Gender Recognition in the Wild with Small Sample Size - A Dictionary Learning Approach

Alessandro D'Amelio[0000−0002−8210−4457], Vittorio Cuculo[0000−0002−8479−9950], and Sathya Bursic[0000−0001−8327−5007]

Dipartimento di Informatica
University of Milan, Milano, Italy
{vittorio.cuculo,alessandro.damelio,sathya.bursic}@unimi.it

Abstract. In this work we address the problem of gender recognition from facial images acquired in the wild. This problem is particularly difficult due to the presence of variations in pose, ethnicity, age and image quality. Moreover, we consider the special case in which only a small sample size is available for the training phase. We rely on a feature representation obtained from the well known VGG-Face Deep Convolutional Neural Network (DCNN) and exploit the effectiveness of a sparse-driven sub-dictionary learning strategy which has proven to be able to represent both local and global characteristics of the train and probe faces. Results on the publicly available LFW dataset are provided in order to demonstrate the effectiveness of the proposed method.

Keywords: Facial Gender Recognition · Sparse Dictionary Learning · Deep Features · Soft Biometrics.

1 Introduction

Human gender recognition is a problem of soft biometrics that has gained a lot of attention in the recent years. In particular the problem of recognizing gender from human faces is typically used in applications like human computer interaction, image retrieval, surveillance, market analysis or for the improvement of traditional biometric recognition systems, moreover, has gained popularity due to large availability of face datasets. In the past few years a lot of research has been carried on, mainly focusing on the problem of gender recognition from faces in a constrained setting (e.g. frontal images, controlled lighting conditions, absence of occlusions). However in order to produce applications that can be used in every day situations (e.g. web pages, webcam, mobile devices) it's necessary to build models which are able to deal with face images in an unconstrained setting; this includes images with occlusions, facial expressions, variation of pose and lighting condition and low resolution images. In the most recent literature, this problems have been addressed using both hand crafted features and Deep Learning and in particular Deep Convolutional Neural Networks (DCNN). In this work we propose a method for the classification of gender from facial images acquired in an unconstrained setting and when only few examples are available

for the training phase. The paper is organized as follows: in the next section we give a brief review of the related work and the state of the art. In Section 3 a detailed description of the proposed algorithm is provided. In Section 4 experimental results of the proposed method on a benchmark dataset are presented with related discussion.

2 Related Work

The problem of gender recognition from face images has been addressed in many works in the recent literature [12], however most of those focus on datasets acquired in a controlled environment, that is out of the scope of the present investigation. To a first approximation, the methods for gender recognition in the wild follow two distinct paths: in one case, a standard classification pipeline is adopted and the dataset is divided in training and test sets. For each image a feature extraction procedure is implemented followed by a machine learning method stated as a binary classification model, eventually preceded by a dimensionality reduction or feature selection module. In this category falls the work of Dago-Casas et al [7] in which Gabor features are extracted and the classification step is carried out using a linear SVM. In order to deal with the imbalance of classes a weighted SVM model is adopted for classification. Shan [16] used Boosted Locally Binary Pattern as features for the classification with a SVM with RBF kernel, while Tapia et al. [17] adopted a feature selection algorithm based on mutual information and fusion of intensity, shape and texture features as input for an SVM classifier. However, both used a subset of the LFW dataset [10] composed of 4500 males and 2943 females, excluding images that did not contain near frontal faces. The second group of methods concerns those who exploit the deep learning for the classification of gender. In [2], Afifi et al rely on the combination of isolated and holistic facial features used to train deep convolutional neural networks followed by an AdaBoost-based score fusion to infer the final gender class. In [15] a Deep Multi-Task Learning Framework called HyperFace is proposed; gender recognition is presented as one of the tasks and is carried out by fusion of different DCNN features (each of which has been trained for a specific task) via a multi-task learning algorithm which exploits the synergy among the tasks to boost up the performances.

3 Proposed method

The proposed method relies on the construction of highly discriminative dictionaries of deep features. Each training (gallery) or test image is first processed using standard image augmentation techniques. For each of those images, the feature characterization from the VGG-Face Net is computed. In the training step only a small subset of images is retained to build the gallery. A sparse-driven sub-dictionary learning strategy is then adopted to build the dictionary. Probe faces are classified via sparse recovery on the learnt dictionary. The overall schema of the model is sketched in Fig. 1. At training stage (red route), each

augmented face image (Section 3.1) is passed through the VGG-Face DCNN and according to the associated label, the obtained feature may be selected to initialize the corresponding sub-dictionary (male or female) prior to the learning stage (Section 3.3). At testing stage (blue route) the probe face image, after augmentation and feature extraction, is classified through sparse recovery on the learnt dictionary (Section 3.4).

