true class of \mathbf{y} :

$$\min_i \|\mathbf{y} - \mathbf{A}\delta_i(\hat{\mathbf{x}}_1)\|_2. \quad (7)$$

3.2 Recognition Using Gabor-based Description and Nearest Neighbor Matching

Daugman observed that simple cells in the visual cortex of mammalian brains can be modeled by Gabor functions [3][4]. and that image analysis based on Gabor functions is deemed to be similar to the perception in the human visual system. Since then, Gabor-based decompositions have been largely reported in the computer vision literature (e.g., [14] and [6]), and particularly in the biometrics domain (e.g., face [26] and fingerprint [23] recognition algorithms).

The impulse response of a Gabor kernel is defined by a harmonic function multiplied by a Gaussian function and these filters have a real and an imaginary component representing orthogonal directions:

$$g(x, y, \omega, \theta, \sigma_x, \sigma_y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{\Phi_1^2}{\sigma_x^2} + \frac{\Phi_2^2}{\sigma_y^2}\right)} e^{i\frac{2\pi\Phi_1}{\omega}}, \quad (8)$$

being $\Phi_1 = x\cos(\theta) - y\sin(\theta)$, $\Phi_2 = -x\cos(\theta) + y\sin(\theta)$, ω the wavelength and θ the orientation. Figure 2 illustrates examples of Gabor filters, where the scale (upper row) and rotation (bottom row) parameters are varying.

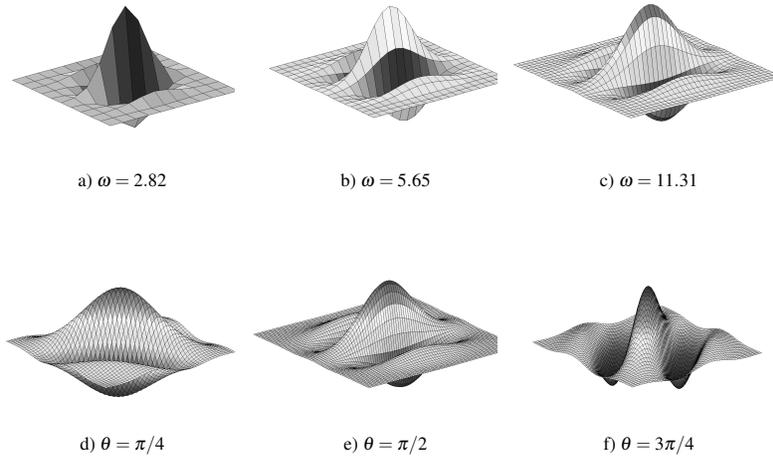


Fig. 2 Illustration of the Gabor filters (real parts) used to decompose the gallery images. The upper row illustrates variations in the *scale* parameter (ω , wavelength), whereas the bottom row illustrates variations in rotation (θ).

A bank of Gabor filters with various scales and rotations was created. The filters were convolved with the image I , resulting in a so-called Gabor space. In our

experiments $\sigma_x = \sigma_y = \omega/2$ was used, to alleviate the computational burden of the optimization process associated to finding the optimal parameters of Gabor kernels

$$\mathbf{g}_i(\mathbf{I}) = \mathbf{g} * \mathbf{I}, \quad (9)$$

being "*" the convolution operator. Let $\mathbf{G}(\mathbf{I}) = [|\mathbf{g}_1^*(\mathbf{I})|, \dots, |\mathbf{g}_n^*(\mathbf{I})|]$ be the concatenation of the magnitude of the $\mathbf{g}_i(\mathbf{I})$ elements, such that

$$\mathbf{g}_i^*(\mathbf{I}) = \mathbf{g}_i(\mathbf{I}) - \mu_{\mathbf{g}_i(\mathbf{I})}, \quad (10)$$

being $\mu_{\mathbf{g}_i(\mathbf{I})}$ the mean of the coefficients of $\mathbf{g}_i(\mathbf{I})$. Hence, $\mathbf{G}(\mathbf{I})$ has zero mean and feeds a matching module that yields the distance between the probe \mathbf{I} and its Gabor representations:

$$\min_k \sum_i \sum_c \|g_i^*(\mathbf{I}) - g_i^*(\mathbf{I}_c^{(k)})\|_1, \quad (11)$$

being $\mathbf{I}_c^{(k)}$ the c^{th} gallery image of the k^{th} class.

Figure 3 illustrates a gallery image \mathbf{I} (leftmost image) and three of its Gabor representations \mathbf{g}_i (images b), c) and d)) obtained when using the parameters $(\omega, \theta) = \{(0.33, \pi/4), (0.28, 3\pi/4), (0.51, \pi/2)\}$ for Gabor kernels (8).

