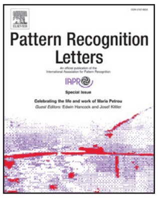
Pattern Recognition Letters 000 (2015) 1–10

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec


Pattern Recognition Letters

Automatic facial attribute analysis via adaptive sparse representation of random patches[☆]

Domingo Mery[a,*], Kevin Bowyer[b]

[a] Departamento de Ciencia de la Computación, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna Santiago 4860 (143), Chile
[b] Department of Computer Science and Engineering, University of Notre Dame, 384 Fitzpatrick, Notre Dame, IN 46556, USA

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Soft biometrics
Expression recognition
Gender recognition
Race recognition
Sparse representations
Facial attribute analysis

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, such as 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+). The proposed method consists of two stages: 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. A stop list is used to remove very common words of the dictionaries. In the testing stage, random test patches of the query image are extracted, and for each non–stopped test patch a dictionary is built concatenating the 'best' representative dictionary of each class. Using this adapted dictionary, each non–stopped 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 eight face databases in order to recognize facial expression, gender, race, disguise and beard. Results show that ASR+ deals well with unconstrained conditions, outperforming various representative methods in the literature in many complex scenarios.

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 [33]). The relevance of this research field is twofold: first, the use of facial attributes, like soft biometric traits (*e.g.*, gender [39], race [15], age [16], etc.), can increase the performance of traditional biometric systems [44] and help recognition based on human descriptions [46]. Second, other facial attributes, like facial expressions, can be used in human–computer interfaces, image retrieval, talking heads and human emotion analysis [62].

Usually, each single facial attribute has been recognized by a specific algorithm. Some examples are the following:

Gender is identified using a SVM classifier with Gaussian RBF kernel [37], a Real AdaBoost classifier with texture features [60], an AdaBoost classifier with a low resolution image [3], SVM classifier of PCA representations [30] and SVM with LBP features using feature selection based on mutual information and feature fusion [50].

Facial expressions are classified using a new feature called 'supervised locally linear embedding' [28], a decomposition into multiple two-class classification problems with 'salient feature vectors' [25], local binary patterns [47], a boosted deep belief network [29], active facial patches [66], and Gabor features [5].

Race is recognized using biologically inspired features [19], an ensemble framework with LDA [31], a probabilistic graphical model [38] and local binary patterns with wavelets features [45].

There are few approaches to estimate age, gender and race together (see for example [20,21]), however, to the best knowledge of the authors, there has been no reported approach that has been tested in recognition of facial attributes in general.

We believe that algorithms based on sparse representations can be used for this general task because in many computer vision applications, under the assumption that natural images can be represented using sparse decomposition, state-of-the-art results have been significantly improved [51]. Thus, it is possible to cast the problem of

[☆] This paper has been recommended for acceptance by Prof. G. Sanniti di Baja.
[*] Corresponding author. Tel.: +56 2 354 5808; fax: +56 2 354 4444.
E-mail address: dmery@ing.puc.cl (D. Mery).

http://dx.doi.org/10.1016/j.patrec.2015.05.005
0167-8655/© 2015 Elsevier B.V. All rights reserved.

recognition of facial attributes into a supervised recognition form with samples (face images) and class levels (*e.g.*, female and male for gender recognition) using learned features in an unsupervised way.

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. Usually, the query image is assigned to the class with the minimal reconstruction error. A very good example is the Sparse Representation Classification (SRC) [56] that has been widely used in face recognition where the dictionary corresponds to the original pixel intensity values of the training face images. Several variations of this approach were recently proposed. To cite a few: In [53], registration and illumination are simultaneously considered in the sparse representation. In [11], an intra-class variant dictionary is constructed to represent the possible variation between gallery and query images. In [54], sparsity and correlation are jointly considered. In [22] and [55], structured sparsity is proposed for dealing with occlusion and illumination. In [12], the dictionary is assembled by the class centroids and sample-to-centroid differences. In [9], SRC is extended by incorporating the low-rank structure of data representation. In [23], a discriminative dictionary is learned using label information. In [42], a linear extension of graph embedding is used to optimize the learning of the dictionary. In [43], a discriminative and generative dictionary is learned based on the principle of information maximization. In [48], a sparse discriminative analysis is proposed using the $\ell_{1,2}$-norm. In [57], a sparse representation in two phases is proposed. In [10], 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 example sunglasses when recognizing gender). 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. Moreover, it would be possible to 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, disguise and beard, obtaining a performance at least comparable with that achieved by state-of-the-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.

A preliminary version of this article was presented in [36]. In this extended version the contributions are: (i) new experiments on AR, UND and FRGC 2.0 databases are included. (ii) The proposed method is evaluated on the recognition of another facial attribute (beard recognition). (iii) The explanation of the proposed method is improved. (iv) In order to compare our method with other methods fairly, we evaluated the accuracy of our proposed method using cross-validation when other methods used cross-validation as well. (v) A method for parameter tuning is proposed. (vi) We discuss the results in greater detail.

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 have 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 [49].

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 non-stopped 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 [56]. Afterwards, the patches are selected according to a discriminative criterion. Finally, the query image is classified by voting for the

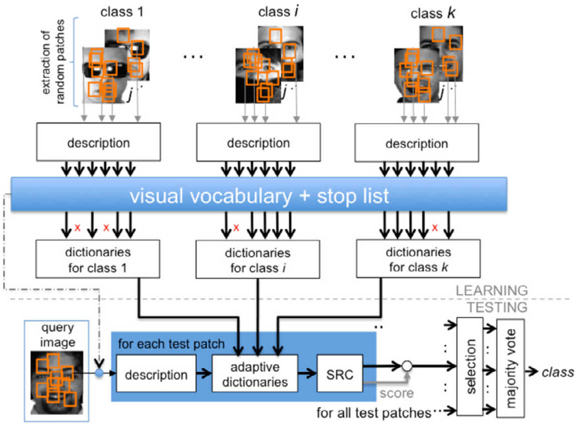


Fig. 1. Overview of the proposed method. The figure illustrates the recognition of disguise. The shown classes are three: sunglasses, scarf and no-disguise. There are two stages: learning and testing. The stop list is used to filter out patches that are not discriminating for these classes. The stopped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

selected patches. Both stages, learning and testing, will be explained in this section in further detail.

