A Deep Joint Learning Approach for Age Invariant Face Verification

Ya Li[1,2], Guangrun Wang[2], Liang Lin[2(✉)], and Huiyou Chang[2]

[1] Guangzhou University, Guangzhou 510006, China
[2] Sun Yat-sen University, Guangzhou 510006, China
liya@gzhu.edu.cn, wanggrun@mail2.sysu.edu.cn,
linliang@ieee.org, isschy@mail.sysu.edu.cn

Abstract. Age-related research has become an attractive topic in recent years due to its wide range of application scenarios. In spite of the great advancement in face related works in recent years, face recognition across ages is still a challenging problem. In this paper, we propose a new deep Convolutional Neural Network (CNN) model for age-invariant face verification, which can learn features, distance metrics and threshold simultaneously. We also introduce two tricks to overcome insufficient memory capacity issue and to reduce computational cost. Experimental results show our method outperforms other state-of-the-art methods on MORPH-II database, which improves the rank-1 recognition rate from the current best performance 92.80% to 93.6%.

Keywords: Face verification · Age invariant · Face recognition · Deep CNN · Joint learning

1 Introduction

Age-related research has become an attractive topic in recent years due to its wide range of application scenarios. Age information is useful in many applications, such as age-specific human-computer interaction, security surveillance monitoring, age-based face images retrieval, automatic face simulation and intelligent advertisement system etc..

In spite of the great advancement in face related works in recent years, face recognition across ages is still a challenging problem. In paper [1], face verification achieved near-human performance on Labeled Faces in the Wild (LFW) dataset using high-dimensional Local Binary Pattern feature (HD-LBP). The work in paper [2,3] even achieved exceed-human ability on face verification using deep learning method. To our best knowledge, there is no such good result on age invariant recognition.

The challenges on age invariant recognition include large intra-subject variations and great inter-subject similarity. As well known, human face appearance will change greatly with the aging process. The changes are different in the different age period as shown in Fig.1(a). From birth to adulthood, the greatest change is the craniofacial growth, that is shape change; and from adulthood to

H. Zha et al. (Eds.): CCCV 2015, Part I, CCIS 546, pp. 296–305, 2015.
DOI: 10.1007/978-3-662-48558-3_30

old age, the most perceptible change becomes skin aging, that is texture change [4]. These changes of same person are the intra-subject variations. Meanwhile, different persons on same age period maybe look like same, that is the inter-subject similarity as shown in Fig.1(b). Therefore, enlarging the inter-subject

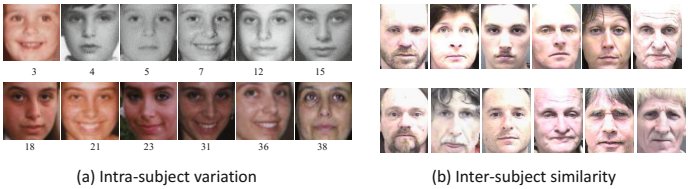

Fig. 1. Example images showing intra-subject variations and the inter-subject similarity. (a) Shows face appearance changes with the aging process. Images come from the FG-NET database [10]. (b)Shows different persons on same age period maybe look like same.

differences while reducing the intra-subject variations is a crucial goal in age invariant recognition as well as face recognition. Many approaches are realized based on this goal in the traditional face recognition such as Linear Discriminant Analysis (LDA) [5], Bayesian face [6,7] and metric learning [8]. However, these models are limited by their linear nature. Of course, many recent studies have made improvements to address these limits. For example, in order to measure similarity between images traditional metric learning methods require a fixed distance threshold. Li et al. [9] proposed to learn a decision function for face matching problem that could be looked as a joint model of a distance metric and a threshold locally adapted rule. But this model is limited by its shallow structures.

In this work, we train a deep convolutional network to learn features, distance metric and threshold function simultaneously. We aim at not only preserving similarity of the same person across ages while discriminating the different individuals, but also learning implicit adaptive thresholds at the same time. Generally it requires positive semi-definite for the Mahalanobis metric, so directly optimizing the metric matrix is computational intensive. Inspired by [9], we learn Mahalanobis metric and distance thresholds jointly, and further factorize the matrix into a fully-connected layer on the top of our deep architecture. In this way, the distance metric and distance thresholds is seamlessly integrated with the image feature represented by the other layers of neural networks. The joint optimization can be then efficiently achieved via the standard backward propagation. Therefore, by means of the nonlinear learning of deep neural networks, we improve the previous models and achieve the state-of-the-art result in age invariant verification.