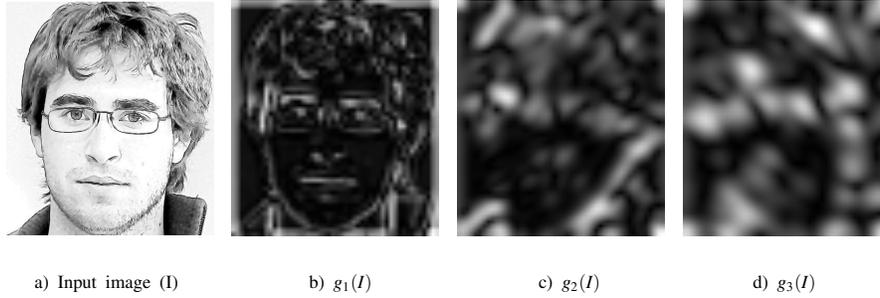


Fig. 3 Example of the decomposition of a gallery image \mathbf{I} (leftmost image) by three different parameterizations of Gabor kernels (images at right).

3.3 Recognition Using Eigenfaces

Firstly proposed by Turk and Pentland [24], the concept of *eigenface* became extremely popular in the computer vision literature and it is reported for many different purposes. The main idea is to simplify the face recognition problem by first mapping the data into a space of lower dimensionality, found by the principal components analysis. Let the j^{th} image from the i^{th} subject be represented by a column

vector $\mathbf{v}_{i,j}$. The first step is to obtain a representation of this column vector subtracted by the mean face of all images in the data base:

$$\mathbf{v}_{i,j}^* = \mathbf{v}_{i,j} - \frac{1}{n_s n_i} \sum_k \sum_r \mathbf{v}_{k,r}, \quad (12)$$

being n_s and n_i the number of subjects and of images per subject available. A dictionary is built from the concatenation of all $\mathbf{v}_{i,j}^*$, i.e., $\mathbf{A} = [\mathbf{v}_{1,1}^*, \dots, \mathbf{v}_{n_i, n_s}^*]$. Next, the covariance matrix is found $\mathbf{C} = \mathbf{A}\mathbf{A}^T$. The eigenvectors $\boldsymbol{\mu}_i$ of \mathbf{C} are found and the best d eigenvectors are kept, corresponding to the d largest eigenvalues:

$$\boldsymbol{\mu}_i = \mathbf{A} \mathbf{v}_i. \quad (13)$$

Each face in the training set can be represented as a linear combination of the best d eigenvectors:

$$\hat{\mathbf{v}}_{i,j}^* = \sum_d \boldsymbol{\omega}_d \boldsymbol{\mu}_d, \quad (\boldsymbol{\omega}_d = \boldsymbol{\mu}_d^T \mathbf{v}_{i,j}^*) \quad (14)$$

yielding to a representation of each face in the principal components space given by: $\boldsymbol{\omega}_{i,j} = [\omega_{i,j_1}, \dots, \omega_{i,j_d}]$. Given an unknown probe image y , it is normalized and projected on the eigenspace $\boldsymbol{\omega}_y = [\omega_{y_1}, \dots, \omega_{y_d}]$. The identity assumed for that probe is given by:

$$\hat{i} = \arg \min_i (\min_j \|\boldsymbol{\omega}_y - \boldsymbol{\omega}_{i,j}\|_2) \quad (15)$$

Figure 4 gives the eigenfaces associated to the four largest eigenvalues of \mathbf{C} , when analyzing the principal components of the gallery data used in our experiments (five images per subject, for one hundred subjects). here, we can perceive that the hair is the most discriminating region (corresponding to $\boldsymbol{\mu}_1$), followed by the low-frequency components of the facial region ($\boldsymbol{\mu}_2$. details that regard the ocular region and eyebrows are highlighted by $\boldsymbol{\mu}_2$ and high frequency information about the chin and nose are highlighted by $\boldsymbol{\mu}_4$).

3.4 Gallery Data

A set of one hundred subjects was used as enrolled identities. For each subject, five frontal images of their faces (mugshots) were acquired under natural and artificial light sources of varying intensity. Images have dimensions $4,368 \times 2,912$, and the upper-right and bottom left corners of the faces were manually annotated, defining regions-of-interest to avoid the effect of data misalignments. Next, each image was converted into grayscale and stored in 8-bit depth uncompressed format. Images at the left columns of figure 5 exemplify six gallery images from three different subjects.



Fig. 4 Examples of the eigenfaces associated to the four largest eigenvalues of the gallery data used in our experiments.

3.5 Sketches

Sketches were created by four humans (non-professional, i.e., they were not familiar with the task), using the *PortraitPad*¹ software toolkit. This is a free access web-based application that enables to customize the appearance of a sketch with high detail, according to the *composite method*. Here, users are requested to iteratively select the type of facial component (e.g., eyes, eyelids, lips,...) and a group of possibilities that exhibit varying textures, shapes and sizes. Next, by selecting one of these possibilities and placing it in the face that is being constructed, it is possible to resemble the appearance of a real human face with high reliability. Figure 6 gives a screenshot of the application used to construct the set of sketches of our experiments.