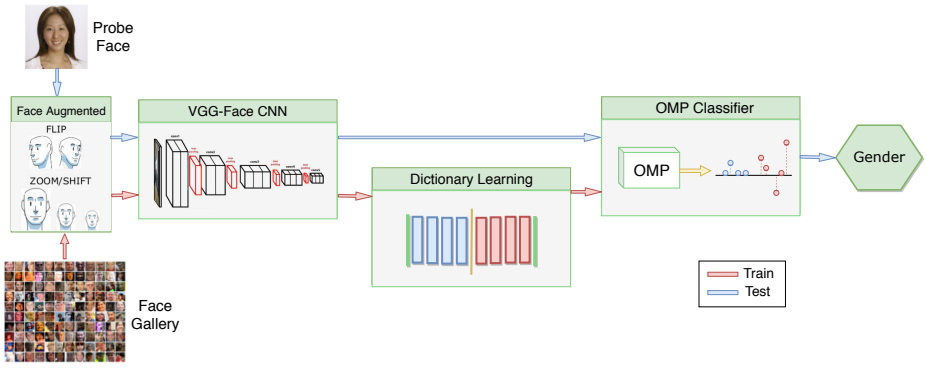

Fig. 1. Pipeline of the model

3.1 Augmented DCNN features

In order to build a dictionary we need to find a feature representation that can be suitable for the discrimination of certain characteristics of human faces. In this work we decided to exploit the effectiveness of Deep Convolutional Neural Networks and in particular the popular architecture VGG-Face Net [13]. Given that we set up the problem to have a small sample size approach, it would be impossible to train a DCNN from scratch; conversely we use a pre-trained network as feature extractor by feeding images into the VGG-Face and taking the output of the network truncated at the last fully connected layer. It is worth noticing that original VGG-Face architecture was trained to recognize face identities with no explicit information about the gender. Moreover there is no overlap between the test dataset and the one used to train the VGG-Net. As pointed out in [5], an augmentation stage which relies on standard image transformations (e.g. flipping, random crops, rescalings) is beneficial for the recognition accuracy. Moreover, adopting augmentation techniques delivers some advantages: first by cropping, scaling and flipping faces at different levels we are able to extract both local and global characteristics for the problem at hand. Secondly, this leads to an increase of the number of training images, thus allowing to have a smaller sample size for the gallery. The augmented feature extraction procedure is carried out by applying to each image the same set of transformations (9 crops, 4 scales, 2 flips), thus producing $L = 72$ transformations. For each "augmented"

image, a feature characterization is extracted using the VGG-Face Net, which delivers a 4096 dimensional feature vector.

3.2 Sparse Representation of DCNN Features

It is assumed that every feature representation of each augmented face image from class i can be recovered from the linear combination of the training data from that class: $\mathbf{y}_{i,j} = x_{i,1}\mathbf{v}_{i,1} + x_{i,2}\mathbf{v}_{i,2} + ... + x_{i,n_i}\mathbf{v}_{i,n_i}$, where $\mathbf{y}_{i,j} \in \mathbb{R}^{\mathrm{m}}$ is the j-th feature vector for the generic augmented face image and x values are the coefficients corresponding to the training data samples for class i. In other words we assume that the feature representation of a given face image with a certain gender can be approximated by linearly combining the features belonging to other images of the same gender. Following this idea, we can represent every generic feature vector $\mathbf{y}_j$ associated to a new test image as:

$$\mathbf{y}_j = \mathbf{D}\mathbf{x}_j \tag{1}$$

Here $\mathbf{D}$ is commonly referred to as the *dictionary* which is, in general, a matrix where each column is a feature vector associated to a training example. From now on, j index will be omitted unless needed. Assume to select n_i training data samples for the i-th class, where each data sample is represented by a vector $\mathbf{v}_{i,j}$ of m elements. These vectors are then used to construct the columns of matrix A_i:

$$\mathbf{A}_i = [\mathbf{v}_{i,1}, \mathbf{v}_{i,2}, ..., \mathbf{v}_{i,n_i}] \tag{2}$$