2.1. Learning

Learning consists of three main steps: description, stop list and dictionaries (see Fig. 1).

2.1.1. Description

In the training stage, a set of n face images of k classes is available, where $\mathbf{I}^i_j$ denotes image j of class i (for $i = 1\dots k$ and $j = 1\dots n$). In each image $\mathbf{I}^i_j$, 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 relative weighting factor between description and location[1]. Using (1), all extracted patches are described as $\mathbf{y}^i_{jp} = f(\mathcal{P}^i_{jp}) = [\mathbf{z}^i_{jp};\ \alpha x^i_{jp};\ \alpha y^i_{jp}]$, for $p = 1\dots m$.

2.1.2. Stop list

In order to eliminate non-discriminative patches, a *stop list* is computed from a *visual vocabulary*. The visual vocabulary is built using all descriptors $\mathbf{Z} = \{\mathbf{z}^i_{jp}\} \in \mathcal{R}^{d\times knm}$, 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 containing N_v visual words is obtained. In order to construct the stop list, the *term frequency* 't_f' is computed: $t_f(d, v)$ 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* 'd_f' is computed: $d_f(v) = \sum_d \{t_f(d, v) > 0\}$, *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 d_f values: on one hand, visual words with highest d_f values are not discriminative because they occur in almost all images. On the other hand, visual words with smallest d_f are so unusual that they correspond in most of the cases to noise. Usually, the top 5% and bottom 10% are stopped [49]. Those patches of $\mathbf{Z}$ that belong to the stopped clusters are not considered in the following steps of our algorithm.

2.1.3. Dictionaries

Now, for class $i = 1\dots k$ an array with the description of all (non stopped) patches $\mathbf{y}^i_{jp}$ 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

$$\mathbf{c}^i_q = k\text{-means}(\mathbf{Y}^i, Q) \tag{2}$$

for $q = 1\dots Q$, where $\mathbf{c}^i_q \in \mathcal{R}^{(d+2)}$ is the centroid of parent cluster q of class i. We define $\mathbf{Y}^i_q$ as the array with all samples $\mathbf{y}^i_{jp}$ that belong to the parent cluster with centroid $\mathbf{c}^i_q$. In order to select a reduced number of samples, each parent cluster is clustered again in R *child* clusters[2]

$$\mathbf{c}^i_{qr} = k\text{-means}(\mathbf{Y}^i_q, R) \tag{3}$$

for $r = 1\dots R$, where $\mathbf{c}^i_{qr} \in \mathcal{R}^{(d+2)}$ 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

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

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 $\{\mathbf{c}^i_{qr}\}$ 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. Testing stage consists of four main steps: description, adaptive dictionaries, sparse representation classification and selection (see Fig. 1).

2.2.1. Description

From the test image, m^t random patches are extracted. From them, s patches are selected (the selection criterion of a test patch will be explained later in Section 2.2.4). A selected patch $\mathcal{P}^t_p$ of size $w \times w$ pixels is described using (1) as $\mathbf{y}^t_p = f(\mathcal{P}^t_p) = [\mathbf{z}^t_p;\ \alpha x^t_p;\ \alpha y^t_p]$ (for $p = 1\dots s$).

2.2.2. Adaptive dictionaries

For each selected test patch with description $\mathbf{y} = \mathbf{y}^t_p$, a distance to each parent cluster q of each class i of the gallery is measured

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

We tested with several distance metrics. The best performance, however, was obtained by:

$$h^i(\mathbf{y}, q) = \min_r ||\mathbf{y} - \mathbf{c}^i_{qr}|| \quad \text{for} \quad r = 1\dots R, \tag{6}$$

which is the smallest Euclidean distance to centroids of child clusters of parent cluster q as illustrated in Fig. 2. For $\mathbf{y}$ and $\mathbf{c}^i_{qr}$ normalized to unit ℓ_2 norm, the following distance can be used based on (6):

$$h^i(\mathbf{y}, q) = \min_r (1 - <\mathbf{y}, \mathbf{c}^i_{qr}>) \quad \text{for} \quad r = 1\dots R, \tag{7}$$

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

$$\hat{q}^i = \underset{q}{\text{argmin}}\ h^i(\mathbf{y}, q), \tag{8}$$

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

[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).

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

Please cite this article as: D. Mery, K. Bowyer, Automatic facial attribute analysis via adaptive sparse representation of random patches, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.05.005

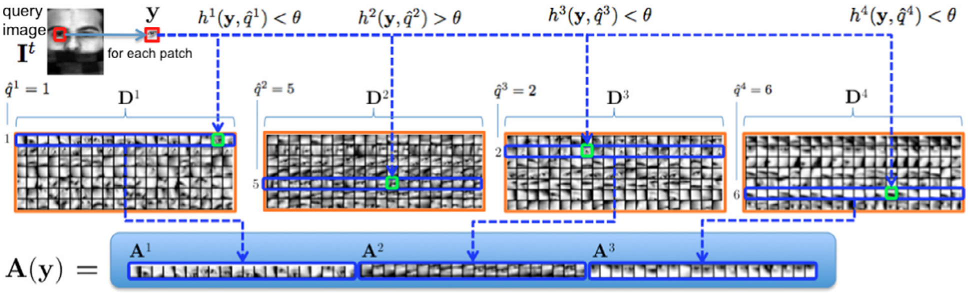


Fig. 2. Adaptive dictionary **A** of patch **y**. In this example there are $k = 4$ classes in the gallery. For this patch only $k' = 3$ classes are selected. Dictionary **A** is built from those classes by selecting all child clusters (of a parent cluster—see blue rectangles) which have a child with 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 **y**, *i.e.*, $h^2(\mathbf{y}, \hat{q}^2) > \theta$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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 **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 $\mathbf{A}^{i'} := \bar{\mathbf{A}}^{v_{i'}}_{u_{i'}}$.