A total of one hundred images was created, having as main concern that each sketch should resemble one of the identities enrolled in the database. Figure 7 illustrates pairs of sketches and of the corresponding identities in the gallery data that they aim to resemble.

3.6 Correlation Between the Responses Given by Humans and Automated Methods

Initially, all the sketches were considered as probe data and presented to the human volunteers. They were asked to match each sketch against one of the identities enrolled in the gallery. No more than one possibility should be considered and - in case of doubt - volunteers had to choose the identity that they perceive to resemble the sketch *more closely*. Let $\varphi_{i,j}(x)$ be a characteristic function that gives the probability that a human observer matches the i^{th} probe to the j^{th} identity in the gallery set. Fig-

¹ <http://portraitpad.com/>

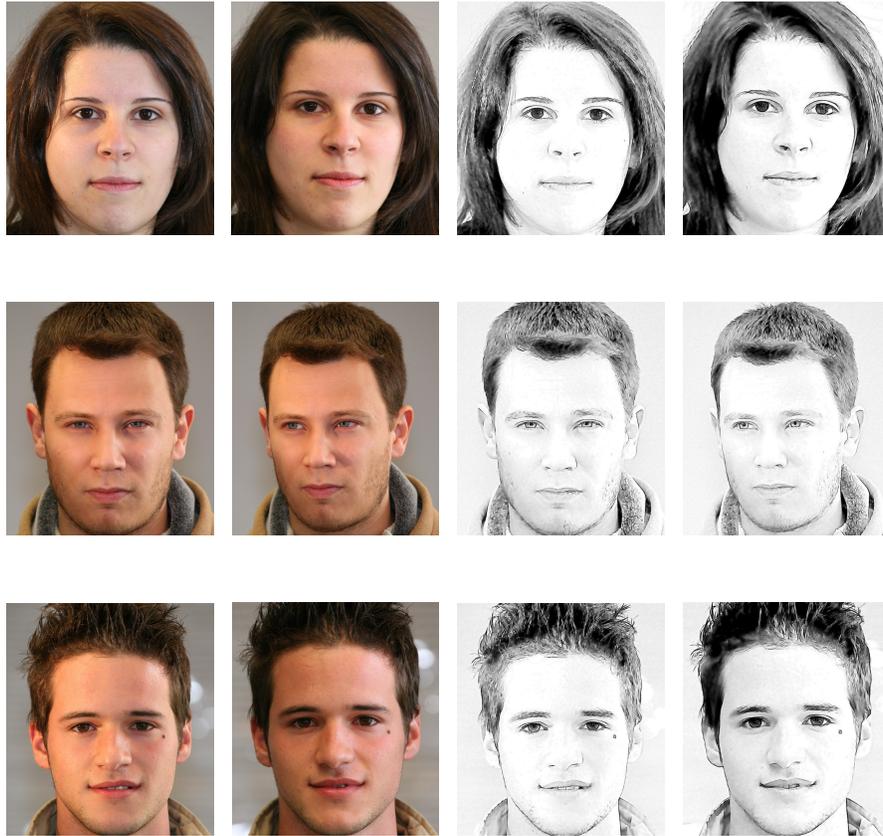


Fig. 5 Examples the images used as gallery data in our experiments. The two leftmost columns illustrate the used gallery samples, whereas the rightmost columns give the corresponding data translated into the sketch domain.

ure 8 gives extreme cases of the statistic of $\varphi_{i,j}$, where the horizontal axis denotes the class. Nearby each plot, the corresponding sketch is given, enabling to perceive the main features of the sketches where users had reduced uncertainty (upper row) and of the cases that raised more doubts (bottom rows).

Based on the empirical observation of the typical appearances of the subjects that produced the extreme $\varphi_{i,c}(x)$ statistics, we highlight the following factors:

- The hair style is particularly important to humans, when being asked to identify an individual. Obviously, this raises the concerns about counterfeiting measures from criminals, in order to prevent identification.
- Particular localized marks (e.g., spots) are also specially helpful to humans, when attempting to identify others.



Fig. 6 Screenshot of the web-based application (*PortraitPad* used to create the set of sketches used in our experiments).

- At a secondary level, the shapes and sizes of eyes, nose and lips also play an important role in the reliability of the resulting sketches. Every time a subject had a particularly big/small nose, ears,...., volunteers had no difficulty in matching the sketch against the correct identity.
- The texture of the skin is particularly hard to reproduce in a sketch. In several circumstances, the volunteers found extremely hard to define the skin texture of a sketch, which they thought to maximally resemble the identity of the person of interest.

