Recognition of Facial Attributes using Adaptive Sparse Representations of Random Patches

Domingo Mery[1] and Kevin Bowyer[2]

[1] Department of Computer Science
Pontificia Universidad Católica de Chile
http://dmery.ing.puc.cl

[2] Department of Computer Science & Engineering
University of Notre Dame
http://www.nd.edu/~kwb

Abstract. It is well known that some facial attributes –like soft biometric traits– can increase the performance of traditional biometric systems and help recognition based on human descriptions. In addition, other facial attributes –like facial expressions– can be used in human–computer interfaces, image retrieval, talking heads and human emotion analysis. This paper addresses the problem of automated recognition of facial attributes by proposing a new general approach called Adaptive Sparse Representation of Random Patches (ASR+). In the learning stage, random patches are extracted from representative face images of each class (*e.g.*, in gender recognition –a two-class problem–, images of females/males) in order to construct representative dictionaries. In the testing stage, random test patches of the query image are extracted, and for each test patch a dictionary is built concatenating the 'best' representative dictionary of each class. Using this adapted dictionary, each test patch is classified following the Sparse Representation Classification (SRC) methodology. Finally, the query image is classified by patch voting. Thus, our approach is able to learn a model for each recognition task dealing with a larger degree of variability in ambient lighting, pose, expression, occlusion, face size and distance from the camera. Experiments were carried out on seven face databases in order to recognize facial expression, gender, race and disguise. Results show that ASR+ deals well with unconstrained conditions, outperforming various representative methods in the literature in many complex scenarios.

Keywords: sparse representation, soft biometrics, gender recognition, race recognition, facial expression recognition.

1 Introduction

Automated recognition of facial attributes has been a relevant area in computer vision, making many important contributions since the 1990s (see for example

[19]). The relevance of this research field is twofold: First, the use of facial attributes, like soft biometric traits (*e.g.*, gender [24], race [8], age [9], etc.), can increase the performance of traditional biometric systems [26] and help recognition based on human descriptions [28]. Second, other facial attributes, like facial expressions, can be used in human–computer interfaces, image retrieval, talking heads and human emotion analysis [36].

Usually, each single facial attribute has been recognized by a specific algorithm. Some examples are the following: a) Gender is identified using a SVM classifier with Gaussian RBF kernel [22], a Real AdaBoost classifier with texture features [35], an AdaBoost classifier with a low resolution image [3], and a SVM classifier of PCA representations [16]. b) Facial expressions are classified using a new feature called 'supervised locally linear embedding' [14], a decomposition into multiple two-class classification problems with 'salient feature vectors' [13], local binary patterns [29], a boosted deep belief network [15], active facial patches [37], and Gabor features [4]. c) Race is recognized using biologically inspired features [11], an ensemble framework with LDA [17], a probabilistic graphical model [23] and local binary patterns with wavelets features [27].

There are few approaches to estimate age, gender and race together (see for example [12]), however, to the best knowledge of the authors, there has been no reported approach, that can be used to recognize facial attributes in general. We believe that algorithms based on sparse representations can be used for this task because in many computer vision applications, under assumption that natural images can be represented using sparse decomposition, state-of-the-art results have been significantly improved [31]. Algorithms based on Sparse Representation Classification (SRC) [33] have been widely used in face recognition. In the sparse representation approach, a dictionary is built from the gallery images, and matching is done by reconstructing the query image using a sparse linear combination of the dictionary. The identity of the query image is assigned to the class with the minimal reconstruction error. Several variations of this approach were recently proposed. In [34], a sparse representation in two phases is proposed. In [7], sparse representations of patches distributed in a grid manner are used. These variations improve recognition performance as they are able to model various corruptions in face images, such as misalignment and occlusion.

Reflecting on the problems confronting recognition of facial attributes, we believe that there are some key ideas that should be present in new proposed solutions. First, it is clear that certain parts of the face are not providing any information about the class to be recognized. For this reason, such parts should be detected and should not be considered by the recognition algorithm. Second, in recognizing any class, there are parts of the face that are more relevant than other parts (for example the mouth when recognizing an expression like happiness). For this reason, relevant parts should be class-dependent, and could be found using unsupervised learning. Third, in the real-world environment, and given that face images are not perfectly aligned and the distance between camera and subject can vary from capture to capture, analysis of fixed sub-windows can lead to misclassification. For this reason, feature extraction should not be

in fixed positions, and can be in several random positions, and use a selection criterion that enables selection of the best regions. Fourth, the expression that is present in a query face image can be subdivided into 'sub-expressions', for different parts of the face (*e.g.*, eyebrows, nose, mouth). For this reason, when searching for images of the same class it would be helpful to search for image parts in all images of the gallery instead of similar gallery images.

Inspired by these key ideas, we propose a new general method for recognition of facial attributes. Three main contributions of our approach are: 1) A new general algorithm that is able to recognize a wide range of facial attributes: it has been evaluated in the recognition of expressions, gender, race and disguise obtaining a performance at least comparable with that achieved by state-of-art techniques. 2) A new representation for the classes to be recognized: this is based on representative dictionaries learned for each class of the gallery images, which correspond to a rich collection of representations of selected relevant parts that are particular to a specific class. 3) A new representation for the query face image: this is based on *i*) a discriminative criterion that selects the best test patches extracted randomly from the query image and *ii*) and an 'adaptive' sparse representation of the selected patches computed from the 'best' representative dictionary of each class. Using these new representations, the proposed method (ASR+) can achieve high recognition performance under many complex conditions, as shown in our extensive experiments.