There are two tricks to overcome insufficient memory capacity issue and to reduce computational cost. Considering the capacity of memory, we scale up our approach to the large amount of training data by using a batch-process strategy.

In each round of training, we randomly select a relatively small number of images, and use them to organize the training pairs. By taking the training pairs as the inputs, we update the model parameters by the stochastic gradient descent (SGD) algorithm and use backpropagation algorithm to compute gradient. In order to reduce the computational cost, we calculate the gradients on the images instead of the proceeded image pairs. Thus we can avoid computing the gradients repeatedly, because one image can be included into more than one training pairs. So the computational cost is reduced by only depending on the number of the selected images.

The reminder of this paper is organized as follows. In the next section, we first review related works in age invariant recognition. Then in Section 3, we introduce our deep joint metric learning method and the network framework. And we present our experiment in Section 4. Finally, Section 5 gives some conclusions.

2 Related Work

In the literature, most age-related works focus on age estimation before [10–18] , including exact age estimation and age group estimation. Lanitis et al. [10] first present an exact age estimation method using the statistical face model which established according to the aging function. Geng et al. [11] propose an algorithm named AGES to learn aging pattern subspace which can reconstruct the face image missing in the training samples. Fu et al. [12] involve the manifold ways for age estimation and compared experiment results by several manifold methods. Guo et al. [13] propose using biological inspired features (BIF) and principal component analysis (PCA) dimension reduction for facial image description, and they use support vector machine (SVM) classification method for age estimation. In paper [15], Guo et al. focus on cross-data age estimation and introduce a "correlation" item to measure two different populations' correlation and project two different aging patterns into a common space. More recently, Geng et al. [16] extend their label distribution learning algorithms [14] and propose two adaptive label distribution learning algorithms IIS-ALDL and BFGS-ALDL which can learn the parameters of label distribution adaptively. Li et al. [17] propose a novel hierarchical feature composition and selection model used in facial age estimation. To the best of our knowledge, until now there is only one paper in literature using deep model for age estimation. In [18], Yan et al. use deep Convolutional Neural Network (CNN) to extract facial features and SVM classifier to estimate age groups.

Recent years, more and more works focus on age invariant recognition. In paper [19], Park et al. propose a 3D facial aging model and simulation method for age-invariant face recognition. This is a generative approach, which try to compensate the 3D facial images of lacked age before recognition. Ling et al. [20] combine gradient orientation pyramid (GOP) with SVM for face verification. For an image pair they used the cosine distances between each one's multi scale gradient orientation as the feature vector, and then they use SVM for classification. In [21], Li et al. propose a new method to used in age invariant recognition, which

is a variation of random subspace LDA. What's more, they represent each face using two patch-based local feature descriptors SIFT and LBP, so their method is named multi-feature discriminant analysis (MFDA). Based on the observation: different person usually share common characteristics and the same person contain intrinsic features which are relatively invariant across ages, Gong et al. [22] express a facial image using age component, identity component and a noise term. And they adopt Expectation Maximization (EM) algorithm to estimate this generative model parameters. More recently, Chen et al. [23] suppose that if two young persons look alike, it is likely that they also look similar when they are old. And they propose a data-driven approach called cross-age reference coding (CARC) for age invariant recognition.

3 Deep Joint Metric Learning Framework

we train a deep convolutional network to learn features, distance metrics and threshold simultaneously. In this section, we first introduce the classical metric learning method and the optimization objective of our model. Then we present our deep model. Finally, we show how to use our method for age invariant face recognition.

3.1 Optimization Objective

Following the early work of Xing et al. [24], for pairwise images (x, y), most distance metrics learning approaches learn a Mahalanobis-like distance: $d(x, y) = (x - y)^t M(x - y)$, where M is a positive semi-definite (PSD) matrix. Suppose category labels of pairwise images are $c(x)$ and $c(y)$. x, y are in the same class or are similar, if $c(x) = c(y)$. A simple way is minimum the distance between samples in the same class.

