HAL
open science

A JOINT LEARNING APPROACH FOR CROSS DOMAIN AGE ESTIMATION

Binod Bhattarai, Gaurav Sharma, Alexis Lechervy, Frédéric Jurie

▶ **To cite this version:**

Binod Bhattarai, Gaurav Sharma, Alexis Lechervy, Frédéric Jurie. A JOINT LEARNING APPROACH FOR CROSS DOMAIN AGE ESTIMATION. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Mar 2016, Shanghai, China. hal-01301390

HAL Id: hal-01301390

https://hal.science/hal-01301390v1

Submitted on 12 Apr 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A JOINT LEARNING APPROACH FOR CROSS DOMAIN AGE ESTIMATION

Binod Bhattarai[1], *Gaurav Sharma*[2], *Alexis Lechervy*[1], *Frederic Jurie*[1]

[1]CNRS UMR 6072, University of Caen Normandy, ENSICAEN, France
[2]Max Planck Institute for Informatics, Saarbrücken, Germany

ABSTRACT

We propose a novel joint learning method for cross domain age estimation, a domain adaptation problem. The proposed method learns a low dimensional projection along with a regressor, in the projection space, in a joint framework. The projection aligns the features from two different domains, i.e. source and target, to the same space, while the regressor predicts the age from the domain aligned features. After this alignment, a regressor trained with only a few examples from the target domain, along with more examples from the source domain, can predict very well the ages of the target domain face images. We provide empirical validation on the largest publicly available dataset for age estimation i.e. MORPH-II. The proposed method improves performance over several strong baselines and the current state-of-the-art methods.

1. INTRODUCTION AND RELATED WORK

Automatic age estimation from face images has become a popular research problem [1, 2, 3, 4, 5]. It has various important applications such as age specific human-computer interaction [6], business intelligence [7] etc. Previous studies [8, 9, 10, 11] have shown that the rate of aging among different groups of people is different. This is because, aging patterns are directly affected by genes, dieting habits, culture, weather, race, gender etc. Thus, it has been more challenging to design an age prediction model which generalizes for people from such different categories. In addition, it has been shown that, training a single model on all different groups together, affect the performance that separate specialized models for different groups can give, due to the differences in aging patterns [9].

Training separate model for each and every group of people has its own limitations. It is difficult, expensive and time consuming to collect and annotate face images. Moreover, due to privacy related concerns, people may not be keen to share information about them such as ages, race etc. Thus, it would be ideal to utilise the training examples available for one group of people to improve performance in another group which has a very limited number of training examples. In this paper we are interested in such a setting as illustrated in Fig. 1.

This project is funded in part by the ANR (grant ANR-12-SECU-0005)

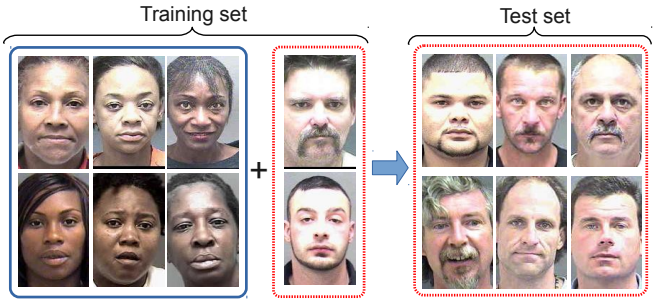

Fig. 1. Illustration of the proposed setting of cross domain age estimation. The algorithm learns a projection and a regressor jointly, to align source and target face domains and predict ages in the target domain. The training is mainly with source domain examples complemented very few target domain examples, while testing is done on target domain images only. The source and target domains may differ in age range, sex, race etc.

As explained above, we are interested in the problem of estimating age from face images, in a cross-population setting i.e. we have a large number of training examples available in one domain (the source domain) but only a very few ones in another domain (the target domain). We would like to utilise the training examples of the source domain to improve the performance of age estimation on the target domain. This problem was first posed and addressed by Guo et al. [12]. In their approach, they used a variant of LDA (Linear Discriminant Analysis) to learn common projection matrix which aligns aging patterns from source and target. However, they need a large number of target instances to learn target domain aging pattern, which are often not available in practice. Similarly, Alnajar et al. [13] proposed a method to do cross expression age estimation. But, the datasets they used for their experiments, FACES and LifeSpan are rather small and do not reflect the situation where abundant training data is available in the source domain.