The rest of the paper is organized as follows: in Section 2, the proposed method is explained in further detail. In Section 3, the experiments and results are presented. Finally, in Section 4, concluding remarks are given.

2 Proposed Method

According to the motivation of our work, we believe that facial attributes can be recognized using a patch-based approach. Thus, following a sparse representation methodology, in a learning stage a number of random patches can be extracted from each training image, and a dictionary can be built for each class by concatenating its patches (stacking in columns). In the testing stage, several patches can be extracted and each of them can be classified using its sparse representation. The final decision can be made by majority vote. This baseline approach, however, shows four important disadvantages: *i*) The location information of the patch is not considered, *i.e.*, a patch of one part of the face could be erroneously represented by a patch of a different part of the face. This first problem can be solved by considering the (x, y) location of the patch in its description. *ii)* The method requires a huge dictionary for reliable performance, *i.e.*, each sparse representation process would be very time consuming. This second problem can be remedied by using only a part of the dictionary *adapted* to each patch. Thus, the whole dictionary of a class can be subdivided into sub-dictionaries, and only the 'best' ones can be used to compute the sparse representation of a patch. *iii*) Not all query patches are relevant, *i.e.*, some patches of the face do not provide any discriminative information of the class (*e.g.*, sunglasses when identifying

gender). This third problem can be addressed by selecting the query patches according to a score value. *iv*) It is likely that many images of different classes has common patches, such as similar skin textures when identifying gender, which occur in most faces of all classes and are therefore not discriminating for a particular class. This fourth issue can be addressed using a text retrieval approach including a *visual vocabulary* and a *stop list* to reject those common words [30].

In this section we describe our approach taking into account the four mentioned improvements. As illustrated in Fig. 1, in the learning stage, for each class of the gallery, several random small patches are extracted and described from their images (using both intensity and location features). However, only those patches that are not filtered out by the stop list are considered to build representative dictionaries. In the testing stage, random test patches of the query image are extracted and described. A patch that belongs to the stop list is not considered. For each (considered) test patch a dictionary is built concatenating the 'best' representative dictionary of each class. Using this adapted dictionary, each test patch is classified in accordance with the Sparse Representation Classification (SRC) methodology [33]. Afterwards, the patches are selected according to a discriminative criterion. Finally, the query image is classified by voting for the selected patches. Both stages will be explained in this section in further detail.

2.1 Learning

In the training stage, a set of n face images of k classes is available, where $\mathbf{I}_j^i$ denotes image j of class i (for $i = 1 \dots k$ and $j = 1 \dots n$). In each image $\mathbf{I}_j^i$, m patches are randomly extracted. In this work, the description of a patch $\mathcal{P}$ is defined as vector:

$$\mathbf{y} = f(\mathcal{P}) = [\ \mathbf{z}\ ;\ \alpha x\ ;\ \alpha y\ ] \in \mathcal{R}^{d+2} \tag{1}$$

where $\mathbf{z} = g(\mathcal{P}) \in \mathcal{R}^d$ is a descriptor of patch $\mathcal{P}$; (x, y) are the image coordinates of the center of patch $\mathcal{P}$; and α is a weighting factor between description and location[1].Using (1) all extracted patches are described as $\mathbf{y}_{jp}^i = f(\mathcal{P}_{jp}^i) = [\ \mathbf{z}_{jp}^i\ ;\ \alpha x_{jp}^i\ ;\ \alpha y_{jp}^i\ ]$, for $p = 1 \dots m$.

In order to eliminate non-discriminative patches, a *stop list* is computed from a *visual vocabulary*. The visual vocabulary is built using all descriptors $\mathbf{Z} = \{\mathbf{z}_{jp}^i\} \in \mathcal{R}^{d \times knm}$, for $i = 1 \dots k$, for $j = 1 \dots n$ and for $p = 1 \dots m$. Array $\mathbf{Z}$ is clustered using a k-means algorithm in N_v clusters. Thus, a visual vocabulary $\mathcal{V}$ containing N_v visual words is obtained. In order to construct the stop list, the *term frequency* 'tf' is computed: $\mathsf{tf}(d, v)$ is defined as the number of occurrences of word v in document d, for $d = 1 \dots K$, $v = 1 \dots N_v$. In our case, a document corresponds to a face image, and $K = kn$ is the number of faces in the gallery. Afterwards, the *document frequency* 'df' is computed: $\mathsf{df}(v) = \sum_d \{\mathsf{tf}(d, v) > 0\}$,

[1] In our experiments, the size of the patch is $w \times w$. The descriptor $\mathbf{z}$ corresponds to the intensity values of the patch subsampled by 2 in both directions, *i.e.*, $d = (w \times w)/4$ given by stacking its columns normalized to unit length in order to deal with different illumination conditions; (x, y) are normalized coordinates (values between 0 and 1).

i.e., the number of faces in the gallery that contain a word v, for $v = 1 \dots N_v$. The stop list is built using words with highest and smallest df values: On one hand, visual words with highest df values are not discriminative because they occur in almost all images. On the other hand, visual words with smallest df are so unusual that they correspond in most of the cases to noise. Usually, the top 5% and bottom 10% are stopped [30]. Those patches of $\mathbf{Z}$ that belong to the stopped clusters are not considered in the following steps of our algorithm.