If metric learning is used in matching problem, it requires a threshold to decide whether x and y are matched. We formulate it as following:

$$(x - y)^t M(x - y) \leq d, \quad M \succeq 0. \tag{1}$$

However, it is inappropriate for a fixed threshold d. Because maybe the distance of intra-subject is lager than the distance of inter-subject. Li et al. [9] propose to learn threshold adaptively, d is a function related with (x, y) instead of a constant, so inequation (1) becomes to $(x - y)^T M(x - y) \leq d(x, y)$. Thus, the decision function $f(x, y)$ can be written as:

$$f(x, y) = d(x, y) - (x - y)^t M(x - y) \begin{cases} \geq 0 & \text{if } c(x) = c(y) \\ < 0 & \text{otherwise} \end{cases}. \tag{2}$$

Since the metric M itself is quadratic, we assume $d(x, y)$ as a simple quadratic form, i.e.,

$$d(x, y) = \frac{1}{2} x^t \widetilde{A} x + \frac{1}{2} y^t \widetilde{A} y + x^t \widetilde{B} y + c^t (x + y) + b. \tag{3}$$

Substitute Eq.(3) in Eq.(2), we get:

$$\begin{aligned} f(x,y) &= \frac{1}{2}x^t(\widetilde{A}-2M)x + \frac{1}{2}y^t(\widetilde{A}-2M)y + x^t(\widetilde{B}+2M)y + c^t(x+y) + b \\ &= \frac{1}{2}x^tAx + \frac{1}{2}y^tAy + x^tBy + c^t(x+y) + b, \end{aligned} \tag{4}$$

where $A = (\widetilde{A} - 2M)$ and $B = (\widetilde{B} + 2M)$. Suppose A is PSD and B is negative semi-definite (NSD), A and B can be factorized as $L_A^T L_A$ and $L_B^T L_B$. Eq.(4) can be further written as:

$$\begin{aligned} f(x,y) &= \frac{1}{2}x^tL_A^TL_Ax + \frac{1}{2}y^tL_A^TL_Ay - x^tL_B^TL_By + c^t(x+y) + b \\ &= \frac{1}{2}(L_Ax)^t(L_Ax) + \frac{1}{2}(L_Ay)^t(L_Ay) - (L_Bx)^t(L_By) + c^tx + c^ty + b \end{aligned} \tag{5}$$

Through above transformation, face recognition is cast into computing the decision function (5). For an age instance z of person P, we wish to learn a reidentification model to successfully identify another age instance z' of the same person. This can be achieved by learning metrics L_A, L_B and vector c, subject to the value of $f(z, z')$ as large as possible for same person while as small as possible for different person. Given a training set $Z = \{(z_i, y_i)\}_{i=1}^N$, where y_i is the class label, i.e. person ID, N is the total person number, we define a pairwise set $\Omega = \{\Omega_k = (z_i, z_j)\}$. Taking pairwise set Ω as input, we can maximize the sum of $l(\Omega_k) \times f(\Omega_k)$, where $l(\Omega_k)$ is the label of image pair Ω_k, if z_i and z_j come from the same person, $l(\Omega_k) = -1$, otherwise, $l(\Omega_k) = 1$. So our objective is maximize the sum of $l(\Omega_k) \times f(\Omega_k)$. Further more, we ignore the pairs that $f(\Omega_k) > 1$, here the choice of the constant 1 is arbitrary but not important, and changing it to any other positive constant c results only in the matrices being replaced by c times. We denote $(L_A, L_B, c)^T$ as W, our hinge-loss like objective function is:

$$H(W) = \sum_{\Omega} \max\{0, l(\Omega_k) \times f(\Omega_k) + 1\} \quad k = 1, 2...N^2 \tag{6}$$

3.2 Deep Architecture

L_A, L_B and c discussed in previous section can be looked as the weight of the fully-connected layer in deep CNN. Using one CNN network we realize the feature extracting, metric and threshold learning simultaneously. The parameters of the three components can be obtained by using network propagation algorithms.

