AGE AND GENDER RECOGNITION USING INFORMATIVE FEATURES OF VARIOUS TYPES

Ehsan Fazl-Ersi, M. Esmaeel Mousa-Pasandi, Robert Laganière

School of Electrical Engineering and Computer Science
University of Ottawa
Ottawa, ON K1N 6N5 Canada

Maher Awad

ADitude Media inc.
Ottawa, ON
K2H 8K7 Canada

ABSTRACT

Gender recognition and age classification are important applications of face analysis. The vast majority of the existing solutions focus on a single visual descriptor which often encodes only a certain characteristic of the image regions (e.g., shape, or texture, or color, etc.). In this paper, we propose a novel framework for gender and age classification, which facilitates the integration of multiple feature types and therefore allows for taking advantage of various sources of visual information. Furthermore, in the proposed method, only the regions that can best separate face images of different demographic classes (with respect to age and gender) contribute to the face representations, which in turn, improves the classification and recognition accuracies. Experiments performed on a challenging publicly available database validate the effectiveness of our proposed solution and show its superiority over the existing state-of-the-art methods.

Index Terms— gender recognition, age classification, uniform LBP, face processing, color histogram, feature selection

1. INTRODUCTION

Age and gender recognition has long been recognized as an important module for many computer vision applications, such as human-robot interaction, visual surveillance and passive demographic data collections. More recently, the growing interest in the advertising industry for launching demographic-specific marketing and targeted advertisements in public places has attracted the attention of more researchers in the field of computer vision to the problem of age and gender recognition.

A key component in any gender classification system is face representation. While some methods choose to use raw pixels (e.g., [1], [2] and [3]) without any modification, the majority of the existing methods use local visual descriptors to produce stronger and often more compact representations for face images. Examples of visual descriptors commonly used for age and/or gender recognition are SIFT (e.g., used in [4] and [5]), LBP (e.g., used in [6] and [6]) and color histograms (e.g., used in [7]). In these methods, local descriptors are often extracted from a dense regular grid over the entire image and then the face representation is built by concatenating these extracted descriptors into a single vector. A key issue in this framework is to determine the optimal grid parameters (e.g., spacing, size, number of grids in multi-resolution/pyramid approaches, etc.). Dago-Casas et al. [8] proposed to use raw pixels, Gabor jets and LBPs on Gallaghers database for gender recognition. They reduced the size of extracted features by using Principle Component Analysis and showed that by using Gabor jets followed by SVM high gender classification accuracy is obtained. While previous methods used fixed settings and performed trial-and-error heuristics to determine the right grid parameters, in this paper, we suggest using feature selection to allow the most informative image regions (or grid cells) to contribute to the face representation, i.e., those that can best separate face images that belong to different demographic classes (with respect to age and gender). This approach further facilitates the integration of different types of descriptors (e.g., color based, shape based, texture based, etc.) and allows for more compact representations by preventing redundant features from contributing to the face representation. The approach in [9] also uses feature selection but in the context of gender recognition only. They used LBP, intensity histogram and gradient orientation features while here we added SIFT and color descriptions. We also used the more challenging Gallagher's database on which we performed both gender recognition and age classification.

In section 4, we show that our proposed face representation approach, combined with an SVM classifier, outperforms the existing age and gender classification methods on a challenging database developed by Gallagher et al. [10], by a decent margin.

The remainder of this paper is organized as follows: in Section 2, we describe the different steps of the proposed face representation method. Section 3 describes the learning and classification modules, and Section 4 presents the implementation details and experimental results. Finally, we conclude the paper and discuss some directions for future work in Section 5.

2. FACE REPRESENTATION

Unlike the vast majority of the existing methods that use a single type of descriptor based on a fixed setting (in terms of grid parameters), in our proposed method, a face image is represented by a collection of different types of local descriptors extracted from various regions across the image. This is due to the fact that each type of visual descriptor only captures certain information from an image region and can be used to complement the information captured by another type of descriptor. For example while Local Binary Pattern (LBP) descriptor encodes spatial relations between neighboring pixels and is useful to describe the texture of an image patch, a Scale Invariant Feature Transform (SIFT) builds local histograms of gradient orientations and is best to capture the shape attributes of an image patch. Therefore, extracting SIFT and LBP descriptors from proper locations in the face image (e.g., cheeks for LBP descriptor to distinguish between faces with and without beard, and nose and mouth for SIFT descriptor to distinguish between different faces based on the shape characteristics of these facial features) allows the produced face representations to take advantage of both sources of information and provide better distinctiveness for classification and recognition purposes.