Therefore, a dictionary for patch **y** is built using the best representative patches as follows (see Fig. 2)

$$\mathbf{A}(\mathbf{y}) = [\mathbf{A}^1 \dots \mathbf{A}^{i'} \dots \mathbf{A}^{k'}] \in \mathcal{R}^{(d+2)\times Rk'} \tag{9}$$

2.2.3. Sparse representation classification (SRC)

With this adaptive dictionary **A**, built for patch **y**, we can use SRC methodology [56]. That is, we look for a sparse representation of **y** using the ℓ_1-minimization approach:

$$\hat{\mathbf{x}} = \text{argmin}||\mathbf{x}||_1 \quad \text{subject to} \quad \mathbf{Ax} = \mathbf{y}. \tag{10}$$

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

$$r_{i'}(\mathbf{y}) = ||\mathbf{y} - \mathbf{A}\delta_{i'}(\hat{\mathbf{x}})||, \tag{11}$$

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 **y** will be the class that has the minimal residual, that is it will be

$$\hat{i}(\mathbf{y}) = v(\hat{i}'), \tag{12}$$

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 $\mathbf{y}^t_p$, for $p = 1 \dots s$[3]

$$\text{identity}(\mathbf{I}^t) = \text{mode}(\hat{i}(\mathbf{y}^t_1), \dots \hat{i}(\mathbf{y}^t_p), \dots \hat{i}(\mathbf{y}^t_s)). \tag{13}$$

2.2.4. Selection and majority vote

The selection of s patches of query image $\mathbf{I}^t$ is as follows:

(1) From query image $\mathbf{I}^t$, m^t patches are randomly extracted and described using (1): $\mathbf{y}^t_j$, for $j = 1 \dots m^t$, with $m^t \geq s$.
(2) Those patches $\mathbf{y}^t_j$ that belong to the stopped clusters of our visual vocabulary are not considered.
(3) Each remaining patch $\mathbf{y}^t_j$ is represented by $\hat{\mathbf{x}}^t_j$ using (10).
(4) The *sparsity concentration index* (SCI) of each patch is computed in order to evaluate how spread are its sparse coefficients [56]. SCI is defined by

$$S_j := \text{SCI}(\mathbf{y}^t_j) = \frac{k\ \max_{i'}(||\delta_{i'}(\hat{\mathbf{x}}^t_j)||_1)/||\hat{\mathbf{x}}^t_j||_1 - 1}{k - 1}, \tag{14}$$

where SCI ∈ [0, 1]. $\text{SCI}(\mathbf{y}^t_j) = 1$ means that all sparse coefficients of $\hat{\mathbf{x}}^t_j$ correspond to the same class. On the other hand, $\text{SCI}(\mathbf{y}^t_j) = 0$ means that the sparse coefficients are evenly spread between the classes [56]. 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.
(5) Array $\{S\}^m_{j=1}$ is sorted into descending 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 (13) will be taken with the first s' patches.

3. Experimental results

In this section, we report the experiments that we conducted in several recognition tasks, we discuss the obtained results, and we give some details about parameter tuning and implementation.

3.1. Experiments

ASR+ was evaluated in the recognition of several facial attributes: facial expressions (Section 3.1.1), gender (Section 3.1.2), and other attributes such as race, disguise and beard (Section 3.1.3). Experiments were carried out on eight 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 same 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 [7] 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 [56] 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 [57], and (v) LAD [10] 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 comparing our proposed approach with 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

[3] If there is a tie, we randomly chose one among the winners.

Please cite this article as: D. Mery, K. Bowyer, Automatic facial attribute analysis via adaptive sparse representation of random patches, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.05.005

Table 1
Recognition of seven expressions in JAFFE.

Method	Reference	η [%]
SLLE	[27]	92.7[1]
LP-LBP	[13]	93.8[1]
Boosted-LBP	[47]	81.0[1]
Hybrid filter	[26]	96.7[1]
L-SVM	[18]	92.4[1]
PDM-Gabor	[24]	90.2[1]
Ensamble	[61]	96.2[1]
ASR+	(Ours)	96.7

[1] Result from cited paper.

Table 2
Recognition of six expressions in CK+[1].

Method	Reference	η [%]
CSPL	[66]	89.9[2]
CPL	[66]	88.4[2]
MCPL	[65]	90.3[2]
MCSPL	[65]	91.5[2]
Boosted-LBP	[47]	92.6[2]
Ensamble	[61]	99.4[2]
SVM-RBF	[4]	86.9[2]
ASR+	(Ours)	99.3

[1] Only two of six expressions are shown.
[2] Result from cited paper.

Table 3
Happiness recognition in SmileFlick.

Method	Reference	η [%]
NBNN	[7]	73.1
LBP	[1]	87.5
SRC	[56]	96.8
TPTSR	[57]	91.2
LAD	[10]	97.5
ASR+	(Ours)	97.5

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 these 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 each experiment, 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, α, w, N_v)'[4].

3.1.1. *Facial expression*

Recognition of facial expression was evaluated on following databases: JAFFE, CK+ and SmileFlick.