We propose a joint learning method which (i) learns a subspace for aligning features from source and target domain and (ii) learns a regressor in this subspace for predicting ages. Our projection learning approach is similar to the metric learning method of Mignon and Jurie [14] – the projection matrix is

learnt to satisfy sparse pairwise (dis)similar constraints and age prediction based constraints simultaneously. We show empirically that the proposed method is consistently better than several strong baselines including those based on discriminative metric learning. We obtain state-of-the-art performance on the largest publicly available age estimation dataset. In the following, we discuss the proposed method in Sec. 2 then in Sec. 3 we provide the experimental results and, finally, in we conclude in Sec. 4.

2. PROPOSED METHODS

We now explain the proposed method in detail. We first introduce Metric Learning (ML) in general and then we explain how it can be used for learning a projection to align features from source and target domains. Finally, we explain the proposed Joint Learning (JL) algorithm.

2.1. Metric Learning and its application to cross-domain classification

Metric Learning (ML) has been quite successful in various facial analysis tasks such as face recognition [14, 15] and face retrieval [16]. Mahalanobis-like ML can be seen as learning a projection to map high dimensional features into a lower dimensional subspace where the pairwise constraints are better satisfied. For a pair of descriptors $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^{d'}$, ML involves the task of learning a Mahalanobis-like metric of the form $D^2_M(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top M(\mathbf{x}_i - \mathbf{x}_j)$, parameterized by positive semi-definite matrix M. As M is PSD, it can be decomposed as $M = L^\top L$. The problem can then be re-formulated as that of finding a linear subspace, into which features are first mapped and then compared with Euclidean distance i.e.,

$$D^2_L(\mathbf{x}_i, \mathbf{x}_j) = ||L\mathbf{x}_i - L\mathbf{x}_j||^2_2 \tag{1}$$

In the present case, we are given a training set of face images represented by their feature vectors and annotated with their ages i.e. $\mathcal{T} = \{(X, Y) : X \in \mathbb{R}^{d' \times N}, Y \in \mathbb{N}^N\}$. We construct two other sets from this information, set of *similar* vectors $\mathcal{S}$ annoated as $y_{ij} = 1$ and that of *dissimilar* ones $\mathcal{D}$, annotated as $y_{ij} = -1$, given by

$$\mathcal{S} = \{(i, j) : |y_i - y_j| \leq \delta\} \tag{2}$$

$$\mathcal{D} = \{(i, j) : |y_i - y_j| > \delta\} \tag{3}$$

with $\delta = 0$. We are interested in learning a mapping $f : \mathbb{R}^d \to \mathbb{R}^+$ to predict the age of new test faces where $d \ll d'$. We impose pairwise similarity and dissimilarity constraints, in the present case, and formulate the learning similar to the approach of Mignon and Jurie [14] i.e. optimize the objective function given as,

$$\min_L \mathcal{L}(\mathcal{T}, \mathcal{S}, \mathcal{D}; L) = \sum_{\mathcal{S} \cup \mathcal{D}} \ell_L(\mathbf{x}_i, \mathbf{x}_j, y_{ij}) \tag{4}$$

$$\ell_L(\mathbf{x}_i, \mathbf{x}_j, y_{ij}) = \max[0, m - y_{ij}(b - D^2_L(\mathbf{x}_i, \mathbf{x}_j))],$$