Now, for class i an array with the description of all (non stopped) patches $\mathbf{y}_{jp}^i$ is defined as $\mathbf{Y}^i$.The description $\mathbf{Y}^i$ of class i is clustered using a k-means algorithm in Q clusters that will be referred to as *parent* clusters:

$$\mathbf{c}_q^i = \text{kmeans}(\mathbf{Y}^i, Q) \tag{2}$$

for $q = 1 \dots Q$, where $\mathbf{c}_q^i \in \mathcal{R}^{(d+2)}$ is the centroid of parent cluster q of class i. We define $\mathbf{Y}_q^i$ as the array with all samples $\mathbf{y}_{jp}^i$ that belong to the parent cluster

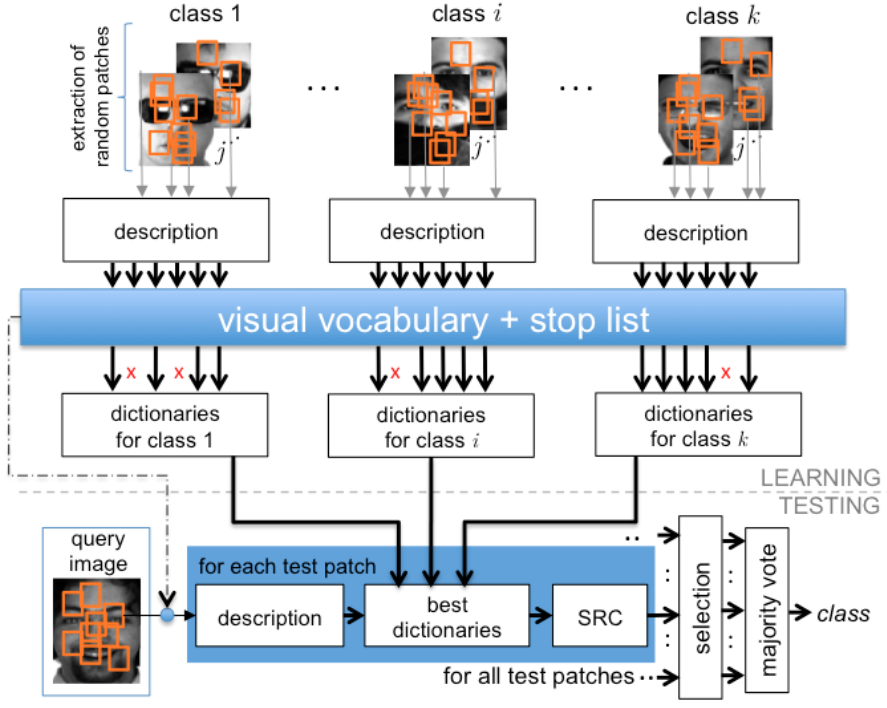

Fig. 1. Overview of the proposed method. The figure illustrates the recognition of disguise. The shown classes are three: sunglasses, scarf and no-disguise. The stop list is used to filter out patches that are not discriminating for these classes. The stopped patches are not considered in the dictionaries of each class and in the testing stage.

with centroid $\mathbf{c}_q^i$. In order to select a reduced number of samples, each parent cluster is clustered again in R *child* clusters[2]

$$\mathbf{c}_{qr}^i = \text{kmeans}(\mathbf{Y}_q^i, R) \tag{3}$$

for $r = 1 \dots R$, where $\mathbf{c}_{qr}^i \in \mathcal{R}^{(d+2)}$ is the centroid of child cluster r of parent cluster q of class i. All centroids of child clusters of class i are arranged in an array $\mathbf{D}^i$, and specifically for parent cluster q are arranged in a matrix:

$$\bar{\mathbf{A}}_q^i = [\mathbf{c}_{q1}^i \dots \mathbf{c}_{qr}^i \dots \mathbf{c}_{qR}^i]^\mathsf{T} \in \mathcal{R}^{(d+2)\times R} \tag{4}$$

Thus, this arrangement contains R representative samples of parent cluster q of class i as illustrated in Fig. 2. The set of all centroids of child clusters of class i ($\mathbf{D}^i$), represents Q representative dictionaries with R descriptions $\{\mathbf{c}_{qr}^i\}$ for $q = 1 \dots Q, r = 1 \dots R$.

2.2 Testing

In the testing stage, the task is to determine the class of the query image $\mathbf{I}^t$ given the model learned in the previous section. From the test image, s selected test patches $\mathcal{P}_p^t$ of size $w \times w$ pixels are extracted and described using (1) as $\mathbf{y}_p^t = f(\mathcal{P}_p^t) = [\ \mathbf{z}_p^t;\ \alpha x_p^t;\ \alpha y_p^t\ ]$ (for $p = 1 \dots s$). The selection criterion of a test patch will be explained later in this section. For each selected test patch with

[2] If n_q^i, the number of samples of $\mathbf{Y}_q^i$, is less than R, $\mathbf{c}_{qr}^i$ is built by taking the R first samples of a replicated version of the samples $[\mathbf{Y}_q^i\ \mathbf{Y}_q^i \dots]$. This dictionary with R words is equivalent to have a dictionary of n_q^i words only.