In the specific scenario adopted here, we have two classes, so $i \in \{\male, \female\}$, $n_i = q_i * K$ where q_i is the number of images from the i-th class selected to join the gallery (this number is chosen to be reasonably small and equal for both classes, so $q_{\male} = q_{\female}$) and $K << L$ is the number of augmented features to keep in the dictionary for every training image. Each feature vector of 4096 elements is reduced dimensionally using Principal Component Analysis (PCA) in order to retain 95% of the variance. The concatenation of the $\mathbf{A}_i$ matrices yields the dictionary:

$$\mathbf{D} = [\mathbf{A}_{\male}, \mathbf{A}_{\female}] \in \mathbb{R}^{\mathrm{m}\times\mathrm{n}} \tag{3}$$

Where m is the result of the dimensionality reduction step and $n = n_{\male} + n_{\female}$. The solution of the linear equation (1) boils down to the problem of solving an under determined system given that the matrix $\mathbf{D}$ is a $m \times n$ matrix with $m < n$. The "common way" of solving this kind of systems is by defining an optimization problem of the form: $\min_{\mathbf{x}} \mathbf{J}(\mathbf{x}) \quad s.t. \quad \mathbf{y} = \mathbf{D}\mathbf{x}$.

The form of $\mathbf{J}(\mathbf{x})$ governs the kind of solutions we may obtain. If we choose $\mathbf{J}(\mathbf{x})$ to be the squared Euclidean norm $\|\mathbf{x}\|_2^2$, then the solution to the optimization problem can be obtained in closed form by $\mathbf{x} = \mathbf{D}^{\dagger}\mathbf{y}$ which is the standard least-squares solution, where $\mathbf{D}^{\dagger}$ is the pseudo-inverse of the matrix $\mathbf{D}$. However this kind of solution is not suitable for the problem of recognition. In fact, we

wish to obtain the sparsest solution in order to "select" only those atoms of the matrix $\mathbf{D}$ that correspond to the correct class at hand. In practice, if we have a probe image containing, say, a female face, we (ideally) wish to recover a solution for the vector $\mathbf{x}$ in which the support of $\mathbf{x}$ is non-zero only at the columns of $\mathbf{D}$ that belong to $\mathbf{A}_{\female}$.

In order to yield the sparsest solution for $\mathbf{x}$, an $\mathcal{L}_0$ norm is chosen for $\mathbf{J}(x)$: $\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad s.t. \quad \mathbf{y} = \mathbf{D}\mathbf{x}$. Moreover, when at most k atoms are sufficient to represent the sample $\mathbf{y}$, the previous problem can be rewritten as in the following:

$$\mathbf{y} = \mathbf{D}\mathbf{x} \quad s.t. \quad \|\mathbf{x}\|_0 \leq k \tag{4}$$

Since data in real applications often contains noise, the model appearing in the previous equation is somewhat unrealistic. Thus, it is reasonable to revise such exact model introducing a noise assumption:

$$\min_{\mathbf{x}} \|\mathbf{D}\mathbf{x} - \mathbf{y}\|_2^2 \quad s.t. \quad \|\mathbf{x}\|_0 \leq k \tag{5}$$

Finding a solution to this optimization problem is an NP-Hard problem, but approximations can be found using approximate algorithms [18].

3.3 Sparse sub-Dictionary Learning

In this section we aim at building class specific sub-dictionaries of the form of Eq. 2 able to capture the sparsity patterns for the gender classification problem. In the vein of [6], this can be achieved by learning such sub-dictionaries to well represent face characteristics through the sparse vectors $\mathbf{x}$. To this end the *sparse dictionary learning problem* can be defined as follows:

$$\min_{\mathbf{X},\mathbf{D}} \|\mathbf{D}\mathbf{X} - \mathbf{Y}\|_F^2 \quad s.t. \quad \|\mathbf{x}_u\|_0 \leq k, \quad \|\mathbf{y}_u\|_2 = 1 \tag{6}$$

where $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_q] \in \mathbb{R}^{\mathrm{m}\times\mathrm{q}}$ is the data matrix obtained by concatenating column-wise all the $q = (q_{\male} + q_{\female}) \times L$ training feature vectors and similarly, $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_q] \in \mathbb{R}^{\mathrm{n}\times\mathrm{q}}$ is the matrix of the corresponding sparse representations. There is no closed form solution for the problem defined by Eq. 6 in the same way as there isn't for the problem of Eq. 5. However, this can be heuristically solved by adopting the well-established alternating optimization scheme consisting in repeatedly executing these two steps until a stop condition is met:

- *Sparse coding:* solve problem (6) for $\mathbf{X}$, fixing the dictionary $\mathbf{D}$
- *Dictionary update:* solve (6) for each $\mathbf{A}_i$ separately, fixing $\mathbf{X}_i$, then build the new dictionary as in Eq. 3

At the very first step, the dictionary $\mathbf{D}$ may be initialized by randomly selecting training feature vectors; the sparse coding step can then be solved resorting to standard sparse approximation algorithms like [1] or [14] as well as for the dictionary update rule that can be casted into different forms (e.g. [3, 9, 8]).