Algorithm 1 Joint learning of projection and regressor

Input: (i) Projection matrix, L ; Regressor $\mathbf{w}$, ii) Set of face features $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{d' \times N}$, Set of age annotations $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_N] \in \mathbb{R}^N$ (ii) Sparse pairwise age annotation $\mathcal{S}, \mathcal{D}$ (iii) maximum iterations `max-iters`, $\epsilon, \alpha, \beta, \gamma$, `learn-rate:`$r$
Output: L, $\mathbf{w}$
while $it <$ `max-iters` **do**
  $\Delta y_i \leftarrow |\mathbf{w}L\mathbf{x}_i - y_i|$
  if $\Delta y_i > \epsilon$ **then**
    $L_{it} \leftarrow L_{it-1} - \beta r \mathbf{w}_{it-1}\mathbf{x}_i^\top$
    $\mathbf{w}_{it} \leftarrow \mathbf{w}_{it-1} - r(\beta L\mathbf{x}_i + \lambda \mathbf{w}_{it-1})$
  end if
  $\Delta y_j \leftarrow |\mathbf{w}L\mathbf{x}_j - y_j|$
  if $\Delta y_j > \epsilon$ **then**
    $L_{it} \leftarrow L_{it-1} - \beta r \mathbf{w}_{it-1}\mathbf{x}_j^\top$
    $\mathbf{w}_{it} \leftarrow \mathbf{w}_{it-1} - r(\beta L\mathbf{x}_j + \lambda \mathbf{w}_{it-1})$
  end if
  $D^2_L(\mathbf{x}_i, \mathbf{x}_j) \leftarrow \|L\mathbf{x}_i - L\mathbf{x}_j\|^2$
  if $y_{ij}(1 - D^2_L(\mathbf{x}_i, \mathbf{x}_j)) < 0.2$ **then**
    $L_{it} \leftarrow L_{(it-1)} - \gamma r y_{ij} L_{(it-1)}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top$
  end if
end while

using stochastic gradient descent. In this equation, m and b are called margin and bias respectively and are free parameters. We generate the pairwise constraints from the large number of examples from source domain and a limited number of examples from the target domain. This is similar to the approach of Saenko et al. [17], who use ML for cross-domain image classification. It is important to note here that, the pairs they generated were from the examples belonging to two different domains. In [17], after learning projection matrix, training examples are projected into this subspace and classifier is trained in this subspace.

2.2. Proposed joint learning for cross-domain regression

An immediate extension of the approach of Saenko et al. [17] for regression could be similar ML projection followed by regressor learning. The problem with such approach is that it would not directly address the main goal of minimizing the absolute age difference between the ground truth age and predicted age. Moreover, pairwise constraints try to bring images belonging to same age categories together but push away the images belonging to different age categories. They push dissimilar pair away equally i.e. without taking into consideration the difference in their ages. For example, two pairs of images with the ages (25, 26) and (25, 55) are equally pushed apart. Unlike classification tasks, it is important to address this issue in regression tasks. Incorporating the regressor while learning projection matrix address this problem by pushing the ages with lesser difference comparatively less

farther.

We are thus interested in learning a projection L and a regressor $\mathbf{w}$, in the resulting space, jointly. We propose to minimize the following objective for learning $\mathbf{w}, L$,

$$\min_{L,\mathbf{w}} \mathcal{L}(\mathcal{T}, \mathcal{S}, \mathcal{D}; L, \mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \beta \sum_k \ell_{\mathbf{w}}(L\mathbf{x}_k, y_k) + \gamma \sum_{\mathcal{S}\cup\mathcal{D}} \ell_L(\mathbf{x}_i, \mathbf{x}_j, y_{ij}) \quad (5)$$

where, the first term is ℓ^2 regularization on $\mathbf{w}$, $\lambda, \beta, \gamma \in \mathbb{R}$ are free parameters controlling the relative contributions of the different terms, $\ell_{\mathbf{w}}$ is the support vector regression loss which aims to bring the predicted age within $\pm\epsilon \in \mathbb{R}^+$ of the true age, given by:

$$\ell_{\mathbf{w}}(L\mathbf{x}, y) = \max(0, |\mathbf{w}^\top L\mathbf{x} - y| - \epsilon) \quad (6)$$

where $\ell_L(\mathbf{x}_i, \mathbf{x}_j, y_{ij})$ is the loss which aims at bringing similar age pairs together while pushing dissimilar age pairs away from each other. In practice, we optimize the objective using a stochastic gradient based solver, which is detailed in Alg. 1.

3. EXPERIMENTS

Dataset. We use the largest publicly available dataset for age estimation, the MORPH-II dataset, to evaluate the proposed method. We followed the experimental setup of Guo et al. [12] and compared the performance of our method with their method. We computed Local Binary Patterns (LBP) [18] of face images instead of Biologically Inspired Features (BIF) [10] which they used for their experiments. The database contains around 55k images from different races ('Black', 'White', 'Caucasian', etc.) and genders ('Male', 'Female'). Similar to [12], we took randomly sampled subsets of the database for the experiments. We took images from two races 'Black', and 'White', and two genders 'Male', and 'Female'. This subset contains 2,570 White Female (WF), 7,960 White Male (WM), 2,570 Black Female (BF), and 7,960 Black Male (BM) face images. Each of these categories is called a domain. From each of these domains, 50% of randomly sampled images are used for training and validation purposes and the rest 50% are used for testing. We used SVM regressor for predicting ages. The performance is calculated by Mean Absolute Error (MAE). MAE is the mean of absolute difference between the ground truth age and the predicted age.