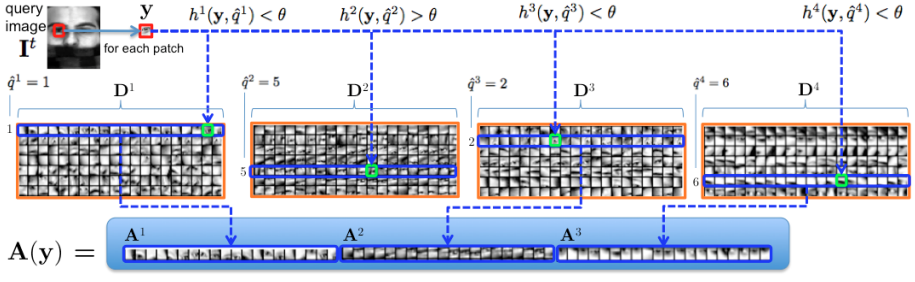

Fig. 2. Adaptive dictionary $\mathbf{A}$ of patch $\mathbf{y}$. In this example there are $k = 4$ classes in the gallery. For this patch only $k' = 3$ classes are selected. Dictionary $\mathbf{A}$ is built from those classes by selecting all child clusters (of a parent cluster -see blue rectangles-) which have a child with the smallest distance to the patch (see green squares). In this example, class 2 does not have child clusters that are similar enough to patch $\mathbf{y}$, *i.e.*, $h^2(\mathbf{y},\hat{q}^2) > \theta$.

description $\mathbf{y} = \mathbf{y}_p^t$, a distance to each parent cluster q of each class i of the gallery is measured:

$$h^i(\mathbf{y}, q) = \text{distance}(\mathbf{y}, \bar{\mathbf{A}}_q^i). \tag{5}$$

We tested with several distance metrics. The best performance, however, was obtained by $h^i(\mathbf{y}, q) = \min_r ||\mathbf{y} - \mathbf{c}_{qr}^i||$, which is the smallest distance to centroids of child clusters of parent cluster q as illustrated in Fig. 2. Normalizing $\mathbf{y}$ and $\mathbf{c}_{qr}^i$ to have unit ℓ_2 norm, (5) can be rewritten as:

$$h^i(\mathbf{y}, q) = 1 - \max < \mathbf{y}, \mathbf{c}_{qr}^i > \text{ for } r = 1 \dots R \tag{6}$$

where the term $< \bullet >$ corresponds to scalar product that provides a similarity (cosine of angle) between vectors $\mathbf{y}$ and $\mathbf{c}_{qr}^i$. The parent cluster that has the minimal distance is searched:

$$\hat{q}^i = \underset{q}{\text{argmin}} \; h^i(\mathbf{y}, q), \tag{7}$$

which minimal distance is $h^i(\mathbf{y}, \hat{q}^i)$. For patch $\mathbf{y}$, we select those gallery classes that have a minimal distance less than a threshold θ in order to ensure a similarity between the test patch and representative class patches. If k' classes fulfill the condition $h^i(\mathbf{y}, \hat{q}^i) < \theta$ for $i = 1 \dots k$, with $k' \leq k$, we can build a new index $v_{i'}$ that indicates the index of the i'-th selected class for $i' = 1 \dots k'$. For instance in a gallery with $k = 4$ classes, if $k' = 3$ classes are selected (*e.g.*, classes 1, 3 and 4), then the indices are $v_1 = 1$, $v_2 = 3$ and $v_3 = 4$ as illustrated in Fig. 2. The selected class i' for patch $\mathbf{y}$ has its dictionary $\mathbf{D}^{v_{i'}}$, and the corresponding parent cluster is $u_{i'} = \hat{q}^{v_{i'}}$, in which child clusters are stored in row $u_{i'}$ of $\mathbf{D}^{v_{i'}}$, *i.e.*, in $\mathbf{A}^{i'} := \bar{\mathbf{A}}_{u_{i'}}^{v_{i'}}$.

Therefore, a dictionary for patch $\mathbf{y}$ is built using the best representative patches as follows (see Fig. 2):

$$\mathbf{A}(\mathbf{y}) = [\ \mathbf{A}^1 \dots \mathbf{A}^{i'} \dots \mathbf{A}^{k'}\ ] \in \mathcal{R}^{(d+2) \times Rk'} \tag{8}$$

With this adaptive dictionary $\mathbf{A}$, built for patch $\mathbf{y}$, we can use SRC methodology [33]. That is, we look for a sparse representation of $\mathbf{y}$ using the ℓ_1-minimization approach:

$$\hat{\mathbf{x}} = \text{argmin}||\mathbf{x}||_1 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{y} \tag{9}$$