To determine what type of visual descriptors and which regions in the image are most informative to contribute to the face representation, we suggest using feature selection (where a feature is defined as the couple descriptor type and image region) to choose the optimal set of features from a pool of candidate features. In the next two sub-sections, we explain how the pool of candidate features can be generated and how informative features can be selected, respectively.

2.1. Pool of Candidate Features

To generate the pool of candidate features, for each aligned face image (based on affine transformation determined from three facial landmarks, i.e., left eye, right eye and mouth center) in the training set, an image pyramid is built and then different types of visual descriptors are extracted from dense regular grids (with size and spacing of pixels) over the image at each level of the pyramid. This results in a large number of descriptors being extracted from various regions of the face images. In this paper, we consider three types of features, each encoding a certain characteristic (e.g., color, shape and texture) of an image region.

Local Binary Pattern (LBP): LBP [11] is a powerful texture-encoding descriptor based on occurrence statistics of a set of local binary patterns. In our implementation, we only take into account the uniform patterns. A uniform pattern is a Local Binary Pattern (LBP) with at most two bitwise transitions (or discontinuities) in the circular presentation of the pattern. When using a 3×3 neighborhood, for example, only 58 of the 256 total patterns are uniform, which yields in 59-dimensional image representation (i.e., histogram), one dimension for each uniform pattern and one dimension for all the non-uniform patterns.

Scale-Invariant Feature Transform (SIFT): SIFT [12] is a powerful description method for characterizing image regions, which has been widely used for various computer vision applications. SIFT produces a 128 dimensional representation for each image region using a 3D (2 locations and 1 orientation) histogram of gradient locations and orientations. The contribution of each pixel to the location and orientation bins is weighted by its gradient magnitude. The quantization of gradient locations and orientations makes SIFT descriptors robust to small geometric distortions and certain illumination variations.

Color Histogram (CH): Modeling color distribution can be very useful in characterizing an image or image region. In our implementation, color distribution is modeled by constructing a histogram with 4 bins per color channel, in the RGB color space. More specifically, the intensity values in each color channel (i.e., red, green and blue) are mapped into 4 values (e.g., intensity values between 0 and 63 are mapped to 0, and so on), and then each of the 64 (4^3) bins stores an integer counting the number of times that the corresponding color triplet occurred.

2.2. Feature Selection

For each candidate feature, n, a response vector, r_n, is generated by computing the similarity between different pairs of training faces, using only the descriptor extracted based on the specifications of the candidate feature, namely the location and the size of a region, and a type of descriptor (e.g., LBP, SIFT, or CH)[1].

A feature selection method based on [13] is then employed to choose the most informative features using the response vectors generated for all candidate features in the pool. First, a binary variable, f_n, is associated to each candidate feature, n, by mapping its response vector, r_n, to 0 (if the response is lower than threshold θ_n) and 1 (if the response is greater than θ_n)[2] .

Given a collection of binary variable, feature selection then attempts to select the most appropriate features that together can best separate the positive training pairs (i.e., pairs with both face images belonging to the same age or gender class) from the negative training pairs (i.e., pairs with both face images belonging to different age or gender class). To this aim, a binary variable C is generated to represent the ground-truth classification, where $C(I) = 1$ if the pair I is positive, and is 0, otherwise.

[1] The similarity between two descriptors is computed as the sum of absolute difference, irrespective of the type of the descriptor.

[2] The threshold θ_n is determined such that the mutual information between the resulting binary variable, f_n, and the class variable, C, is maximum.

The discriminative value of each feature is measured by the amount of mutual information it can deliver about the class:

$$I(f_n; C) = H(C) - H(C|f_n) \tag{1}$$

In the above equation, $I(f_n; C)$ is the mutual information between binary variable f_n and class C, and H denotes entropy. Feature selection starts by identifying the feature, whose binary variable generates the highest mutual information score. It then proceeds by iteratively searching for the next informative feature, f_r, that delivers the maximal amount of additional information with respect to each of the previously selected features:

$$f_r = arg \max_{f_k \in K_r} \min_{f_j \in S_r} (I(f_k, f_j; C) - I(f_j; C)) \tag{2}$$

Here K_r and S_r are the set of features not yet selected, and the set of features already selected at iteration r, respectively.

The feature selection process ends when the increment in mutual information gained by selecting a new feature is less than a certain threshold, or until the number of selected features reaches a certain limit.

3. RECOGNITION AND CLASSIFICATION

The feature selection process provides us with a set of features, each representing a certain region in the face image and specifying a particular descriptor type to be extracted from that region.