- **JAFFE**: The database [33] contains 7 expressions ('neutral' and six basic emotions: 'fear', 'happiness', 'sadness', 'surprise', 'anger' and 'disgust' as shown in Fig. 3) 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. In our experiments, we used a 10-fold cross validation protocol for recognition of the seven expressions with ($Q = 200, R = 80, m = 320, \alpha = 2, w = 36, N_v = 200$). For fair comparison, Table 1 shows only results using the same experimental protocol. It is worth mentioning that ASR+ is able to select automatically those patches that are relevant for the recognition task as illustrated in Fig. 3. In this figure, the center of mass of each patch that belongs to the majority voted class is represented by a small 2D Gaussian mask (the size is the width of the patch). The visualization shows the addition of these Gaussian masks using a heat color map. Thus, in red we have the regions of the face with more selected patches, whereas in blue the regions with no selected patch. It is clear that the red regions correspond to those areas of the face that are more distinctive for the classified expression. For example, we can observe how relevant are the patches of the mouth region by recognizing the expression 'happiness'.
- **CK+**: The database [32] consists of eight expressions ('contempt' was added to the 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 six basic emotions. In our experiments, we used a 10-fold cross validation protocol for recognition of the six expressions with ($Q = 100, R = 80, m = 100, \alpha = 0.25, w = 20, N_v = 400$). For fair comparison, Table 2 shows only results using the same experimental protocol.
- **SmileFlick**: 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[5]. 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 Table 3. Face images are not shown in this paper due to copyright restrictions.

3.1.2. *Gender*

Gender recognition was evaluated on following databases: AR, FERET, UND and GROUPS.

- **AR**: The images of database 'AR' [34] 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). For gender recognition, we follow the protocol of [59] that uses the non-occluded subset (14 images per subject). In this experiment, the first 25 males and 25 females were used for training and the last 25 males and 25 females were used for testing. In our experiments, ($Q = 200, R = 220, m = 300, \alpha = 1, w = 30, N_v = 300$). See results in Table 4.
- **FERET**: The database [41] contains more than 3500 face images from women and men (with different races such as African, Asian and Caucasian) involving different expressions and illumination

[4] Additional to the parameters (Q, R, m, α, w, N_v) given in each experiment, we set the other parameters (for all experiments) as follows: 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 (10) 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.

[5] See `vision.CascadeObjectDetector` object on http://www.mathworks.com/products/computer-vision/.

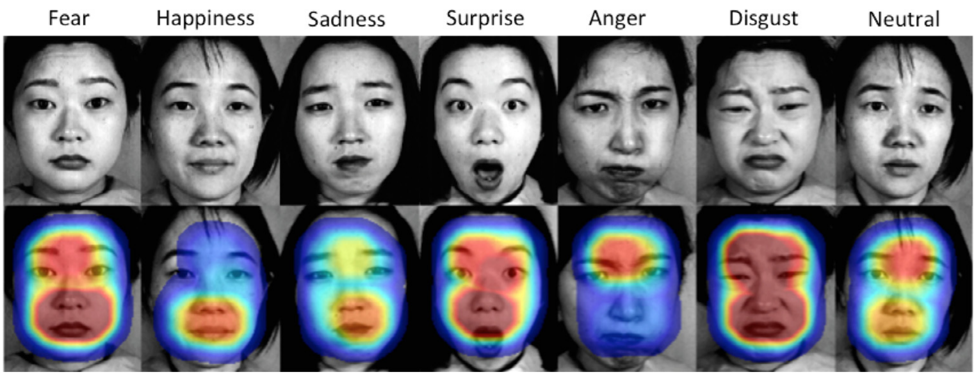


Fig. 3. In the recognition of expressions, there are parts of the face that are more relevant than other parts. This representation shows a 'heat map' of the location of the patches that where selected by ASR+ for each query image (one for each expression). They correspond to the centers of mass of the patches of the majority vote class (see Section 2.2.4).

Table 4
Gender recognition in AR.

Method	Reference	η [%]
LDL	[58]	95.3[1]
LC-KSVD	[23]	86.8[2]
FDDL	[59]	95.4[1]
DKSVD	[64]	86.1[2]
LGBP+SRC	[8]	97.7[1]
L$_2$R	[8]	94.0[1]
L$_2$R	[63]	94.9[1]
ASR+	(Ours)	97.6

[1] Result from cited paper.
[2] Result cited in [58].

Table 5
Gender recognition in FERET.

Method	Reference	Images	Eval	Mixed	η [%]
SVM-RBF	[37]	1755	CV	No	96.6[1]
Real AdaBoost	[60]	3529	CV	Yes	93.8[1]
AdaBoost	[3]	2409	HO	No	94.4[1]
AdaBoost	[3]	3529	CV	Yes	97.1[1]
2DPCA-SVM	[30]	800	CV	No	94.8[1]
ASR+	(Ours)	1040	CV	No	94.1

[1] Result from cited paper.

conditions. In the literature, there are many reported experimental protocols including different number of images, number of females and males, 'mixed' or 'unmixed' datasets[6], and evaluation methodology (5-fold cross validation (CV) and 80–20% hold out (HO)). In order to compare the performance of our approach, Table 5 shows the results obtained by other state-of-the-art methods, however, the evaluation protocols are not exactly the same. In [37], 1,044 males and 711 females were tested and the accuracy was estimated using a 5-fold cross validation strategy. In [60], 3529 images were used and the accuracy was estimated using a 5-fold cross validation strategy. In [3], 2409 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 the same work there is an experiment where a subject may appear in both train and test set using 5-fold cross validation with a significant increase in the accuracy). In [30], 400 males and 400 females were used and the accuracy was estimated using a 5-fold cross validation strategy[7]. We tested ASR+ on 1040 unmixed images (600 males and 440 females) using 5-fold cross validation with ($Q = 200, R = 80, m = 320, \alpha = 3.1, w = 36, N_v = 200$). The results are not conclusive because the experimental protocols (and the selected images) are not exactly the same, however, we can observe that the accuracies are very similar.