In this work we exploit the well established Orthogonal Matching Pursuit (OMP) [14] and K-SVD [3] algorithm for the sparse coding and dictionary update steps, respectively.

3.4 Classification via Sparse Recovery

The problem of recognizing the gender of a new probe subject can be casted to the recovery of the sparsest solution of a linear system. In particular, given a dictionary $\mathbf{D}$, of the form of Eq. 3, and a new probe image I, the augmentation step is computed yielding L transformed images (I_l). The feature extraction is then performed via the VGG-Face DCNN. For each augmented feature vector, the PCA projection computed in the dictionary learning phase is applied, thus obtaining $L = 72$ (9 crops $\times$ 4 scales $\times$ 2 flips) feature vectors ($\mathbf{y}_l$). For each of the obtained $\mathbf{y}_l$, the sparse recovery on the learnt dictionary $\mathbf{D}$ is computed using OMP, following Eq. 5. The solutions have the following form: $\mathbf{x}_l = \left[\mathbf{x}_{l,\male}; \mathbf{x}_{l,\female}\right]$; for each $\mathbf{x}_l$ vector ($l \in \{1, ..., 72\}$), the number of non-zero elements in $\mathbf{x}_{l,\male}$ and $\mathbf{x}_{l,\female}$ is counted and classification is performed by majority voting.

4 Experimental Results

The effectiveness of the method is assessed on the Labeled Faces in the Wild (LFW) dataset [10]. This dataset contains more than 13000 images of 5749 different subjects acquired in uncontrolled conditions. The pose, illumination, and expression variations, together with the possible presence of partial occlusions and disguised faces make the gender recognition problem challenging.

The original release of the LFW dataset does not contain gender labels, however Afifi et al. [2] used an estimation method for the gender label based on the first name of the subjects; the obtained labels where then reviewed three times to completely eliminate any incorrect labels. Besides the difficulties outlined in the previous paragraph, the LFW dataset adds another hitch to the gender recognition problem, namely a huge imbalance between the two classes. In particular the dataset is composed by 10268 examples for the male category and only 2966 for the female one. For what concerns the cardinality of the gallery, we conducted 3 experiments using $q_{\male} = q_{\female} = [50, 100, 200]$, in order to asses the importance of the size of the gallery on the classification accuracy. In other words, among the 13234 images of the dataset, only, 50, 100 or 200 are in turn selected from the male and female category respectively and used for training, while all the others are used for test. We experimentally set $K = 5$, that is for every gallery image, 5 feature vectors out of the 72 are randomly chosen to join the appropriate subdictionary, prior to the learning phase. All the three experiments are repeated 10 times each. In each trial the set of images to be put in gallery is selected by uniformly sampling 50, 100 or 200 images from both classes. This ensures that in each trial the model potentially has a different set of images, identities, occlusions and ethinicities in gallery, while maintaining the distribution of males and females constant. For each trial a learning and testing phase is executed and the results are averaged. In Table 4 mean results are displayed alongside comparison with state of the art methods on LFW proposed in literature. Precision, recall and F1-measure are reported if available.

As shown in the table, the method proposed here yields comparable results with other models in the literature. Notably, by augmenting the size of the

Method	Accuracy	Precision	Recall	F1-measure
Our {50} (10268m 2966f)	91.73	80.57	91.23	85.57
Our {100} (10268m 2966f)	94.43	84.71	95.10	89.60
Our {200} (10268m 2966f)	95.13	86.10	94.42	90.06
Gabor+W-SVM (10129m 2959f) [7]	92.96	94.10	89.05	91.50
Boosted LBP+SVM (4500m 2943f) [16]	94.81	-	-	-
LBP+SVM (4500m 2943f) [17]	98.01	-	-	-
$AFIF^4$ [2]	95.98	-	-	-
HyperFace [15]	94.00	-	-	-