The residuals are calculated for the reconstruction for the selected classes $i' = 1 \dots k'$:

$$r_{i'}(\mathbf{y}) = ||\mathbf{y} - \mathbf{A}\delta_{i'}(\hat{\mathbf{x}})|| \tag{10}$$

where $\delta_{i'}(\hat{\mathbf{x}})$ is a vector of the same size as $\hat{\mathbf{x}}$ whose only nonzero entries are the entries in $\hat{\mathbf{x}}$ corresponding to class $v(i') = v_{i'}$. Thus, the class of selected test patch $\mathbf{y}$ will be the class that has the minimal residual, that is it will be

$$\hat{i}(\mathbf{y}) = v(\hat{i}') \tag{11}$$

where $\hat{i}' = \text{argmin}_{i'}\, r_{i'}(\mathbf{y})$. Finally, the identity of the query class will be the majority vote of the classes assigned to the s selected test patches $\mathbf{y}_p^t$, for $p = 1 \ldots s$:

$$\text{identity}(\mathbf{I}^t) = \text{mode}(\hat{i}(\mathbf{y}_1^t), \ldots \hat{i}(\mathbf{y}_p^t), \ldots \hat{i}(\mathbf{y}_s^t)) \tag{12}$$

The selection of s patches of query image is as follows:
i) From query image $\mathbf{I}^t$, m^t patches are randomly extracted and described using (1): $\mathbf{y}_j^t$, for $j = 1 \ldots m^t$, with $m^t \geq s$.
ii) Those patches $\mathbf{y}_j^t$ that belong to the stopped clusters of our visual vocabulary $\mathcal{V}$ are not considered.
iii) Each remaining patch $\mathbf{y}_j^t$ is represented by $\hat{\mathbf{x}}_j^t$ using (9).
iv) The *sparsity concentration index* (SCI) of each patch is computed in order to evaluate how spread are its sparse coefficients [33]. SCI is defined by

$$S_j := \text{SCI}(\mathbf{y}_j^t) = \frac{k \cdot \max(||\delta_{i'}(\hat{\mathbf{x}}_j^t)||_1)/||\hat{\mathbf{x}}_j^t||_1 - 1}{k - 1} \tag{13}$$

If a patch is discriminative enough it is expected that its SCI is large. Note that we use k instead of k' because the concentration of the coefficients related to k classes must be measured.
iv) Array $\{S\}_{j=1}^m$ is sorted into a descended order of SCI value. The first s patches in this sorted list in which SCI values are greater than a τ threshold are then selected. If only s' patches are selected, with $s' < s$, then the majority vote decision in (12) will be taken with the first s' patches.

3 Experimental Results

ASR+ was evaluated in the recognition of several facial attributes: facial expressions (Section 3.1), gender (Section 3.2), race (Section 3.3) and disguise (Section 3.4). Experiments were carried out on seven databases under varying conditions. We demonstrate the performance of our ASR+ approach with a combination of two types of experiments: 1) When it is possible, we compare performance of ASR+ against recent published performance results of a variety of algorithms using the database and similar experimental protocol used in the paper about each algorithm. 2) We compare performance of ASR+ to performance of five 'baseline methods'. They are re-implemented versions of five well-known general recognition algorithms that have been used in face recognition problems. In this case, the methods are the following: *i)* NBNN [6] using intensity features normalized to the unit length in 6×6 partitions, *ii)*, NBNN using LBP-based features [1] with 6×6 partitions, *iii)* SRC [33] where the images were sub-sampled to 22×18 pixels building features of dimension $d = 396$, *iv*) TPTSR based on a two-phase test sample sparse representation approach [34], and *v)* LAD [7] based on locally adaptive sparse representation of patches distributed in a grid. We coded these methods in Matlab according to the specifications given by the authors in their papers.

The used protocol –when evaluating our proposed approach and the baseline methods– is the following: In the databases, there were face images from k classes (*e.g.*, in gender recognition $k = 2$, for female and male) and more than n images per class. All face images were resized to 110 × 90 pixels and converted to a grayscale image if necessary. From each class, n images were randomly chosen for training and one for testing. In order to obtain a better confidence level in the accuracy, the test was repeated N times by randomly selecting $n + 1$ faces images per class each time. The reported accuracy η in all of our experiments is the average calculated over the N tests. In order to report the number of training images and runs of each experiment, we use the notation '$[n|N]$'.

In addition, we report other parameters of our method that depend on the alignment of the face images, the number of training images and the size of the local information of the face that is used in the recognition task. They are the number of parent and child clusters (Q and R), the number of patches extracted in each training image (m), the weighting factor for location coordinates (α), the size of patches (w) and the size of the visual vocabulary (N_v). We use the notation '$\{Q, R, m, \alpha, w, N_v\}$'.

3.1 Facial Expression

The performance of our method was evaluated on three databases: *i)* JAFFE database [20]: It contains 7 expressions ('neutral' and six basic emotions: 'anger', 'disgust', 'fear', 'happiness', 'sadness' and 'surprise') captured from 10 Japanese women. For each subject, there are 3–4 face images for the non-neutral and one for the neutral expressions, *i.e.*, the database consists of 213 images. Results are summarized in Tab. 1. In our case, we used $[n = 29|N = 50]$ and $\{Q = 100, R = 80, m = 250, \alpha = 3, w = 40, N_v = 400\}$. *ii)* CK+ database [18]: It consists of 8 expressions ('contempt' was added to the six basic emotions) captured from 100 subjects as sequences (starting with a neutral face and ending with the peak of a facial expression). In order to compare our method with other methods fairly, a common experimental protocol was followed: The first frame of the sequence (neutral face) and the three last frames (emotion faces) were used. Experiments were carried out to recognize the 6 basic emotions using a leave-one out strategy. Results are summarized in Tab. 1. In our case, we used $[n = 74|N = 50]$ and $\{Q = 100, R = 80, m = 120, \alpha = 0.25, w = 18, N_v = 400\}$. *iii)* SmileFlick (own database): In this experiment, the idea was to detect smiling faces. For this end, 52 face images with smile and 57 face images with neutral expression were collected manually from frontal portraits published in Flickr including subjects from different age, race, gender and illumination. The faces were detected automatically using Computer Vision Toolbox of Matlab[3]. In our experiments, we used $[n = 49|N = 60]$ and $\{Q = 80, R = 50, m = 300, \alpha = 3, w = 40, N_v = 400\}$. The results of our method compared with the baseline methods are summarized in Tab. 1.