In order to assess the levels of correlation between the responses given by the human volunteers and by the automated methods, the agreement of their answers was measured. It was assumed that any eventual dependence between responses would be at most linear, which justifies the use of the phi coefficient ϕ . This coefficient, also known as *mean square contingency coefficient* is a measure of association between two binary variables. Let X_i and Y_i be two random variables that denote the rank-1 identity of the i^{th} probe, given by experts X and Y . The ϕ coefficient between X and Y is given by:

$$\phi = \frac{\sum_i \mathbb{1}_{\{X_i=Y_i\}}}{\sum_i \mathbb{1}_{\{X_i=Y_i\}} + \sum_i \mathbb{1}_{\{X_i \neq Y_i\}}} \quad (16)$$

where $\mathbb{1}$ is a characteristic function, X_i and Y_i denote the system outputs, \bar{X}, \bar{Y} are the sample means and σ_X, σ_Y the standard deviations.

During the experiments we noted that the agreement between experts varies strongly with respect to the degree of reliability of the sketch. In case of sketches evidently similar to a single identity, both humans and algorithms tend to agree on their answer, whereas in cases where the identity of the sketch is not so evident, automated methods and humans often disagree. Hence, resembling the *sparsity co-*

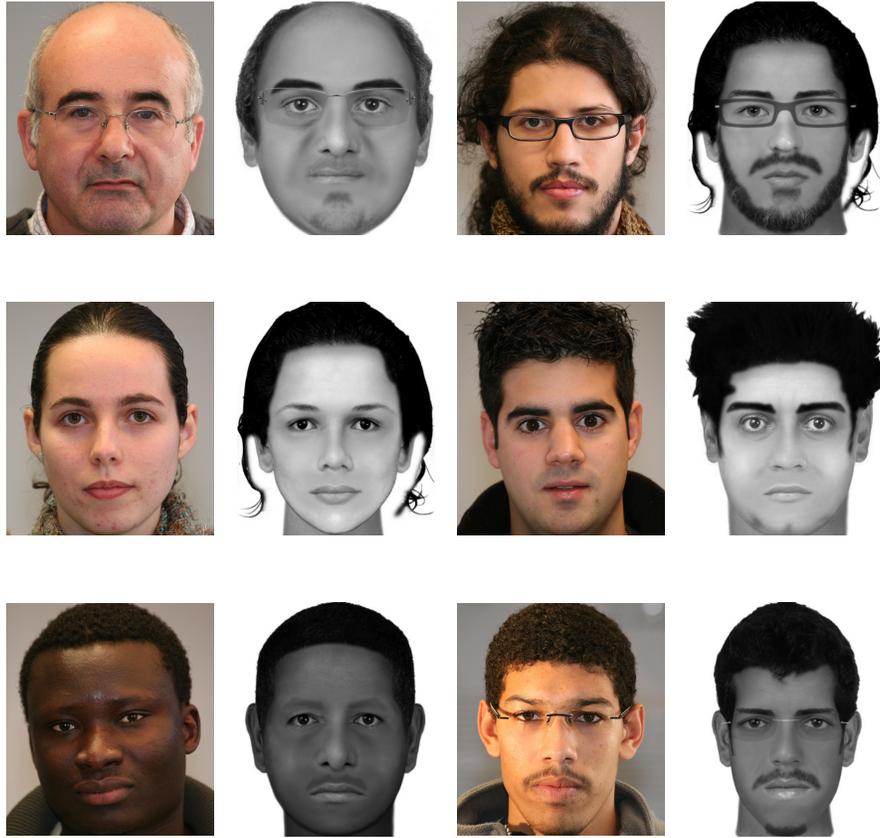


Fig. 7 Examples of the sketches and of the identities in gallery data that they aim to resemble.

efficient index, the following function measures the certainty of the classification of the i^{th} sketch by humans:

$$Q(i) \doteq \frac{k \|\max_c \varphi_{i,c}(x)\|_1 / \|\varphi_{i,c}(x)\|_1 - 1}{k - 1}, \tag{17}$$

where k denotes the number of classes (100 in our experiments). $Q(i) \approx 0$ corresponds to complete uncertainty and $Q(i) = 1$ occurs when all the human volunteers matched the i^{th} sketch against the same identity. Hence, the value of $Q(\cdot)$ enables to perceive how the uncertainty of humans is related to the agreement of the responses of automated recognition algorithms. Figure 9 expresses the relationship between the value of $Q(I)$ and the ϕ coefficient, showing broad evidence about a direct relationship between both values. The Gabor-based description technique was observed to be the algorithm that maximally resembles the responses given by humans (*HG*), specially for highly reliable sketches. In general, moderate to low levels of agree-