Table 6
Gender recognition in UND.

Method	Reference	η [%]	Method	Reference	η [%]
L1: Intensity	[2]	85.5[1]	L2: Shape	[50]	84.6[1]
L2: Shape	[2]	84.6[1]	L3: Texture	[50]	92.1[1]
L3: Texture	[2]	86.8[1]	L4: 20 × 20	[50]	84.2[1]
L4: 20 × 20	[2]	73.6[1]	L5: 36 × 36	[50]	88.2[1]
L5: 36 × 36	[2]	80.6[1]	L6: 128 × 128	[50]	92.1[1]
L6: 128 × 128	[2]	79.3[1]	L7: L1 …L6	[50]	92.5[1]
L7: L1 …L6	[2]	91.2[1]	best fea	[50]	94.0[1]
L1: Intensity	[50]	87.8[1]	ASR+	(ours)	92.5

[1] Result cited in [50].

- **UND** (Collection B): The database [14] contains a face image of 487 subjects (301 males and 186 females). For gender recognition, we follow the evaluation protocol suggested by [50] that uses a stratified 5-fold cross validation. That means in each fold we have 241 males and 149 females for training and 60 males and 37 females for testing. Results are summarized in Table 6. The reported accuracies for the method developed by [2] include several types of features: L1, L2 and L3 include intensity, shape and texture features respectively in three different scales (images of 20 × 20, 36 × 36 and 128 × 128 pixels); L4, L5 and L6 include a concatenation of intensity, shape and texture features of only one scale (images of 20 × 20, 36 × 36 and 128 × 128 pixels respectively); and L7 is a concatenation of all mentioned features. On the other hand, the reported accuracies for

[6] For the 'unmixed' protocol, images of a particular subject appear only in the training set or in the testing set, whereas for 'mixed' protocol a subject may appear in both training and testing sets.

[7] There are other experiments on FERET database reported in the literature that are not included in Table 5 because the testing protocols are significantly different. In [2,50], 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 95.3% and 97.8% respectively).

Table 7
Gender recognition in GROUPS.

Method	Reference	η [%]
NBNN	[7]	84.2
LBP	[1]	83.3
SRC	[56]	86.9
TPTSR	[57]	85.8
LAD	[10]	87.5
ASR+	(Ours)	93.3

Table 8
Race recognition in FRGC 2.0[1].

Method	Reference	η [%]
NBNN	[7]	61.3
LBP	[1]	63.0
SRC	[56]	62.0
TPTSR	[57]	65.3
LAD	[10]	85.7
ASR+	(Ours)	87.1

[1] Only two of five races are shown.

Table 9
Disguise recognition in AR[1].

Method	Reference	η [%]
NBNN	[7]	97.8
LBP	[1]	96.1
SRC	[56]	98.3
TPTSR	[57]	97.8
LAD	[10]	96.7
ASR+	(Ours)	97.8

[1] Only two of three classes are shown.

the method developed by [50] include the same L1 ...L7 features, however, the features were selected using different feature selection algorithms before the classification was performed (the accuracies shown in Table 6 are the best accuracies obtained by the authors in each case). In addition, [50] reported a method called 'best fea', where the best features obtained from the different feature selection algorithms for L1 ...L6 were used. In our experiments, we tested ASR+ on images of 128 × 128 using intensity features only. The obtained accuracy of ASR+ is higher than the accuracy of the other methods when comparing at the same scale or with the same type of features. Moreover, a better accuracy was only achieved when a sophisticated approach including all features (intensity, shape and texture) in all scales (20 × 20, 36 × 36 and 128 × 128) and the best results of four different feature selection approaches was used. In our experiments, $(Q = 170, R = 50, m = 300, \alpha = 0.6, w = 30, N_v = 300)$.

• **GROUPS**: The database [17] consists of 28,231 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 Table 7. Our method is compared with the basis methods[8].

3.1.3. Other attributes

In this Section, we report the results obtained in race recognition, and disguise and beard detection.

• **Race**: For human beings it is very difficult to distinguish a race, because it depends on how people self identify[9], however, in our paper, the term 'race' –as in [15]– 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 FRGC 2.0 [40] and 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). The faces were detected automatically using Computer Vision Toolbox of Matlab[5]. 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 Table 8.

• **Disguise**: In this experiment, the idea was to distinguish faces with certain kind of occlusion. For this purpose, the database AR [34] 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 classes: 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 Table 9.

• **Beard**: For beard recognition we use FRGC 2.0 [40] with only 30 men with beard and 30 men without beard. In this experiment, we used $(n = 30|N = 60)$ and $(Q = 90, R = 90, m = 700, \alpha = 3, w = 48, N_v = 500)$. In our experiments, we use the third bottom image. The results of our method compared with the baseline methods are summarized in Table 10. As we can see in the face images of this database, this experiment is challenging since some men have an extremely short beard (called *dirty shave*).

3.1.4. Discussion

Our method is a general algorithm that is able to recognize many facial attributes automatically in cases with less constrained conditions, including some variability in ambient lighting, pose, expression and size of the face. Ten experiments were carried out on eight databases in order to recognize facial expression, gender, race, disguise and beard.

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

[8] There is another experiment on GROUPS database reported in [6], 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 Table 7.

[9] 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/).

Please cite this article as: D. Mery, K. Bowyer, Automatic facial attribute analysis via adaptive sparse representation of random patches, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.05.005

Table 10
Beard recognition in FRGC 2.0.

Method	Reference	η [%]
NBNN	[7]	75.0
LBP	[1]	75.0
SRC	[56]	78.9
TPTSR	[57]	62.5
LAD	[10]	80.8
ASR+	(Ours)	90.4

both scenarios, ASR+ can deal with the unconstrained conditions extremely well.