[3] http://www.mathworks.com/products/computer-vision/

Table 1. Recognition of Expressions

Database	Method	Ref	η [%]
JAFFE	SLLE	[12]	86.8^{+}
	SFRCS	[13]	86.0^{+}
7	Ada+SVM(RBF)	[14]	81.0^{+}
classes	$BDBN_J$	[15]	91.8^{+}
	$BDBN^{*}_{J+C}$	[15]	93.0^{+}
	ASR+	(ours)	94.3
CK+	CSPL	[16]	89.9^{+}
	CPL	[16]	88.4^{+}
6	AdaGabor	[17]	93.3^{+}
classes	LBPSVM	[14]	95.1^{+}
	BDBN	[15]	96.7^{+}
	ASR+	(ours)	97.5
SmileFlick	NBNN	[28]	73.1
	LBP	[29]	87.5
2	SRC	[24]	96.8
classes	TPTSR	[25]	91.2
	LAD	[26]	97.5
	ASR+	(ours)	97.5

(*): It was improved using CK+ database.
(+): Result from cited paper.

Table 2. Recognition of Gender

Database	Method	Ref	η [%]
FERET	SVM-RBF	[8]	96.6^{+}
	Real AdaBoost	[9]	93.8^{+}
2	AdaBoost	[10]	94.4^{+}
classes	2DPCA-SVM	[11]	94.8^{+}
	ASR+	(ours)	95.0
GROUPS	NBNN	[28]	84.2
	LBP	[29]	83.3
2	SRC	[24]	86.9
classes	TPTSR	[25]	85.8
	LAD	[26]	87.5
	ASR+	(ours)	93.3

(+): Result from cited paper. Evaluation protocols are not exactly the same (see text).

Table 3. Recognition of Race

Database	Method	Ref	η [%]
WebRace	NBNN	[28]	61.3
	LBP	[29]	63.0
5	SRC	[24]	62.0
classes	TPTSR	[25]	65.3
	LAD	[26]	85.7
	ASR+	(ours)	87.1

Table 4. Recognition of Disguise

Database	Method	Ref	η [%]
AR	NBNN	[28]	97.8
	LBP	[29]	96.1
3	SRC	[24]	98.3
classes	TPTSR	[25]	97.8
	LAD	[26]	96.7
	ASR+	(ours)	97.8

3.2 Gender

The performance of our method was evaluated on two databases: *i)* FERET database [25]: It contains more than 3,500 face images from women and men (with different races such as African, Asian and Caucasian) involving different expressions and illumination conditions. We used a subset of 1,050 images (602 male and 448 female) where each subject has only one image. We used $[n = 440|N = 200]$ and $\{Q = 160, R = 80, m = 120, \alpha = 3, w = 36, N_v = 200\}$. Results are summarized in Table 2. In order to compare the performance of our approach, Table 2 shows the results obtained by other state-of-art methods, however, the evaluation protocols are not exactly the same. In [22], 1,044 males and 711 females were tested and the accuracy was estimated using a five-fold cross validation strategy. In [35], 3,529 images were used and the accuracy was estimated using a five-fold cross validation strategy. In [3], 2,409 images were used and 80% was used for training and 20% for testing ensuring that images of a particular individual appear only in the training set or test set. In [16], 400 males and 400 females were used and the accuracy was estimated using a five-fold cross validation strategy[4]. *ii)* GROUPS database [10]: It consists of 28,231

[4] There are other experiments on FERET database reported in the literature that are not included in Tab. 2 because the testing protocols are significantly different: In [3], there is an experiment where a subject may appear in both train and test set (in

face images collected from Flickr images. It is a real–world database containing several facial expressions, face poses, illumination conditions and races. We used the labeled data contained in 'MATLAB DATA' file with 1978 face images (946 males and 1032 females). We used in this case $[n = 700|N = 100]$ and $\{Q = 80, R = 50, m = 80, \alpha = 3, w = 16, N_v = 200\}$. Results are summarized in Tab. 2. Our method is compared with the basis methods[5].

3.3 Race

For human beings it is very difficult to distinguish a race, because it depends on how people self identify[6], however, in our paper, the term 'race' –as in [8]– refers to a person's physical appearance rather than sociological and cultural concepts like ethnicity. For this end, we manually built a database from frontal portraits from the web. The images were subjectively collected and categorized in five very different 'races'. The collected races and the number of images per class are the following: 'Asian' (80), 'Black' (89), 'Hispanic' (85), 'Indian' (84) and 'White' (90). We call this database WebRace. The faces were detected automatically using Computer Vision Toolbox of Matlab[3]. In this case, we used $[n = 79|N = 60]$ and $\{Q = 90, R = 90, m = 700, \alpha = 3, w = 48, N_v = 500\}$. The results of our method compared with the baseline methods are summarized in Tab. 3.