Face Description. We used Viola and Jones face detector [19] to compute the bounding boxes of faces. These bounding boxes were resized to the size of 250×250. We computed facial landmarks using publicly available state-of-art facial landmark detector [20][1]. With the help of these facial landmarks we align the faces if required. The aligned faces are then centre cropped into the size of 160×100. We then compute local binary patterns (LBP) for each of these images using the publicly available `vlfeat` [21] library. We set cell size is equal to 10 as parameter and obtain a signature for each of the images which are of 9280 dimensions. Note however, the proposed method can work with other types of features e.g. LQP [22], LHS [23] or Fisher Vectors [24].

[1] https://github.com/soundsilence/FaceAlignment

3.1. Baselines

As a first reference we used the full features without any projection learning and hence without any compression. In addition, we compared with the following competitive baselines.

Unsupervised compression. We used Whitened Principal Components (WPCA) to compress high dimensional LBP to 64 dimensions. For training and testing, these representations are very efficient but suboptimal, as they may remove some discriminative information for age prediction.

Supervised Compression with ML. We used ML to learn compact representation of images which retains some discriminative information. We initialized with WPCA and learned the projection with stochastic gradient descent. This approach not only samples features that are useful for age estimation, but also aligns the features between the source and target domains.

After compressing, and potentially aligning the domains, for all these baselines, we use the publicly available SVR from `scikit-learn` [25] to learn the model on projected features to predict the ages. For all the experiments reported, we chose a linear kernel. We split train set into two halves for cross-validation. We set $\epsilon = 0.1$ and select the C parameter for SVR by cross-validation.

3.2. Proposed joint approach

Joint Learning (JL) learns the regressor and projection in with an integrated objective function. The advantage of JL in comparison to ML is that it takes care of dissimilarity constraint between the ages. As mentioned before in the Section 2, ML pushes the dissimilar images equally farther irrespective of difference between the ages. We trained JL identically cf. ML; we used the same training pairs that were used for ML and initialized the projection matrix with WPCA and regressor by mean of the principal components of WPCA. Since we learned a projection matrix of dimensions 64, our regressor has 64 dimensions. The initial values of regressor are mean values of 64 principal components. We set learning rate to 0.001 and the number of maximum iterations to 2×10^5. For the regressor, we set $\epsilon = 0.1$, similar to that of standard SVR we used for all the baselines.