As the publicly available datasets with person age information are relatively small, we use a relatively small network for our model. Fig. 2 shows the overall network architecture, which contains 7 layers. The first layer is a convolutional layer including 32 kernels of size $5 \times 5 \times 3$ with a stride of 2 pixels. The second layer is a max-pooling layer. The third layer is also a convolutional layer taking the max-pooling output as input and filters it with 32 kernels of size $5 \times 5 \times 16$ with a stride of 1 pixel. The fourth layer is a max-pooling layer followed by three

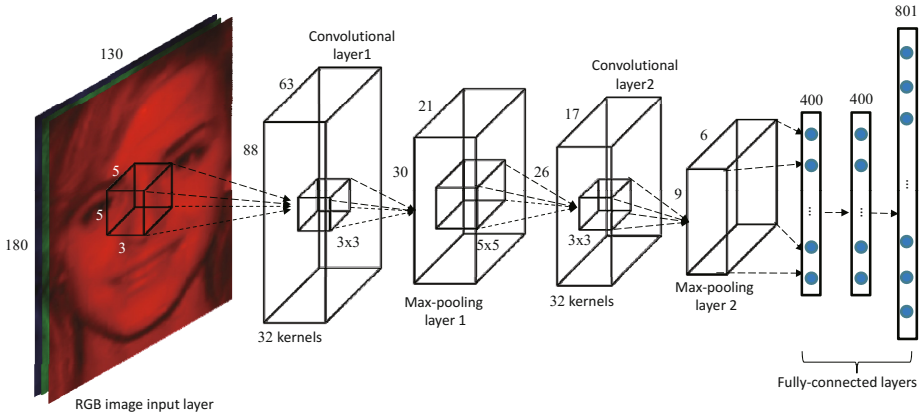

Fig. 2. Architecture of our model. The first and the third layers are convolutional layers, the second and the fourth are max-pooling layers. The last three layers are fully-connected layers.

fully-connected layers including two 400 dimension fully-connected layers and one 801 dimension fully-connected layer. We use rectified linear units (ReLU) for neurons in the convolutional layers. In fact, we can look feature extraction stage extracts 400 dimensional features by stacking two convolution-pooling layers and a fully-connected layer. The second fully-connected layer takes the role of projecting features to a common space. And the metric and threshold learning stage is realized by the third fully-connected layer with 801 dimension outputs, which learns the matrices L_A, L_B and vector c simultaneously.

3.3 Learning Algorithm

Batch Process. We apply batch learning strategy to optimize the parameters, due to the limited memory. For N training images, the pairs number is $O(N^2)$. Even for a moderate dataset, it is impossible to load all image pairs into memory to train the network. How to choose pairs used in batch process? A simple way is to generate pairs randomly. However, this method makes the negative pairs are far more than positive pairs, because the person number be approximately double the pairs number and the likelihood of two pairs sharing the same person is very small.

To solve this problem we propose pairs generation rule as follows. In each iteration, we select a fixed number of persons(classes), and generate image pairs only using these persons. In order to ensure the training samples without loss of generality, we randomly select samples according to one ratio of positive and negative samples.

Parameter Optimization. As well know, the key step of CNN is updating the model parameters by the stochastic gradient descent (SGD) algorithm.

A straight method is to calculate the gradient for each pair according to the loss function and sum these gradients to get the overall gradient. Since each pair contains two images, it will incur twice network propagation in this approach. It is inefficient without sharing the propagation of the same image in different pairs. In fact, we can optimize this process by computing the gradient of images rather than image pairs.

The objective function Eq.(6) can be written as following form:

$$H(W) = \sum_{i,j \in m} loss(F_W(I_i), F_W(I_j)), \tag{7}$$

where $< I_i, I_j >$ is the image pair. Based on the batch processing strategy, Eq.(7) can be further written as:

$$H(W) = H(F_W(I_1), F_W(I_2), ..., F_W(I_m)), \tag{8}$$

where $\{I_i\}$ represents the set of all the distinct images in the pairs, and m denotes the number of the distinct images. Compute the desired partial derivatives, which are given as:

$$\frac{\partial H}{\partial W^l} = \sum_{i=1}^{m} \frac{\partial H}{\partial {X_i}^l} \cdot \frac{\partial {X_i}^l}{\partial W^l}, \tag{9}$$

$$\frac{\partial H}{\partial {X_i}^l} = \frac{\partial H}{\partial {X_i}^{l+1}} \cdot \frac{\partial {X_i}^{l+1}}{\partial {X_i}^l}, \tag{10}$$

where W represents the network parameters, X_i^l represents the feature maps of the image I_i at the l^{th} layer. The right item of Eq.(9) is the sum of $\frac{\partial H}{\partial {X_i}^l} \cdot \frac{\partial {X_i}^l}{\partial \mathbf{W}^l}$, it is denoted as $\frac{\partial H}{\partial \mathbf{W}^l}({I_i}')$ shortly.