Thus, the significance of our results is twofolds: first, ASR+ is a general recognition algorithm that can be used in different facial attribute analysis tuning six parameters only (see Section 3.2 for details). Second, results show that ASR+ deals well with unconstrained conditions in every experiment, achieving a high recognition performance in many complex conditions and obtaining similar or better performance in comparison with other representative methods in the literature.

3.2. Parameter tuning

Our proposed method ASR+ has six parameters, $\theta = (Q, R, m, \alpha, w, N_v)$[4]. Similar to other published methods, in Section 3.1 our parameters were manually tuned in order to maximize the accuracy.

Regarding the parameters used by our method, the number of extracted patches per class in the training stage is $m \times n$, where m is the number of patches per training image, and n is the number of training images per class. Consequently, the total number of clusters per class $Q \times R$ –where Q (2) and R (3) are the number of parent and child clusters respectively– must be less than $m \times n$. In our experiments, the ratio $m \times n$ to $Q \times R$ is between 1.1 and 10, and $Q \geq R$. Parameter α (1) in our experiments is between 0.25 and 4. By adapting this weighting parameter, an optimal compromise between appearance description and location of the patch is possible. For highly aligned faces, α may be high, however, for not aligned faces or for attributes that do not depend so much on the location, α may be low. Parameter w[1] means the size of the patch. It depends on the size of the face and the size of the area in which the attribute can be distinguishable. Finally, N_v (see Section 2.1.2) means the number of visual words of the dictionary that is used to construct the stop-list. In our experiments, N_v is a number between 100 and 500.

In this section, we propose a methodology to automatically tune parameter vector θ, that avoids possible overfitting. Given a dataset $\mathbf{X}$ with N face images for a specific recognition task, *e.g.*, gender recognition, we divide it into two disjoint subsets: tuning dataset $\mathbf{X}_1$ and validation dataset $\mathbf{X}_2$ with N_1 and N_2 samples respectively ($N = N_1 + N_2$). On first subset, we estimate the best parameter vector $\hat{\theta}$ that maximizes the accuracy of ASR+ (using cross-validation). Afterwards, we evaluate the accuracy of ASR+ on subset $\mathbf{X}_2$ using $\hat{\theta}$. Thus, the parameters of the method are estimated using $\mathbf{X}_1$ only and the final accuracy is evaluated using an independent dataset $\mathbf{X}_2$.

In order to test the proposed methodology, a dataset with a large number N is required. In this experiment, our dataset $\mathbf{X}$ is the FERET database used in Section 3.1.2 with $N = 1040$ unmixed subjects (600 males and 440 females). The tuning dataset ($\mathbf{X}_1$) consists of 75% of the images of $\mathbf{X}$ (the first 450 males and 330 females), and the validation dataset ($\mathbf{X}_2$) consists of the remaining 25% of the images of $\mathbf{X}$ (the last 150 males and 110 females). Thus, $N_1 = 780$, $N_2 = 260$ samples.

In order to maximize the accuracy on $\mathbf{X}_1$, we use exhaustive search for $Q = 30, 60, \ldots 240$; $R = 30, 60, \ldots 150$; $m = 100, 200, 300$; $\alpha = 0.25, 0.5, 1, 2, 3, 4$; $w = 28, 32, 36, 40$; $N_v = 0, 100, \ldots, 500$. Using 5-fold cross-validation, the highest accuracy on tuning dataset was 93.2% for $\hat{\theta} = (Q = 210, R = 90, m = 300, \alpha = 3, w = 36, N_v = 200)$[10]. For this set of parameters, the accuracy on the validation dataset was $\eta = 93.8\%$. Thus, the accuracy obtained in both tuning and validation datasets are very similar. Moreover, the reported accuracy in Section 3.1 was $\eta = 94.1\%$ (see Table 5), using 5-fold cross-validation on the whole dataset $\mathbf{X}$ with $N = 1040$ samples, with a similar set of parameters that were manually obtained by maximizing the accuracy.

It is worth mentioning that the best accuracy obtained in the tuning dataset without considering stop-list on $\mathbf{X}_1$, *i.e.*, $N_v = 0$, was $\eta_1 = 89.7\%$, that is 3.5% below of the accuracy considering a stop-list. This reduction of approximately 3–5% in the accuracy (when the stop-list was not considered), was observed in every experiment reported in this paper.

3.3. Implementation details

In the implementation of ASR+, we used open source libraries like VLFeat [52] for k-means and SPAMS for sparse representation[11]. 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 10.10.1, processor 2.6 GHz Intel Core i7 with 4 cores and memory of 16GB RAM 1600MHz DDR3. The remaining algorithms were implemented in MATLAB 2014b. The code of the MATLAB implementation is available on our webpage[12].

4. Conclusions

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.

We believe that ASR+ can be used to solve other kinds of recognition problems (*e.g.*, recognition of faces with glasses, mustaches or estimation of age). Preliminary results have shown that ASR+ can be used to recognize specific individuals as well [35]. The proposed model is very flexible and obviously it can be used with other descriptors.

In terms of future work, we will extend this approach to face recognition using videos and other object-recognition problems. We will accelerate computation by using faster proximity search algorithms.

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

[10] In these experiments the minimal accuracy was 80.4%, the mean was 88.5% and the standard deviation was 1.7%.

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

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

Please cite this article as: D. Mery, K. Bowyer, Automatic facial attribute analysis via adaptive sparse representation of random patches, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.05.005

Engineering at the Pontificia Universidad Catolica de Chile and the College of Engineering at the University of Notre Dame. The authors would also like to thank J. Tapia for providing the experimental protocol on UND.