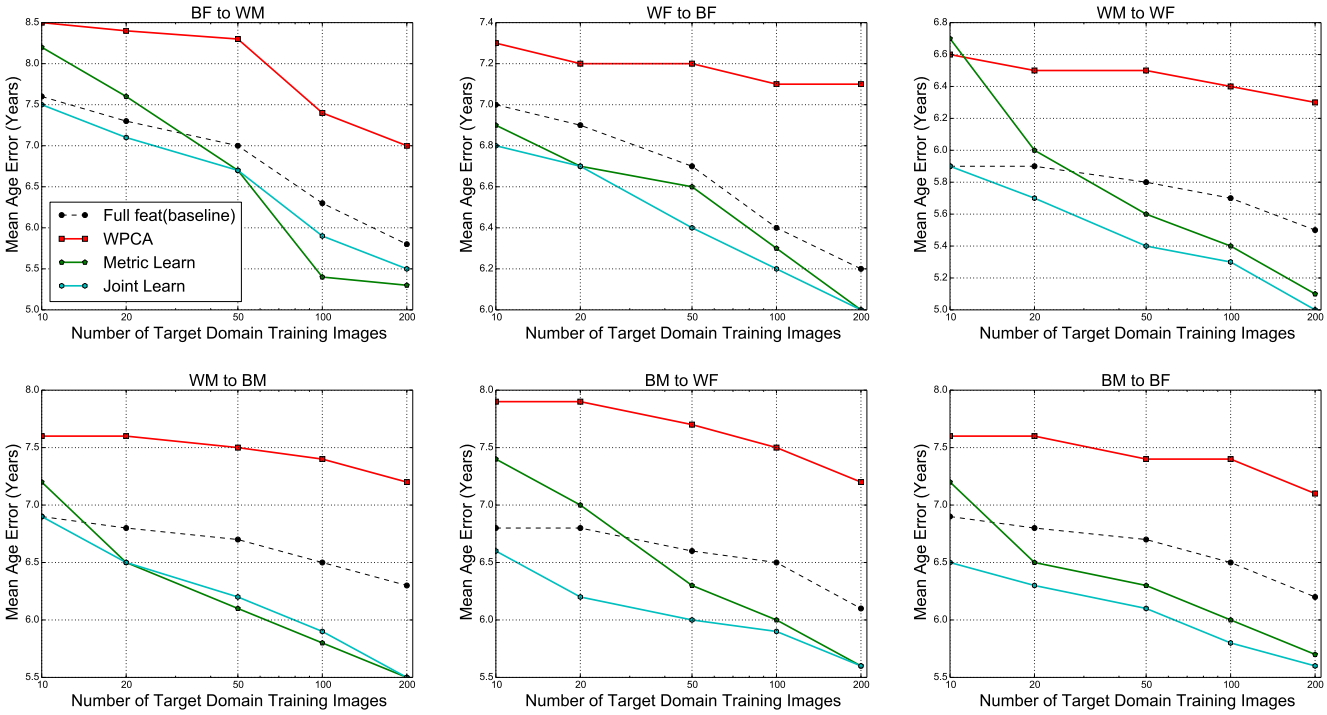

Fig. 2. Graphs showing performance of different approaches vs. the number of target training examples.

3.3. Experimental Results

Fig. 2 shows the performance of all the baselines and the one of our approach w.r.t. the size of the number of target training examples in 6 unique domain pairs. When we exchange the role of source and target of these 6 pairs, we get 12 domain pairs, which constitutes the total number domain pairs in our experiments. Tab. 1 shows the performances of our method along with those of the baselines and the current state-of-art method of Guo et al. [12]. The values in the table shows the Mean (over 12 domain pairs) of the MAE (mean average error over examples) in years in relation with the number of Target Training Examples (TTE) used. It usually requires a large number of labeled examples per class to compute scatter matrix using LDA, so we assume Guo et al. used more than 200 examples. In the domains, WF and BF, 200 examples counts around $(200/1285) \times 100 = 15.6\%$ and in WM, BM, it counts $(200/3980) \times 100 = 5\%$ of the training examples.

We note that, in comparison to the baselines i.e. LBP and WPCA, the proposed method consistently performs better. In comparison to ML, it performs better when the training examples from target domain is very small; whereas ML performs even worse than WPCA in such case (e.g. source target pair WM and WF). ML overfits when the positive training pairs are very small in number. This is an important practical use case, as often obtaining annotated examples of a new target domain is expensive. With the increasing size of target examples, the performance of ML ultimately converges to that of JL. Finally, the proposed approach clearly out-performs previous state-of-the-art method [12] by just taking 20 training examples from the target domain.

Method	LBP	WPCA	ML	JL
Dimensions	9280	64	64	64

TTE	Method	Mean of MAE (y)
>200	[12]	**6.6 ± 1.0**
0	LBP	6.8 ± 0.8
	WPCA	7.4 ± 0.7
10	LBP	6.8 ± 0.7
	WPCA	7.4 ± 0.7
	ML	7.2 ± 0.7
	JL	**6.7 ± 0.7**
20	LBP	6.7 ± 0.7
	WPCA	7.3 ± 0.7
	ML	6.7 ± 0.5
	JL	**6.5 ± 0.6**

TTE	Method	Mean of MAE (y)
50	LBP	6.5 ± 0.5
	WPCA	7.3 ± 0.7
	ML	6.2 ± 0.4
	JL	**6.1 ± 0.4**
100	LBP	6.2 ± 0.4
	WPCA	7.0 ± 0.6
	ML	**5.8 ± 0.4**
	JL	**5.8 ± 0.4**
200	LBP	5.9 ± 0.5
	WPCA	6.8 ± 0.6
	ML	**5.5 ± 0.4**
	JL	**5.5 ± 0.4**

Table 1. Performance comparison between different baselines, our approach and previous state-of-art method [12].