Table 1. Experiments on LFW dataset and comparisons. For each method we report the accuracy, precision, recall and F1 measure. The cardinality of the two classes is shown in brackets. For the proposed method the number of images per class that compose the gallery is put in curly brackets

gallery, both the accuracy and the F1-measure increase, reaching results that sensibly outperform those in [7] in terms of accuracy. Among the other methods outlined, [7] is the method that uses the biggest subset of LFW; in fact, many of the analyzed models act on a subset of the LFW dataset by rejecting images for which face detection fails; we believe that this would lead to exclude from the analysis the most challenging images. Moreover in some works the most numerous class is sub-sampled in order to obtain a class balanced dataset. To the best of our knowledge, the method proposed in [7] is the only one that clearly provides results for the whole LFW dataset on the gender recognition problem.

5 Conclusions

In this work a method for the classification of gender from facial images in the wild is proposed. The method exploits the effectiveness of the sparse-driven sub-dictionary learning strategy on DCNN features formerly presented in [6]. The experimental results show that the proposed method is able to deal with variations in pose, lighting, occlusions, facial expressions and ethnicity while using a training set (gallery) with a small sample size. The results obtained are comparable with the state of the art on the LFW dataset, despite the huge difference in the cardinality of both the training set and the test set used. In future work, we plan to explicitly inquire the impact of specific hurdles (facial expressions, occlusions, etc.) by relying on appropriate datasets [11, 4].

References

1. Adamo, A., Grossi, G., Lanzarotti, R., Lin, J.: Sparse decomposition by iterating lipschitzian-type mappings. Theoretical Computer Science **664**, 12–28 (2017)

2. Afifi, M., Abdelhamed, A.: Afif4: Deep gender classification based on adaboost-based fusion of isolated facial features and foggy faces. arXiv preprint arXiv:1706.04277 (2017)
3. Aharon, M., Elad, M., Bruckstein, A., et al.: K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on signal processing **54**(11), 4311 (2006)
4. Boccignone, G., Conte, D., Cuculo, V., Lanzarotti, R.: Amhuse: a multimodal dataset for humour sensing. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. pp. 438–445. ACM (2017)
5. Bodini, M., D'Amelio, A., Grossi, G., Lanzarotti, R., Lin, J.: Single sample face recognition by sparse recovery of deep-learned lda features. In: International Conference on Advanced Concepts for Intelligent Vision Systems. pp. 297–308. Springer (2018)
6. Cuculo, V., D'Amelio, A., Grossi, G., Lanzarotti, R., Lin, J.: Robust single-sample face recognition by sparsity-driven sub-dictionary learning using deep features. Sensors **19**(1), 146 (2019)
7. Dago-Casas, P., González-Jiménez, D., Yu, L.L., Alba-Castro, J.L.: Single-and cross-database benchmarks for gender classification under unconstrained settings. In: Computer vision workshops (ICCV Workshops), 2011 IEEE international conference on. pp. 2152–2159. IEEE (2011)
8. Engan, K., Aase, S.O., Husoy, J.H.: Method of optimal directions for frame design. In: 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258). vol. 5, pp. 2443–2446. IEEE (1999)
9. Grossi, G., Lanzarotti, R., Lin, J.: Orthogonal procrustes analysis for dictionary learning in sparse linear representation. PloS one **12**(1), e0169663 (2017)
10. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition (2008)
11. Martinez, A.M.: The ar face database. CVC Technical Report24 (1998)
12. Ng, C.B., Tay, Y.H., Goi, B.M.: A review of facial gender recognition. Pattern Analysis and Applications **18**(4), 739–755 (2015)
13. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: BMVC. vol. 1, p. 6 (2015)
14. Pati, Y.C., Rezaiifar, R., Krishnaprasad, P.S.: Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In: Proceedings of 27th Asilomar conference on signals, systems and computers. pp. 40–44. IEEE (1993)
15. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(1), 121–135 (2019)
16. Shan, C.: Learning local binary patterns for gender classification on real-world face images. Pattern recognition letters **33**(4), 431–437 (2012)
17. Tapia, J.E., Perez, C.A.: Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of lbp, intensity, and shape. IEEE transactions on information forensics and security **8**(3), 488–499 (2013)
18. Zhang, Z., Xu, Y., Yang, J., Li, X., Zhang, D.: A survey of sparse representation: algorithms and applications. IEEE access **3**, 490–530 (2015)