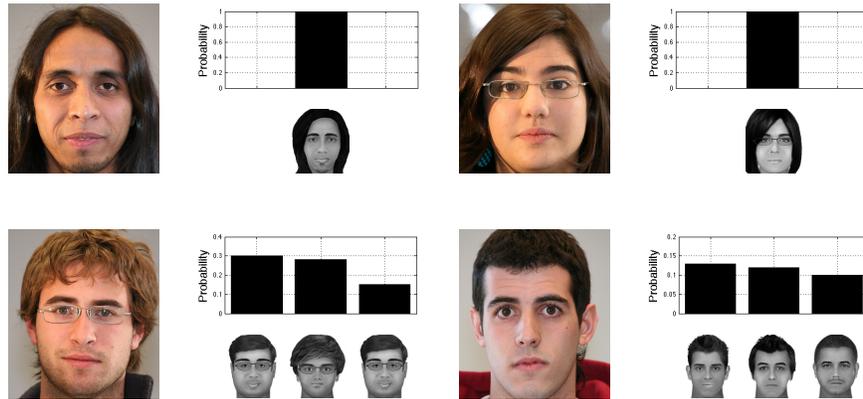


Fig. 8 Examples of extreme cases of the $\varphi_{r,c}(x)$ statistic, corresponding to cases where the human observers had minimal (upper row) and maximal uncertainty (bottom row).

ment between humans and automated methods were observed: for Q values lower than 0.3 the agreement between both types of experts was residual. Oppositely, values raise significantly for high $Q(\cdot)$ values, as it is evident in the rightmost histogram of the figure.

In summary, it can be concluded that Humans and the *Gabor-based* algorithm have a rough linear relationship in their levels of correlation, with respect to the level of reliability of the sketches, i.e., as the reliability of the sketches augments, there is a substantially higher probability that the response attained by the automated algorithm and of a human is the same. At the other extreme, the correlation of the responses given by the eigenfaces algorithms was always low, both for sketches of low, medium or high reliability.

3.7 Performance of Automated Methods

The performance of the automated face recognition methods was tested with respect to the level of agreement of humans in correctly identifying sketches, using different $Q(i)$ minimal values. Figure 10 gives the receiver operating characteristic curves observed for each algorithm, when using $Q(i) \geq 0.0$ (upper left plot), $Q(i) \geq 0.5$ (upper right) and $Q(i) \geq 0.75$ (bottom plot). The sparse representation method corresponds to the continuous lines, the Eigenfaces technique is represented by the dashed lines and the Gabor-based technique by the dotted lines. A relatively poor performance was observed when all sketches and identities were considered (i.e., $Q \geq 0$). In this case, each algorithm outperformed in different regions of the performance space. However, as the degree of evidence of the sketches' identity increases ($Q \geq 0.5$ and $Q \geq 0.75$) the performance attained by the sparse representation technique starts to

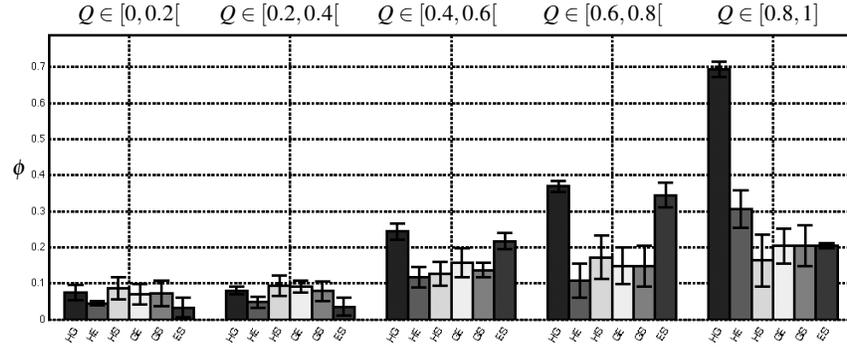


Fig. 9 Relationship between the agreement of the responses given by the automated sketch recognition strategies (G denotes the Gabor-based recognition, S regards the sparse representation technique and E the Eigenfaces method) and human responses (denoted by H). XY is the agreement between the responses given by X and Y experts). Results are given with respect to the levels of uncertainty $Q(\cdot)$ of human observers in matching sketches to identities in gallery data (17).

be consistently the best. In the case of sketches with maximum reliability ($Q \geq 0.75$) the difference to the remaining methods is evident, and full sensitivity is attained at a false positive rate of about 0.25. In the case of the eigenfaces and Gabor-based recognition methods, the improvements in performance also occurred, but not in such an evident way.

In a subsequent evaluation, we evaluated whether the translation of the gallery data into the sketch domain carries significant advantages over the comparison *image-against-sketch*. All the gallery images were converted to sketches using the *AKVIS* software package², which was considered to faithfully resemble the appearance of the sketches among the options freely available. Here, as illustrated in Figure 11, the idea is that, by converting first the photos into the sketch domain, the amount of information is substantially reduced and only the most representative features of each face are retained.

