Age regression from soft aligned face images using low computational resources

Juan Bekios-Calfa[1], José M. Buenaposada[2], and Luis Baumela[3]

[1] Dept. de Ingeniería de Sistemas y Computación, Universidad Católica del Norte
Av. Angamos 0610, Antofagasta, Chile
juan.bekios@ucn.cl
[2] Dept. de Ciencias de la Computación, Universidad Rey Juan Carlos
Calle Tulipán s/n, 28933, Móstoles, Spain
josemiguel.buenaposada@urjc.es
[3] Dept. de Inteligencia Artificial, Universidad Politécnica de Madrid
Campus Montegancedo s/n, 28660 Boadilla del Monte, Spain
lbaumela@fi.upm.es

Abstract. The initial step in most facial age estimation systems consists of accurately aligning a model to the output of a face detector (e.g. an Active Appearance Model). This fitting process is very expensive in terms of computational resources and prone to get stuck in local minima. This makes it impractical for analysing faces in resource limited computing devices. In this paper we build a face age regressor that is able to work directly on faces cropped using a state-of-the-art face detector. Our procedure uses K nearest neighbours (K-NN) regression with a metric based on a properly tuned Fisher Linear Discriminant Analysis (LDA) projection matrix. On FG-NET we achieve a state-of-the-art Mean Absolute Error (MAE) of 5.72 years with manually aligned faces. Using face images cropped by a face detector we get a MAE of 6.87 years in the same database. Moreover, most of the algorithms presented in the literature have been evaluated on single database experiments and therefore, they report optimistically biased results. In our cross-database experiments we get a MAE of roughly 12 years, which would be the expected performance in a real world application.

1 Introduction

Age is a demographic variable that can be estimated using visual cues such as facial appearance, gait, clothing or hair style and non-visual cues like the voice. Automatic age estimation has interesting applications to enforce legal age restrictions in vending machines, automate marketing studies in shopping centres, measure tv audience or recognise faces automatically from videos. The aim of this paper is to use facial appearance as a visual cue to estimate the age of a person.

The facial age estimation problem is difficult since we are trying to estimate the real age from the face appearance, which depends on environmental

conditions like health, eating habits, sun exposure record, etc [13]. Facial age estimation can be seen either as a classification problem (i.e. different age groups or ranges) or a regression problem.

The state-of-the-art on age estimation can be organised into hard aligned (AAMs or manually) results and soft aligned results. There are two key references in the hard aligned group: the Bio-inspired Features (BIF) [5] and the Regression from Patch Kernel (RPK) [12]. The BIF approach uses a bank of Gabor filters at different scales and orientations with a combination layer and a PCA reduction step over manually aligned faces of 60×60 pixels. Although the result is 4.77 years of MAE in leave-one-person-out cross-validation, the best reported so far, the computational requirements of the method are quite high. The RPK approach breaks the 32×32 pixels input image into equally sized patches (8×8 pixels each). Then each patch is described by Discrete Cosine Transform (DCT) and the position in the image plane is added to the descriptor. The probability distribution of the patch descriptors within an image is modelled by a mixture of Gaussians and the age is finally estimated by Kernel Regression [12]. This approach achieves a MAE of 4.95 years on FG-NET, with standard leave-one-subject-out cross-validation.

Concerning soft aligned results, [6] performs training and testing directly on the output of the face detector. They extract Histogram of Oriented Gradients (HoG), Locally Binary Patterns (LBP) and local intensity differences from local patches in a regular image grid. The regressor is based on a Random Forest trained with 250 randomly selected images from FG-NET. They achieve a MAE of 7.54 years. Their result is optimistically biased since the same subject may be in the training and testing sets. In [3], they use semi-supervised learning using web queries, multiple face detectors and robust multiple instance learning. They use DCT local image descriptors and a fully automated pipeline from database collection to age regression estimation. The main limitation of this approach for a resource limited device is its computational complexity.

An important issue to consider is whether it is worth using computationally intensive face alignment procedures rather than learning to estimate face age with unaligned images. Most face age estimation results use Active Appearance Models (AAMs) for face alignment [13]. Unfortunately, fitting an AAM to unseen faces is prone to get stuck in local minima [10]. Moreover, fitting an AAM can be a computationally prohibitive task when there are many faces in the image or when the computation is performed on a resource limited device, such as a smart phone or an IP camera. An alternative here is using soft aligned algorithms, which require no accurate alignment to the input face image [3, 6].

In this paper we follow a soft alignment approach and train our regressor with cropped faces obtained from a face detector. We use K-NN regression for age estimation using a learned metric. Our metric is derived from Fisher Linear Discriminant Analysis. By computing the LDA projection matrix using age groups we impose that similar aged faces be close to each other and far apart from different aged ones. With this approach we can get roughly state-of-the-art age estimation. By dealing with the misalignment during training, the on-line

classification algorithm is quite simple and efficient. We train our algorithms in one database and test them in a different one (see section 3.2). With leave-one-person-out cross-validation in FG-NET we get a MAE of 5. With cross-database tests we achieve a MAE of 12 years, which is a more realistic value for a real application.