The Eq.(9) shows that the overall gradient is the sum of the image-based gradient. The Eq.(10) shows that the partial derivative of each image respect to the feature maps can be calculated recursively. So the gradients of network parameters can be obtained by back propagation algorithm. Algorithm 1 shows the detail.

4 Experiment

The experiment is carried on MORPH-II database [25]. MORPH-II contains more than 55,000 face images of more than 13,000 individuals and ages range from 16 to 77. The average number of images per individual is 4. The individuals come from different races, among them Africans' images accounted for about 77%, the European images about 19%, and the remaining includes Hispanic, Asian and other races. The training data consists of 20000 face images from 10000 subjects, with each subject having two images with the largest age gap. The test data is composed of a gallery set and a probe set from the remaining 3000 subjects. The gallery set is composed of the youngest face images of

Algorithm 1. Deep joint learning algorithm

Input:
Training set X = $\{(x_i, l_i)\}$, initialized parameters W, learning rate $\alpha(t)$, $t \leftarrow 0$
Output:
Network parameters W

while $t < T$ **do**
  $t \leftarrow t + 1$
  Sample training persons randomly from X
  Sample pairwise training set Ω_k from $\{< I_i, I_j >\}$
  for all $\{I_i\}$ **do**
    Calculate the whole network's output $F(I_i)$ and each layer's feature maps X_i by forward propagation
  end for
  for all $\{I_i\}$ **do**
    $\frac{\partial H}{\partial F_W(I_i)} = 0$
    for all image pair Ω_k i.e. $< I_p, I_q >$ **do**
      if $f(\Omega_k) > 1$ **then**
        if $I_i = I_p$ **then**
          $\frac{\partial H}{\partial F_W(I_i)} += \frac{\partial H}{\partial F_W(I_p)}$
        else if $I_i = I_q$ **then**
          $\frac{\partial H}{\partial F_W(I_i)} += \frac{\partial H}{\partial F_W(I_q)}$
        end if
      end if
    end for
    Calculate $\frac{\partial H}{\partial W}(I_i{}')$ using back propagation (Eq.(9) and Eq.(10))
    Sum the partial derivative $\Delta W = \Delta W + \frac{\partial H}{\partial W}(I_i{}')$
  end for
  $W^t = W^{t-1} - \alpha_t \Delta W$
end while

Table 1. Rank-1 identification rates on the MORPH database. Our method achieves the highest recognition rate compared to other state-of-the-art methods

Method	Recognition rate
Park et al. [19]	79.8%
MFDA [21]	83.9%
HFA [22]	91.1%
CARC [23]	92.8%
Ours	**93.6%**

each subject. The probe set is composed of the oldest face images of each subject. This experimental setting is same with [23] and [22].

We compare our deep CNN model against several state-of-the-art methods for age invariant face recognition on MORPH-II, including CARC [23], HFA [22],

Fig. 3. Some examples of rank-1 failed retrievals. The first row are the probe images, the second row is the rank-1 result of our method, and the third row is the ground-truth, i.e. correct matched image in the gallery.

MFDA [21] and method proposed in paper [19]. The comparative results are reported in Table 1. It is encouraging to see that our approach significantly outperforms the current best-performance method CARC by improving the rank-1 identification rate from 92.80% to 93.60%. To our best knowledge, this is the best identification rank-1 result on MORPH-II. For top-10 and top-20, our model achieves 98.8% and 99.34% respectively. Finally, we show some examples of rank-1 failed retrievals in Fig. 3. In spite of the rank-1 retrievals are incorrect in these cases, we can find that the probe images are looked be more similar to the incorrect rank-1 matched images than the true images.

5 Conclusion

In this paper, we propose a new deep CNN model for age-invariant face recognition, which can learn features, distance metrics and threshold simultaneously. Experimental results show our method outperforms other state-of-the-art methods on MORPH-II database. We also introduce two tricks to overcome insufficient memory capacity issue and to reduce computational cost. In the future, we want to investigate other facial attributes recognition, like expression, gender, ethnicity and head pose etc..