4. CONCLUSIONS

We propose a novel joint learning method for cross-domain age estimation. We have evaluated our method on the largest publicly available dataset. The proposed experimental validation shows that our method outperforms wide ranges of strong baselines, improves the performance over the previous state-of-art algorithm and attains a state-of-art performance.

5. REFERENCES

[1] Hu Han, Christina Otto, and Anubhav K Jain, "Age estimation from face images: Human vs. machine performance," in *ICB*, 2013, pp. 1–8.

[2] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy, "Cumulative attribute space for age and crowd density estimation," in *CVPR*, 2013, pp. 2467–2474.

[3] Zheng Song, Bingbing Ni, Dong Guo, Terence Sim, and Shuicheng Yan, "Learning universal multi-view age estimator using video context," in *ICCV*, 2011, pp. 241–248.

[4] Pavleen Thukral, Kaushik Mitra, and Rama Chellappa, "A hierarchical approach for human age estimation," in *ICASSP*, 2012, pp. 1529–1532.

[5] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung, "A ranking approach for human ages estimation based on face images," in *ICPR*, 2010, pp. 3396–3399.

[6] Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles, "Automatic age estimation based on facial aging patterns," *PAMI*, vol. 29, no. 12, pp. 2234–2240, 2007.

[7] Caifeng Shan, Fatih Porikli, Tao Xiang, and Shaogang Gong, *Video Analytics for Business Intelligence*, vol. 409, Springer, 2012.

[8] Guodong Guo and Guowang Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," in *CVPR*, 2011, pp. 657–664.

[9] Guodong Guo, Guowang Mu, Yun Fu, Charles Dyer, and Thomas Huang, "A study on automatic age estimation using a large database," in *ICCV*, 2009, pp. 1986–1991.

[10] Guodong Guo, Guowang Mu, Yun Fu, and Thomas S Huang, "Human age estimation using bio-inspired features," in *CVPR*, 2009, pp. 112–119.

[11] Guodong Guo and Guowang Mu, "Human age estimation: What is the influence across race and gender?," in *CVPR Workshops*, 2010, pp. 71–78.

[12] Guodong Guo and Chao Zhang, "A study on cross-population age estimation," in *CVPR*, 2014, pp. 4257–4263.

[13] Fares Alnajar, Zhongyu Lou, Jose Alvarez, and Theo Gevers, "Expression-invariant age estimation," in *BMVC*, 2014.

[14] Alexis Mignon and Frédéric Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *CVPR*, 2012.

[15] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid, "Is that you? metric learning approaches for face identification," in *ICCV*, 2009, pp. 498–505.

[16] Binod Bhattarai, Gaurav Sharma, Frederic Jurie, and Patrick Pérez, "Some faces are more equal than others: Hierarchical organization for accurate and efficient large-scale identity-based face retrieval," in *ECCV Workshops*, 2014, pp. 160–172.

[17] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell, "Adapting visual category models to new domains," in *ECCV*. 2010, pp. 213–226, Springer.

[18] Matti Pietikäinen, Abdenour Hadid, Guoying Zhao, and Timo Ahonen, *Computer vision using local binary patterns*, vol. 40, Springer, 2011.

[19] Paul Viola and Michael J Jones, "Robust real-time face detection," *IJCV*, vol. 57, no. 2, pp. 137–154, 2004.

[20] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun, "Face alignment by explicit shape regression," *IJCV*, vol. 107, no. 2, pp. 177–190, 2014.

[21] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[22] Sibt Ul Hussain, Thibault Napoléon, Fréderic Jurie, et al., "Face recognition using local quantized patterns," in *BMVC*, 2012.

[23] G. Sharma, S. ul Hussain, and F. Jurie, "Local higher-order statistics (LHS) for texture categorization and facial analysis," in *ECCV*, 2012.

[24] Karen Simonyan, Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Fisher vector faces in the wild," in *BMVC*, 2013.

[25] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., "Scikit-learn: Machine learning in python," *JMLR*, vol. 12, pp. 2825–2830, 2011.