For classification and/or recognition, we use Support Vector Machines (SVM) [14] with RBF kernels. Given that the face representations in our proposed methods are comprised of different types of descriptors, we compute three RBF kernels, one for each descriptor type, normalize them by their respective means, and linearly combine those using weights proportional to the frequency of each descriptor types in the set of selected features.

When the task is to distinguish between more than two classes (e.g., age group identification), we use SVM with one-versus-all rule: a classifier is trained to separate each class from the rest and a test image is assigned to the class whose classifier returns the highest response.

4. EXPERIMENTS

4.1. Database

Unlike most methods (particularly in the context of gender recognition) that have been evaluated only on controlled databases such as FERET [15], we conduct our experiments on a very challenging database with real world images, namely the Gallagher's database [10], to ensure that our proposed system can be generalized well for real world applications.

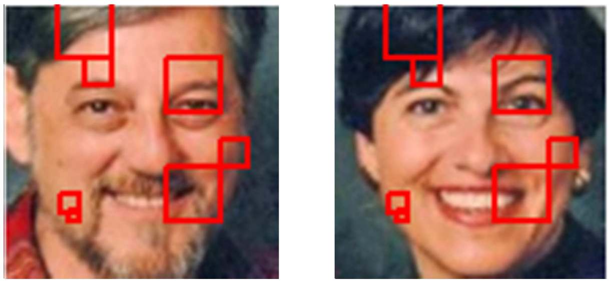

Fig. 1. The seven most informative LBP features as selected by Ullman feature selection technique for gender classification.

The Gallagher's database, which is publicly available, is composed of $28,231$ faces collected from Flickr images taken under unconstrained conditions. The faces are labeled based on their gender and their association to one of 7 age groups (covering from 0 to $+75$ years). In our experiments on this database, we use the folding suggested by Dago-Casas et al. [16], which employs $14,760$ of the higher resolution faces and distribute them into 5 folds, with equal number of males and females in each fold. Following the experimental protocol suggested in [8], images of 4 folds are used to train the models, and images of the remaining fold are used to test the trained models. The experiment is repeated five times, each time with a different fold to be used for testing, and the final result is reported as the mean of the performances obtained in individual runs.

While we use the same experimental procedure for age classification and gender recognition, for age classification, we randomly remove a subset of the face images in the training set that belong to more frequent age groups, to ensure equal number of training images for different age classes.

4.2. Results

In this experiment, results using gender and age recognition based on three different types of descriptors have been gathered. We extracted texture, shape and color features in different spatial scales for each image in Gallagher's dataset and then concatenated them into a single feature vector. Then we select the most 200 informative feature bins among all resulting vectors. As an example, Figure 1 illustrates the most informative LBP features for the case of gender classification. Applying SVM with RBF kernel to the database representing selected feature vectors, outperforms [8] in terms of accuracy.

Table 4.2 shows comparative age and gender recognition results with different methods tested on Gallagher's dataset. The results are averaged over 5 folds of each method for both gender and age recognition rate in percentage. Each fold contains 2952 persons images under various changes of illumination, camera pose and quality. Each method uses 200 feature bins with most discriminative power selected from different

Table 1. Comparative results of the gender and age recognition systems on Gallagher's dataset.

Method	gender recognition				age recognition			
	accuracy	feature bins			accuracy	feature bins		
LBP	90.43	200			55.88	200		
CH	82.82	200			42.30	200		
SIFT	89.61	200			54.18	200		
LBP+CH+SIFT	91.59	130 LBP	15 CH	55 SIFT	63.01	110 LBP	35 CH	55 SIFT
Pixels+PCA [8]	80.11	Not reported			N/A	N/A		
Gabor Jets+PCA [8]	86.61	Not reported			N/A	N/A		
LBPs+PCA [8]	86.69	Not reported			N/A	N/A		

region sizes.

As shown in Table 4.2, the gender recognition rate reached 91.59% by using 130 LBP, 55 SIFT, and 15 CH bins. Similarly, the number of features that are selected in age recognition are 110, 55, and 35 for uniform LBP, SIFT and CH respectively. As a consequence, the combination of uniform LBP with SIFT and Color histogram shows the saliency of texture over shape and color information. Adding shape and color information to the texture descriptor improves the recognition rate by 1.16% with respect to pure LBP.

Table 2. Confusion matrix for five age classes(numbers are normalized).