Acknowledgments. This research is supported by the National High Technology Research and Development Program of China (No.2013AA013801), the Science and Technology Planning Project of Guangdong Province(No. 2013B010406005) and the Guangdong Natural Science Foundation (No. S2013040012570). The authors would like to thank the reviewers for their comments and suggestions.

References

1. Chen, D., Cao, X., Wen, F., Sun, J.: Blessing of dimensionality: high-dimensional feature and its efficient compression for face verification. In: CVPR (2013)
2. Lu, C., Tang, X.: Surpassing human-level face verification performance on LFW with GaussianFace. ArXiv e-prints (2014)

3. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: NIPS (2014)
4. Fu, Y., Guo, G., Huang, T.S.: Age synthesis and estimation via faces: A survey. TPAMI **32**(11), 1955–1976 (2010)
5. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. TPAMI **19**(7), 711–720 (1997)
6. Moghaddam, B., Jebara, T., Pentland, A.: Bayesian face recognition. Pattern Recognition **33**, 1771–1782 (2000)
7. Chen, D., Cao, X., Wang, L., Wen, F., Sun, J.: Bayesian face revisited: a joint formulation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 566–579. Springer, Heidelberg (2012)
8. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? Metric learning approaches for face identification. In: ICCV (2009)
9. Li, Z., Chang, S., Liang, F., Huang, T.S., Cao, L., Smith, J.R.: Learning locally-adaptive decision functions for person verification. In: CVPR (2013)
10. Lanitis, A., Taylor, C.J., Cootes, T.: Toward automatic simulation of aging effects on face images. TPAMI **24**(4), 442–455 (2002)
11. Geng, X., Zhou, Z.H., Smith-Miles, K.: Automatic age estimation based on facial aging patterns. TPAMI **29**(12), 2234–2240 (2007)
12. Fu, Y., Huang, T.S.: Human age estimation with regression on discriminative aging manifold. TMM **10**(4), 578–584 (2008)
13. Guo, G., Mu, G., Fu, Y., Huang, T.: Human age estimation using bio-inspired features. In: CVPR (2009)
14. Geng, X., Yin, C., Zhou, Z.H.: Facial age estimation by learning from label distributions. TPAMI **35**(10), 2401–2412 (2013)
15. Guo, G., Zhang, C.: A study on cross-population age estimation. In: CVPR (2014)
16. Geng, X., Wang, Q., Xia, Y.: Facial age estimation by adaptive label distribution learning. In: ICPR (2014)
17. Li, Y., Peng, Z., Liang, D., Chang, H. and Cai, Z.: Facial age estimation by using stacked feature composition and selection. The Visual Computer, 1–12 (2015)
18. Yan, C., Lang, C., Wang, T., Du, X., Zhang, C.: Age estimation based on convolutional neural network. In: Ooi, W.T., Snoek, C.G.M., Tan, H.K., Ho, C.-K., Huet, B., Ngo, C.-W. (eds.) PCM 2014. LNCS, vol. 8879, pp. 211–220. Springer, Heidelberg (2014)
19. Park, U., Tong, Y., Jain, A.K.: Age-invariant face recognition. TPAMI **32**(5), 947–954 (2010)
20. Ling, H., Soatto, S., Ramanathan, N., Jacobs, D.W.: Face verification across age progression using discriminative methods. TIFS **5**(1), 82–91 (2010)
21. Li, Z., Park, U., Jain, A.K.: A discriminative model for age invariant face recognition. TIFS **6**(3–2), 1028–1037 (2011)
22. Gong, D., Li, Z., Lin, D., Liu, J., Tang, X.: Hidden factor analysis for age invariant face recognition. In: ICCV (2013)
23. Chen, B.-C., Chen, C.-S., Hsu, W.H.: Cross-age reference coding for age-invariant face recognition and retrieval. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VI. LNCS, vol. 8694, pp. 768–783. Springer, Heidelberg (2014)
24. Xing, E.P., Jordan, M.I., Russell, S., Ng, A.Y.: Distance metric learning with application to clustering with side-information. In: NIPS (2002)
25. Ricanek, K., Tesafaye, T.: Morph: a longitudinal image database of normal adult age-progression. In: FG (2006)