Figure 10 gives the results obtained for the three techniques tested. In the case of the sparse representation technique, the translation of the gallery data into the sketches domain didn't not carried any advantage and even contributed for decreases in performance, independently of the reliability of the sketched considered. In opposition, for the Gabor-based description and matching method, improvements in performance were consistently observed at all levels of sketches' reliability. Finally, differences in performance in any direction were not considered evident in the case of the eigenfaces recognition method, as performance in the sketch domain was observed to be best than in original data domain only for $Q \geq 0.75$, corresponding to the sketches that more evidently match a single identity.

² <http://akvis.com/en/sketch/index.php>

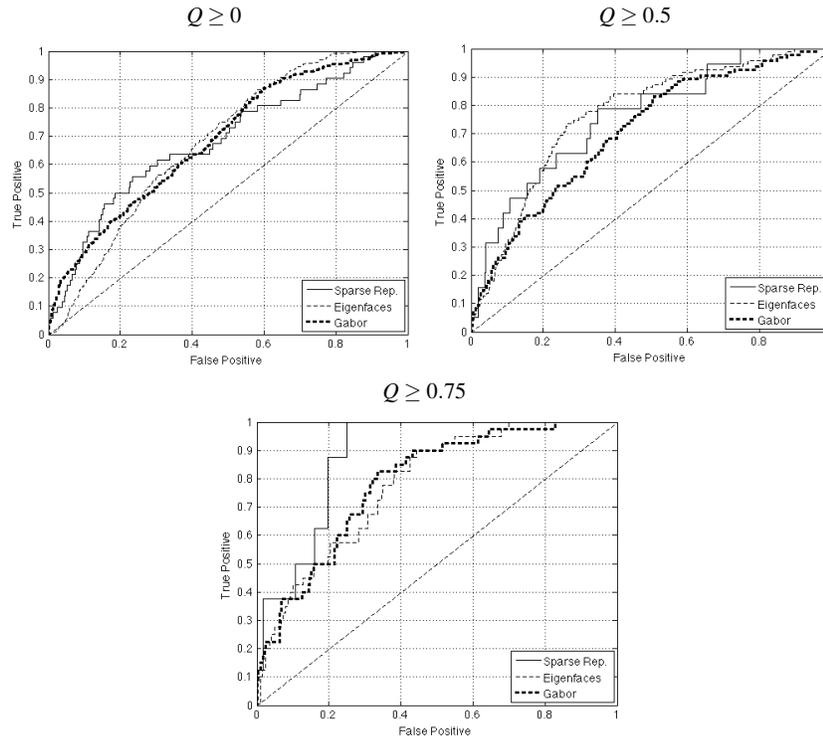


Fig. 10 Comparison of the performance attained by the sparse representation algorithm, eigenfaces and Gabor description algorithms, with respect to the minimal level of agreement of humans in matching a sketch to an identity (minimum value of $Q(i)$ demanded).

As a summary of the experiments carried out, it can be concluded that the intuitive notion of *reliability* of a sketch plays a relevant role in the performance of automated face identification algorithms. Also, the performance of such algorithms significantly improves for reliable sketches, i.e., those that most humans perceive as with high quality ($Q \geq 0.75$). In opposition, there is no significant differences in the performance of automated recognition algorithms among sketches of poor and medium fidelities. When it comes to the algorithms used, it can be concluded that sparse representations are the best technique for high reliability sketches. This observation is in agreement to the functioning of that algorithm, requiring that gallery elements constitute a vector basis (or close to that) of the query data, which should not happen for very bad sketches. At the other extreme, the eigenfaces algorithm outperforms all the other for sketches of moderate reliability. This should be due to the fact that, in opposition to the remaining algorithms, the most important eigenvectors of real faces and sketches of moderate quality still resemble. In this case,