Prediction Actual	(0-12)	(13-19)	(20-36)	(37-65)	(66+)
(0-12)	**84.5187**	10.4980	2.5590	1.4549	0.9694
(13-19)	13.1562	**54.4231**	23.8387	6.8705	1.7116
(20-36)	3.5302	25.8553	**46.4744**	19.6832	4.4569
(37-65)	1.8250	9.5284	21.0849	**44.3041**	23.2575
(66+)	0.5023	1.5804	2.1552	10.4260	**85.3361**

Table 4.2 presents the confusion matrix for five age classes with the method based on combination of LBP, CH and SIFT features. As expected, most of the confusion occurs between adjacent classes. For instance, it is clear from the fifth row of Table 4.2 that mature adults are often misclassified as young adult or senior classes, which is a commonly made mistake.

5. CONCLUSIONS

In this paper we presented a novel gender and age classification method, that unlike the vast majority of the existing solutions that focus on a single visual descriptor (and therefore limiting the face representations by encoding only a certain characteristic of the image regions, such as, shape, or texture, or color), facilitates the integration of multiple feature types and allows for taking advantage of various sources of visual information. The proposed method, based on the selection of informative features, only allows the regions that can best separate face images of different demographic classes (with respect to age and gender) to contribute to the face representations, which in turn, improves the classification and recognition accuracies. A set of experiments conducted on the challenging Gallagher's database validated the effectiveness of our proposed solution in accurately classifying the age and gender of face images taken under unconstrained conditions.

As a potential future direction, we plan to explore the possibility of using a multi-class (versus binary) feature selection method to study its impact on improving the age classification accuracy. We further plan to build a database from images of people captured in real-world scenarios (e.g., images from people watching a public TV display), to generalize our method for real-world applications. By integrating the proposed age and gender classifier with a reliable tracker (e.g. [17]) and, possibly, a face quality assessment measure (e.g. [18]), a real-time demographics visual system can be built.

6. REFERENCES

[1] B. Moghaddam and M. H. Yang, "Learning gender with support faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 707–711, May 2002.

[2] P. Jonathon H. Wechsler, J.R.J. Huang and S. Gutta, "Mixture of experts for classification of gender, ethnic origin, and pose of human faces," *IEEE Transactions on Neural Networks*, vol. 11, pp. 948–960, July 2000.

[3] H. A. Rowley and S. Baluja, "Boosting sex identification performance," *International Journal of Computer Vision*, vol. 71, pp. 111–119, January 2007.

[4] J. G. Wang E. Sung, J. Li and W. Y. Yau, "Boosting dense sift descriptors and shape contexts of face images for gender recognition," *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 96–102, June 2010.

[5] J. G. Wang C. Y. Lee, J. Li and W. Y. Yau, "Dense sift and gabor descriptors-based face representation with applications to gender recognition," *International Conference on Control Automation Robotics and Vision*, pp. 1860–1864, December 2010.

[6] C. Sun C. Zou N. Sun, W. Zheng and L. Zhao, "Gender classication based on boosting local binary pattern," *International Symposium on Neural Networks*, vol. 3972, pp. 194–201, June 2006.

[7] L. Bourdev J. Malik and S. Maji, "Describing people: A poselet-based approach to attribute classification," *IEEE International Conference on Computer Vision (ICCV)*, pp. 1543–1550, November 2011.

[8] L. L. Yu D. Gonzalez-Jimenez, J. L. Alba-Castro and P. Dago-Casas, "Single-and cross-database benchmarks for gender classification under unconstrained settings," *IEEE International Conference on Computer Vision Workshops*, pp. 2152–2159, November 2011.

[9] J.E. Tapia and C.A. Perez, "Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of lbp, intensity, and shap," *IEEE Transactions on Information Forensics and Security*, vol. 8, pp. 488–499, March 2013.

[10] A. Gallagher and T. Chen, "Understanding groups of images of people," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 256–263, June 2009.

[11] T. Maenpaa M. Pietikainen and T. Ojala, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 971–987, July 2002.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, January 2004.

[13] M. Vidal-Naquet and S. Ullman, "Object recognition with informative features and linear classification," *IEEE International Conference on Computer Vision (ICCV)*, pp. 281–288, October 2003.

[14] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[15] P. J. Rauss H. Moon, P. J. Phillips and S. A. Rizvi, "The feret evaluation methodology for face recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1090–1104, October 2000.

[16] H. Ai and Z. Yang, "Demographic classification with local binary patterns," *Advances in Biometrics*, vol. 4642, pp. 464–473, August 2007.

[17] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.

[18] Adam Fourney and Robert Laganiere, "Constructing face image logs that are both complete and concise.," in *CRV*, 2007, pp. 488–494.