References

[1] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: application to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 28 (12) (2006) 2037–2041.
[2] L.A. Alexandre, Gender recognition: a multiscale decision fusion approach, Pattern Recogn. Lett. 31 (11) (2010) 1422–1427.
[3] S. Baluja, H.A. Rowley, Boosting sex identification performance, Int. J. Comput. Vis. 71 (1) (2007) 111–119.
[4] M.S. Bartlett, G. Littlewort, I. Fasel, J.R. Movellan, Real time face detection and facial expression recognition: development and applications to human computer interaction, in: Computer Vision and Pattern Recognition Workshop 2003 CVPRW '03 IEEE, 2003, p. 53.
[5] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Recognizing facial expression: machine learning and application to spontaneous behavior, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005), 2005.
[6] J. Bekios-Calfa, J.M. Buenaposada, L. Baumela, Robust gender recognition by exploiting facial attributes dependencies, Pattern Recogn. Lett. 36 (2014) 228–234.
[7] O. Boiman, E. Shechtman, M. Irani, In defense of nearest-neighbor based image classification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008), 2008, pp. 1–8.
[8] C. Chen, A. Ross, May 2013, Local gradient Gabor pattern (LGGP) with applications in face recognition, cross-spectral matching, and soft biometrics. SPIE Defense, Security, and Sensing 8712, 87120R.
[9] J. Chen, Z. Yi, Sparse representation for face recognition by discriminative low-rank matrix recovery, J. Vis. Commun. Image R. 25 (5) (2014) 763–773.
[10] Y. Chen, T.T. Do, T.D. Tran, Robust face recognition using locally adaptive sparse representation, in: IEEE International Conference on Image Processing (ICIP 2010), 2010, pp. 1657–1660.
[11] W. Deng, J. Hu, J. Guo, Extended SRC: undersampled face recognition via intraclass variant dictionary, IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2012) 1864–1870.
[12] W. Deng, J. Hu, J. Guo, In defense of sparsity based face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013), 2013, pp. 399–406.
[13] X. Feng, M. Pietikainen, A. Hadid, Facial expression recognition based on local binary patterns, Pattern Recognit. Image Anal. 17 (4) (2007) 592–598.
[14] P.J. Flynn, K.W. Bowyer, P.J. Phillips, Assessment of time dependency in face recognition: an initial study, AVBPA'03: Proceedings of the 4th International Conference on Audio- and Video-based Biometric Person Authentication, Springer-Verlag, June 2003.
[15] S. Fu, H. He, Z. Hou, Learning race from face: a survey, IEEE Trans. Pattern Anal. Mach. Intell. (2014), doi:10.1109/TPAMI.2014.2321570.
[16] Y. Fu, G. Guo, T.S. Huang, Age synthesis and estimation via faces: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 1955–1976.
[17] A.C. Gallagher, T. Chen, Understanding images of groups of people, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), 2009, pp. 256–263.
[18] G. Guo, C.R. Dyer, Learning from examples in the small sample case: face expression recognition, IEEE Trans. Syst. Man Cybernet. Part B: Cybernet. 35 (3) (June 2005) 477–488.
[19] G. Guo, G. Mu, 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), 2010, pp. 79–86.
[20] G. Guo, G. Mu, Joint estimation of age, gender and ethnicity: CCA vs. PLS, in: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG 2013). IEEE, 2013, pp. 1–6.
[21] H. Han, C. Otto, X. Liu, A. Jain, Demographic estimation from face images: human vs. machine performance, Pattern Anal. Mach. Intell. 99 (1) (2014). IEEE Transactions on.
[22] K. Jia, T.-H. Chan, Y. Ma, Robust and practical face recognition via structured sparsity, European Conference on Computer Vision (ECCV 2012), Springer, 2012, pp. 331–344.
[23] Z. Jiang, Z. Lin, L.S. Davis, Label consistent k-svd: Learning a discriminative dictionary for recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (11) (2013) 2651–2664.
[24] A. Koutlas, D.I. Fotiadis, An automatic region based methodology for facial expression recognition, in: SMC 2008. IEEE International Conference on Systems, Man and Cybernetics, 2008, pp. 662–666.
[25] M. Kyperountas, A. Tefas, I. Pitas, Salient feature and reliable classifier selection for facial expression classification, Pattern Recognit. 43 (3) (2010) 972–986.
[26] P. Li, S.L. Phung, A. Bouzerdoum, F.H.C. Tivive, Feature selection for facial expression recognition, in: IEEE 2nd European Workshop on Visual Information Processing (EUVIP 2010). IEEE, pp. 35–40.
[27] D. Liang, J. Yang, Z. Zheng, Y. Chang, A facial expression recognition system based on supervised locally linear embedding, Pattern Recognit. Lett. 26 (15) (2005) 2374–2389.
[28] D. Liang, J. Yang, Z. Zheng, Y. Chang, A facial expression recognition system based on supervised locally linear embedding, Pattern Recognit. Lett. 26 (15) (2005b) 2374–2389.
[29] P. Liu, S. Han, Z. Men, Y. Tong, 2014, Facial expression recognition via a boosted deep belief network In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014).
[30] L. Lu, P. Shi, Fusion of multiple facial regions for expression-invariant gender classification, IEICE Electron. Express 6 (10) (2009) 587–593.
[31] X. Lu, A.K. Jain, Ethnicity identification from face images, in: Proceedings of SPIE Defense and Security Symposium, 2004, pp. 114–123.
[32] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, 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.
[33] M.J. Lyons, J. Budynek, S. Akamatsu, Automatic classification of single facial images, IEEE Trans. Pattern Anal. Mach. Intell. 21 (12) (1999b) 1357–1362.
[34] A. Martinez, R. Benavente, The AR face database, CVC Tech. Rep. 24 (June 1998) 8. http://www.cat.uab.cat/Public/Publications/1998/MaB1998/CVCReport24.pdf.