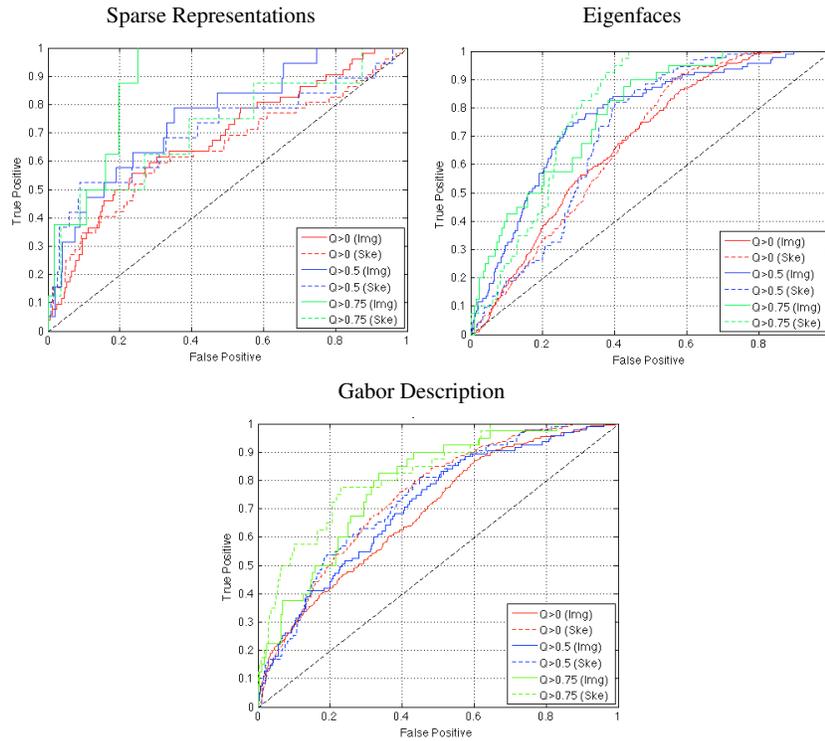


Fig. 11 Comparison of the performance attained when using the original gallery data and when translating it into the sketch domain before the recognition phase (*Img* series represent the comparison image-against-sketch, whereas *Ske* denote the translation into the sketch domain of all gallery data, before the recognition process).

for *better* data, the sparse representation is clearly better, but for *worst* data, the eigenvectors found are not trustworthy anymore.

4 Conclusions and Future Trends

For decades, police and forensic organizations have been searching for potential criminals based on descriptions given by eyewitnesses of some illegal act. However, it is usual that witnesses only observe the criminals for a few seconds, are too nervous, or that the criminals' faces were partially occluded, which decreases the *reliability* of the resulting sketches, i.e., *how much* the sketch and the appearance of the corresponding subject *actually resemble*.

The availability of software frameworks to generate composite facial sketches enables to obtain close representations of real mugshots, which motivated the work

described in this chapter. We assessed the recognition performance attained by three face recognition methods and compared it to the attained by humans in matching composite sketches to real mugshots. Also, the analysis considers the levels of reliability of each sketch, measured by the agreement of humans in matching each sketch against a single identity. Overall, we concluded that the sparse representation for classification algorithm obtained the best results, which was particularly evident for high reliable sketches.

Also, we were particularly concerned about the effects in performance when sketches do not appropriately represent the main features of the identity of interest. i.e., in case of sketches of poor reliability. A set of one hundred sketches was synthesized by non-experts, using a freely available facial composite software system. Then, human volunteers were asked to match each sketch to one of the enrolled identities. When comparing the responses given by humans and by the automated recognition algorithms, we concluded that they agree mostly in cases where the sketch has particular features that turn evident the corresponding identity. Also, the rank-1 identification performance of automated methods is strongly conditioned to the *degree of evidence* of the sketch. Finally, we assessed the improvements in performance due to the translation of the gallery data into the sketch domain before the recognition process. Improvements were not statistically significant in the case of sparse representation and eigenfaces techniques, in opposition to the Gabor-based description and matching strategy, where a consistent improvement was observed.

As a summary of the research conducted in the scope of this chapter, we highlight:

- the *reliability* of the sketches is a major concern, and both automated recognition algorithms and humans have very low effectiveness on bad quality sketches;
- the Gabor-based recognition algorithm produces the responses that more closely resemble those from humans. This is specially evident for sketches of high reliability;
- In general, the sparse representation for classification algorithm produced the best performance, in terms of rank-1 identification accuracy;
- the conversion of the gallery data into the sketches domain not always carries advantages in terms of recognition performance. In several cases, we even observed decreases in the recognition effectiveness.

As future trends for this research topic, it will be important to investigate about automated methods that are able to estimate the reliability of a sketch of unknown identity, which can be helpful to assist the eyewitnesses in describing a criminal. Such quality marks of sketches should be regarded in terms of *discriminating points*, i.e., information that is actually helpful for automated identification experts in distinguishing among a set of identities.

Acknowledgements The financial support given by "FCT-Fundação para a Ciência e Tecnologia" and "FEDER" in the scope of the PTDC/EIA-EIA/103945/2008 ("NECOVID: Negative Covert Biometric Identification") research project is acknowledged.

Also, we acknowledge all the volunteers that participated in the data acquisition sessions and in the synthesis of the facial sketches used in our experiments.