3.4 Disguise

In this experiment, the idea was to distinguish faces with certain kind of occlusion. For this purpose, the database AR [21] was used. The images of this database were taken from 100 subjects (50 women and 50 men) with different facial expressions, illumination conditions, and occlusions with sun glasses and scarf (we used the cropped version). The number of images per subject is 26. We divided the database into three groups: images with scarf (600), images with sunglasses (600) and the rest (1400). In this case, we used $[n = 19|N = 60]$ and $\{Q = 80, R = 50, m = 400, \alpha = 2, w = 16, N_v = 200\}$. The results of our method compared with the baseline methods are summarized in Tab. 4.

3.5 Implementation Details

In the implementation of ASR+, we used open source libraries like VLFeat [32] for k-means and SPAMS for sparse representation[7]. Additional to the parameters

this case, the accuracy is 97.1%). Additionally, in [2], only 304 images (152 males and 152 females) were used for training and 107 images (60 males and 47 females) for testing (in this case, the reported accuracy is 99.1%).

[5] There is another experiment on GROUPS database reported in [5], in which all 28,231 images were used (in this case, the reported accuracy is 76.0%). Since the evaluation protocol is very different, it is not included in Tab. 2.

[6] See for example the educational game 'Guess my race' which aims to show bias tendencies by presenting that race is the result of complex cultural and historical constructions (http://www.gamesforchange.org/play/guess-my-race/).

[7] SPArse Modeling Software available on http://spams-devel.gforge.inria.fr

$\{Q, R, m, \alpha, w, N_v\}$ given in each experiment, the other parameters were (for all experiments): Number of testing patches $m^t = 800$. Threshold for minimal distance between the test patch and child cluster: $\theta = 0.05$. Threshold for SCI $\tau = 0.1$. Number of selected patches $s = 300$. Additionally, the number of words ('atoms') selected from the dictionary in (9) is 20 k'/k, where k' is the number of selected classes for the adaptive sparse representation, and k is the number of classes in the gallery. The time computing depends on the number of classes and the size of the dictionary, however, in order to present a reference, the testing results for the recognition of race were obtained after 0.8s per subject on a Mac Mini Server OS X 10.9.3, processor 2.6 GHz Intel Core i7 with 4 cores and memory of 16GB RAM 1600 MHz DDR3. The remaining algorithms were implemented in MATLAB. The code of the MATLAB implementation is available on our webpage[8].

4 Conclusions

In this paper, we have presented ASR+, a new general algorithm that is able to recognize facial attributes automatically in cases with less constrained conditions, including some variability in ambient lighting, pose, expression, size of the face and distance from the camera. The main contribution of our paper is that the same algorithm can be used in all recognition tasks obtaining a performance at least comparable with that achieved by state-or-art techniques. The robustness of our algorithm is due to three reasons: *i)* the dictionaries learned for each class in the learning stage corresponded to a rich collection of representations of relevant parts which were selected and clustered; *ii)* the testing stage was based on 'adaptive' sparse representations of several patches using the dictionaries estimated in the previous stage which provided the best match with the patches, and *iii)* a visual vocabulary and a stop list used to reject non-discriminative patches in both learning and testing stage.

It is worth mentioning that our extensive empirical evaluation has been performed in two directions: *i)* Other representative methods from the literature have been re-implemented and compared against using our methodology; and *ii)* our algorithm has been evaluated using the methodology of other papers to get a result that can be compared to their published result(s) on the selected datasets. In both scenarios, ASR+ can deal with the unconstrained conditions extremely well, achieving a high recognition performance in many complex conditions and obtaining similar or better performance.

We believe that ASR+ can be used to solve other kinds of recognition problems (*e.g.*, recognition of faces with glasses, mustaches or beards and estimation of age). Preliminary results have shown that ASR+ can be used to recognize specific individuals as well. The proposed model is very flexible and obviously it can be used with other descriptors.

[8] See http://dmery.ing.puc.cl/index.php/material/.

Acknowledgments

This work was supported in part by Fondecyt grant 1130934 from CONICYT–Chile and in part by Seed Grant Program of The College of Engineering at the Pontificia Universidad Catolica de Chile and the College of Engineering at the University of Notre Dame.