[35] D. Mery, K. Bowyer, Face recognition via adaptive sparse representations of random patches, in: IEEE Workshop on Information Forensics and Security (WIFS 2014), 2014a.
[36] D. Mery, K. Bowyer, Recognition of facial attributes using adaptive sparse representations of random patches, in: 1st International Workshop on SoftBiometrics, in conjunction with European Conference on Computer Vision (ECCV 2014), 2014b.
[37] B. Moghaddam, M.-H. Yang, Learning gender with support faces, IEEE Trans. Pattern Anal. Mach. Intell. 24 (5) (2002) 707–711.
[38] H. Moon, R. Sharma, N. Jung, Method and system for robust human ethnicity recognition using image feature-based probabilistic graphical models, 2013, US Patent 8,379,937.
[39] C.B. Ng, Y.H. Tay, B.-M. Goi, Recognizing human gender in computer vision: a survey, Proceedings of 12th Pacific Rim International Conference on Artificial Intelligence, Springer, 2012, pp. 335–346.
[40] P.J. Phillips, P.J. Flynn, W.T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W.J. Worek, Overview of the face recognition grand challenge, CVPR 1 (2005) 947–954.
[41] P.J. Phillips, H. Moon, S.A. Rizvi, P.J. Rauss, The Feret evaluation methodology for face-recognition algorithms, IEEE Trans. Pattern Anal Mach. Intell. 22 (10) (2000) 1090–1104.
[42] R. Ptucha, A. Savakis, LGE-KVSD: flexible dictionary learning for optimized sparse representation classification, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2013), IEEE Computer Society., 2013, pp. 854–861.
[43] Q. Qiu, V.M. Patel, R. Chellappa, Information-theoretic dictionary learning for image classification, in: IEEE Transactions on Pattern Analysis and Machine Intelligence (to be published), 2014.
[44] D.A. Reid, S. Samangooei, C. Chen, M.S. Nixon, A. Ross, Soft biometrics for surveillance: an overview, Handbook of Statistics, Elsevier, 31, 2013, pp. 1–27.
[45] S.H. Salah, H. Du, N. Al-Jawad, Fusing local binary patterns with wavelet features for ethnicity identification, in: Proceedings of IEEE International Conference on Signal and Image Processing (ICSIP 2013), 2013, pp. 330–336.
[46] S. Samangooei, B. Guo, M.S. Nixon, The use of semantic human description as a soft biometric, in: IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS 2008), 2008, pp. 1–7.
[47] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816.
[48] X. Shi, Y. Yang, Z. Guo, Z. Lai, Face recognition by sparse discriminant analysis via joint $l_{2,1}$-norm minimization, Pattern Recognit. (2014).
[49] J. Sivic, A. Zisserman, Video google: a text retrieval approach to object matching in videos, in: International Conference on Computer Vision (ICCV 2003), 2003, pp. 1470–1477.
[50] J.E. Tapia, C.A. Perez, Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of lbp, intensity, and shape, IEEE Trans. Inf. Foren. Sec. 8 (3) (2013) 488–499.
[51] I. Tosic, P. Frossard, Dictionary learning, IEEE Signal Proc. Mag. 28 (2) (2011) 27–38.
[52] A. Vedaldi, B. Fulkerson, Vlfeat: an open and portable library of computer vision algorithms, in: MM '10: Proceedings of the International Conference on Multimedia. New York, October 2010, pp. 1469–1472.
[53] A. Wagner, Z. Zhou, J. Wright, H. Mobahi, A. Ganesh, Y. Ma, Toward a practical face recognition system: robust alignment and illumination by sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2) (2012) 372–386.
[54] J. Wang, C. Lu, M. Wang, P. Li, S. Yan, X. Hu, Robust face recognition via adaptive sparse representation, IEEE Trans. Cybern. 99 (1) (2014).
[55] X. Wei, C.-T. Li, Y. Hu, Robust face recognition under varying illumination and occlusion considering structured sparsity, in: International Conference on Digital Image Computing Techniques and Applications (DICTA 2012), 2012, pp. 1–7.
[56] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[57] Y. Xu, D. Zhang, J. Yang, J.-Y. Yang, A two-phase test sample sparse representation method for use with face recognition, IEEE Trans. Circ. Syst. Video Technol. 21 (9) (2011) 1255–1262.
[58] M. Yang, D. Dai, L. Shen, L.V. Gool, Latent dictionary learning for sparse representation based classification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014), 2014, pp. 4138–4145.
[59] M. Yang, D. Zhang, X. Feng, Fisher discrimination dictionary learning for sparse representation, in: IEEE International Conference on Computer Vision (ICCV 2011), 2011, pp. 543–550.

[60] Z. Yang, M. Li, H. Ai, An experimental study on automatic face gender classification, in: 18th International Conference on Pattern Recognition (ICPR 2006), 3, 2006, pp. 1099–1102.

[61] T. Zavaschi, A.S. Britto Jr, l. Oliveira, Fusion of feature sets and classifiers for facial expression recognition, in: Expert Systems with Applications, 2013.

[62] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 39–58.

[63] L. Zhang, M. Yang, X. Feng, Sparse representation or collaborative representation: which helps face recognition? in: IEEE International Conference on Computer Vision (ICCV 2011), 2011, pp. 471–478.

[64] Q. Zhang, B. Li, Discriminative k-svd for dictionary learning in face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010). IEEE, 2010, pp. 2691–2698.

[65] L. Zhong, Q. Liu, P. Yang, J. Huang, D.N. Metaxas, Learning multiscale active facial patches for expression analysis, IEEE Trans. Cybernet. 99 (2014) 1.

[66] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, D.N. Metaxas, Learning active facial patches for expression analysis, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012), 2012.

Please cite this article as: D. Mery, K. Bowyer, Automatic facial attribute analysis via adaptive sparse representation of random patches, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.05.005