Pattern Recognition Letters 68 (2015) 260–269

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

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 30 May 2015

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/

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