References

1. Chen C., Donoho D., Saunders M.: Atomic Decomposition by Basis Pursuit. *SIAM Review*, 43 (1), 129–159 (2001)
2. Cutler L.B., Penrod S.D., Fisher R.P.: Conceptual, practical and empirical issues associated with eyewitness identification test media. Ross D., Read J., Toglia M. (eds.) *Adult eyewitness testimony: Current trends and developments*, pp. 163–181. New York: Cambridge University Press (1994)
3. Daugman J.: Complete Discrete 2-D Gabor Transforms by Neural Networks for Image Analysis and Compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36 (7), 1169–1179 (1988)
4. Daugman J.: Two-dimensional spectral analysis of cortical receptive field profiles, *Vision Research*, 20 (10): 847–56 (1980)
5. Donoho D.: For Most Large Underdetermined Systems of Linear Equations the Minimal l_1 -Norm Solution Is Also the Sparsest Solution. *Comm. Pure and Applied Math*, 59 (6), 797–829 (2006)
6. Dunn D., Higgins W.E.: Optimal Gabor filters for texture segmentation. *IEEE Transactions on Image Processing*, 4 (7), 947–964(1995)
7. Fisher G.: The Jury's Rise as a Lie Detector. *Yale Law Journal*, 575 (1997)
8. Gurney D., Pine K., Wiseman R.: The gestural misinformation effect: skewing eyewitness testimony through gesture. *American Journal of Psychology*, 126 (3), 301-314 (2013)
9. Han H., Klare B.F., Bonnen K. , Jain A.K.: Matching Composite Sketches to Face Photos. *IEEE Transactions on Information Forensics and Security*, 8 (1), 191–204 (2013)
10. Hancock P., Bruce V., Burton A.: Comparing two computer systems and human perceptions of faces. *Vision Research*, 38, 2277–2288 (1998)
11. Horry R., Halford P., Brewer N., Milne R., Bull R.: Archival Analyses of Eyewitness Identification Test Outcomes: What Can They Tell Us About Eyewitness Memory? *Law and Human Behavior*, (in press) (2013)
12. Jain A.K., Klare B.F.: Matching Forensic Sketches and Mug Shot to Apprehend Criminals. *Computer*, 44 (5), 94–96 (2011)
13. Jain A.K., Chen Y., Demirkus M.: Pores and ridges: Fingerprint matching using level 3 features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29 (1), 15–27 (2007)
14. Kamarainen J.-K., Kyrki V. Kalviainen H.: Invariance properties of Gabor filter-based features overview and applications. *IEEE Transactions on Image Processing*, 15 (5), 1088–1099 (2006)
15. Klare B.F., Li Z., Jain A.K.: Matching Forensic Sketches to Mug Shot Photos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33 (3), 639–646 (2011)
16. Odnot G., Memon A., la Rood A., Millen A.: . Are Two Interviews Better Than One? Eyewitness Memory across Repeated Cognitive Interviews. *PLoS One*, 8 (10), e76305 (2013)
17. Li Y-H. , Savvides M., Bhagavatula V.: Illumination Tolerant Face Recognition Using a Novel Face From Sketch Synthesis Approach and Advanced Correlation Filters. In Proceedings of the *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2, II (2006)
18. Liu Q., Tang X.,Jin H., Lu H., Ma S.: A Nonlinear Approach for Face Sketch Synthesis and Recognition. In Proceedings of the *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1, 1005–1010 (2005)
19. Nejati H., Sim T.: Do You See What I See? A More Realistic Eyewitness Sketch Recognition. In Proceedings of the *International Joint Conference on Biometrics*, pp. 1–8 (2011)
20. Pillai J., Patel V., Chellapa R., Ratha N.: Secure and Robust Iris Recognition Using Random Projections and Sparse Representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33 (9), 1877–1893 (2011)
21. Pramanik S.: Geometric Feature Based Face-Sketch Recognition. In Proceedings of the *International Conference on Pattern Recognition, Informatics and Medical Engineering* , 409–415 (2012)

22. Rush E., Quas J., Yim I., Nikolayev M., Clark S., Larson R.: Stress, Interviewer Support, and Children's Eyewitness Identification Accuracy. *Child Development*, (in press) (2013)
23. Shen L., Bai L.: A review on Gabor wavelets for face recognition. *Pattern Analysis and Applications*, 9, 273–292 (2006)
24. Turk M., Pentland A.: Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3 (1), 71–86 (1991)
25. Tversky B., Marsh E.: Biased Retellings of Events Yield Biased Memories. *Cognitive psychology*, 40 (1), 1–38 (2000)
26. Vinay, K.; Shreyas, B.: Face Recognition Using Gabor Wavelets. in Proceedings of the *Fortieth Asilomar Conference on Signals, Systems and Computers*, 1955–1967 (2006)
27. Wang X., Tang X.: Face Photo-Sketch Synthesis and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31 (11), 1955–1967 (2009)
28. Wright J., Yang A., Ganesh A., Sastry S, Ma Y.: Robust Face Recognition via Face Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31 (2), 210–227 (2009)