References

1. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence 28(12), 2037–2041 (2006)
2. Alexandre, L.A.: Gender recognition: A multiscale decision fusion approach. Pattern Recognition Letters 31(11), 1422–1427 (2010)
3. Baluja, S., Rowley, H.A.: Boosting sex identification performance. International Journal of Computer Vision 71(1), 111–119 (2007)
4. Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Recognizing facial expression: machine learning and application to spontaneous behavior. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005) (2005)
5. Bekios-Calfa, J., Buenaposada, J.M., Baumela, L.: Robust gender recognition by exploiting facial attributes dependencies. Pattern Recognition Letters 36, 228–234 (2014)
6. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008) (2008)
7. Chen, Y., Do, T.T., Tran, T.D.: Robust face recognition using locally adaptive sparse representation. In: IEEE International Conference on Image Processing (ICIP 2010). pp. 1657–1660 (2010)
8. Fu, S., He, H., Hou, Z.: Learning race from face: A survey. IEEE Trans. on Pattern Analysis and Machine Intelligence (2014), (In Press) DOI: 10.1109/TPAMI.2014.2321570
9. Fu, Y., Guo, G., Huang, T.S.: Age synthesis and estimation via faces: A survey. IEEE Trans. on Pattern Analysis and Machine Intelligence 32(11), 1955–1976 (2010)
10. Gallagher, A.C., Chen, T.: Understanding images of groups of people. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009). pp. 256–263 (2009)
11. Guo, G., Mu, G.: A study of large-scale ethnicity estimation with gender and age variations. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2010). pp. 79–86 (2010)
12. Guo, G., Mu, G.: Joint estimation of age, gender and ethnicity: CCA vs. PLS. In: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG 2013). pp. 1–6. IEEE (2013)
13. Kyperountas, M., Tefas, A., Pitas, I.: Salient feature and reliable classifier selection for facial expression classification. Pattern Recognition 43(3), 972–986 (2010)
14. Liang, D., Yang, J., Zheng, Z., Chang, Y.: A facial expression recognition system based on supervised locally linear embedding. Pattern Recognition Letters 26(15), 2374–2389 (2005)

15. Liu, P., Han, S., Men, Z., Tong, Y.: Facial expression recognition via a boosted deep belief network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014) (2014)
16. Lu, L., Shi, P.: Fusion of multiple facial regions for expression-invariant gender classification. IEICE Electronics Express 6(10), 587–593 (2009)
17. Lu, X., Jain, A.K.: Ethnicity identification from face images. In: Proceedings of SPIE Defense and Security Symposium. pp. 114–123 (2004)
18. Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: IEEE workshop on CVPR for Human Communicative Behavior Analysis (2010)
19. Lyons, M.J., Budynek, J., Akamatsu, S.: Automatic classification of single facial images. IEEE Trans. on Pattern Analysis and Machine Intelligence 21(12), 1357–1362 (1999)
20. Lyons, M.J., Budynek, J., Akamatsu, S.: Automatic classification of single facial images. IEEE Trans. on Pattern Analysis and Machine Intelligence 21(12), 1357–1362 (1999)
21. Martinez, A., Benavente, R.: The AR face database (June 1998), cVC Tech. Rep, No. 24
22. Moghaddam, B., Yang, M.H.: Learning gender with support faces. IEEE Trans. on Pattern Analysis and Machine Intelligence 24(5), 707–711 (2002)
23. Moon, H., Sharma, R., Jung, N.: Method and system for robust human ethnicity recognition using image feature-based probabilistic graphical models (2013), uS Patent 8,379,937
24. Ng, C.B., Tay, Y.H., Goi, B.M.: Recognizing human gender in computer vision: a survey. In: Proceedings of 12th Pacific Rim International Conference on Artificial Intelligence. pp. 335–346. Springer (2012)
25. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The feret evaluation methodology for face-recognition algorithms. IEEE Trans. on Pattern Analysis and Machine Intelligence 22(10), 1090–1104 (2000)
26. Reid, D.A., Samangooei, S., Chen, C., Nixon, M.S., Ross, A.: Soft Biometrics for Surveillance: An Overview. In: Handbook of Statistics, vol. 31, pp. 1–27. Elsevier (2013)
27. Salah, S.H., Du, H., Al-Jawad, N.: Fusing local binary patterns with wavelet features for ethnicity identification. In: Proceedings of IEEE International Conference on Signal and Image Processing (ICSIP 2013). pp. 330–336 (2013)
28. Samangooei, S., Guo, B., Nixon, M.S.: The use of semantic human description as a soft biometric. In: IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS 2008). pp. 1–7 (2008)
29. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing 27(6), 803–816 (2009)
30. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: International Conference on Computer Vision (ICCV 2003). pp. 1470–1477 (2003)
31. Tosic, I., Frossard, P.: Dictionary learning. Signal Processing Magazine, IEEE 28(2), 27–38 (2011)
32. Vedaldi, A., Fulkerson, B.: VLfeat: an open and portable library of computer vision algorithms. In: MM '10: Proceedings of the international conference on Multimedia. pp. 1469–1472. New York (Oct 2010)

33. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. on Pattern Analysis and Machine Intelligence 31(2), 210–227 (2009)
34. Xu, Y., Zhang, D., Yang, J., Yang, J.Y.: A Two-Phase Test Sample Sparse Representation Method for Use With Face Recognition. IEEE Trans. on Circuits and Systems for Video Technology 21(9), 1255–1262 (2011)
35. Yang, Z., Li, M., Ai, H.: An experimental study on automatic face gender classification. In: 18th International Conference on Pattern Recognition (ICPR 2006). vol. 3, pp. 1099–1102 (2006)
36. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans. on Pattern Analysis and Machine Intelligence 31(1), 39–58 (2009)
37. Zhong, L., Liu, Q., Yang, P., Liu, B., Huang, J., Metaxas, D.N.: Learning active facial patches for expression analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012) (2012)