2 Age regression from face images

We use a non-linear regressor based on K-NN for age estimation. Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^{M}$ be $p^2 \times 1$ vectors where each $\mathbf{x}_i$ corresponds to the gray levels of a $p \times p$ pixels image scanned by columns and y_i is the age label corresponding to $\mathbf{x}_i$. The euclidean distance in the image space does not take into account the age. This means that with euclidean distance two face image vectors with different age labels could have lower distance than two face images with similar age. Therefore, we use a Mahalanobis-like distance with a learned metric matrix $\mathtt{M}$ to have similar aged face images close to each other and dissimilar aged face images far apart, $d_{\mathtt{M}}(\mathbf{x}_i, \mathbf{x}_j) = ||\mathbf{x}_i - \mathbf{x}_j||^2_{\mathtt{M}} = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathtt{M}(\mathbf{x}_i - \mathbf{x}_j)$. In the following subsections we explain how to learn the metric matrix $\mathtt{M}$ using Fisher Linear Discriminant Analysis and how we make K-NN age regression.

2.1 PCA+LDA projection as the age metric matrix

We use PCA+LDA (Linear Discriminant Analysis after a Principal Component Analysis projection) to compute a projection matrix $\mathtt{W}$. We compute $d_{\mathtt{M}}$ as

$$d_{\mathtt{M}}(\mathbf{x}_i, \mathbf{x}_j) = ||\mathtt{W}(\mathbf{x}_i - \mathbf{x}_j)||^2 = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathtt{W}^\top \mathtt{W}(\mathbf{x}_i - \mathbf{x}_j), \tag{1}$$

which means that the metric matrix is given by $\mathtt{M} = \mathtt{W}^\top \mathtt{W}$.

LDA is a supervised technique for dimensionality reduction that maximises the data separation of different classes. Since age is a continuous variable, to perform LDA first we have to discretise it into c age groups (see section 3 for the actual age groups we use). Given a multi-class problem with c classes and p sample points, $\{\mathbf{x}_i\}_{i=1}^{p}$the basis of the transformed subspace, $\{\mathbf{w}_i\}_{i=1}^{d}$, is obtained by maximising [4] $J(w) = \sum_{i=1}^{d} \frac{\mathbf{w}_i^\top \mathtt{S}_B \mathbf{w}_i}{\mathbf{w}_i^\top \mathtt{S}_m \mathbf{w}_i}$, where $\mathtt{S}_B$ and $\mathtt{S}_m$ are respectively the between-class and full scatter matrices.

Depending on the amount of training data, the performance of the regressor or classifier built on LDA subspace decreases when retaining all eigenvectors associated with non-zero eigenvalues. Thus, a crucial step here is to choose which PCA eigenvectors to keep so that no discriminant information is lost. We select the dimension of the subspace resulting from the PCA step using a cross-validation scheme instead of the usual approach based on retaining the eigenvectors accounting for a given percentage of the variance (usually 95% or 99%) [7]. To this end we sort PCA eigenvectors in descending eigenvalue order. We then perform cross-validation and select the dimension with the best performance for the age regression. This feature selection process is essential to correctly train a PCA+LDA procedure [2].

2.2 K-NN regression

We may interpret (1) as a projection of the face image onto the PCA+LDA subspace with the $\mathtt{W}$ matrix and then a classification in the transformed subspace using the euclidean metric. This is the approach we use in our K-NN regression implementation.

We project each training data vector, $\mathbf{x}_i$, onto the PCA+LDA subspace obtaining $\mathbf{z} = \mathtt{W}\mathbf{x}_i$. Once the optimal number of neighbours, K, is estimated by cross-validation, the regression output for a given input vector $\mathbf{z}$ in the PCA+LDA subspace is given by $\hat{y} = \sum_{i=1}^{K} \hat{w}_i y_i, \quad \hat{w}_i = \frac{w_i}{\sum_{j=1}^{K} w_j}, \quad w_i = \frac{1}{||\mathbf{z}-\mathbf{z}_i||}$, where y_i is the age label (real valued) of the i-th nearest neighbour, $\mathbf{z}_i$, to $\mathbf{z}$. When some or all of the distances are close to zero, or bellow a small threshold α (i.e. $||\mathbf{z} - \mathbf{z}_i|| <= \alpha = 10^{-6}$) we choose the label, y_i, of the nearest neighbour as the regression age label, $\hat{y} = y_i$.

3 Experiments

In this section we evaluate the performance of our age regressor and compare it with other age estimation approaches in the literature. We have used the Productive Aging Lab Face (PAL) database [9], the Images of Groups Dataset [1] and the FG-NET Aging database. To train our algorithm we estimate the number of nearest neighbours, K, and the PCA dimension optimising for the MAE in a cross-validation scheme. We crop and re-size images to a base size of 25×25 pixels using OpenCV's[4] 2.0.0 face detector, which is based on [11]. Then we equalise the histogram to gain some independence from illumination changes. Finally, we also apply an oval mask to prevent the background from influencing our results. Additionally, on FG-NET, we perform two kinds of manual alignment: 1) a similarity transform using the location of the eyes and 2) an affine transform using the location of the eyes and the centre of the mouth.

To train PCA+LDA we have discretised the age of FG-NET and PAL databases into 11 groups: 0-2, 3-7, 8-12, 13-19, 20-28, 29-37, 38-46, 47-55, 56-64, 65-73 and 74-82. On the other hand, the GROUPS database already comes with discrete age labels, which are organised in groups 0-2, 3-7, 8-12, 13-19, 20-36, 37-65, and 66+. In our experiments we use those face detections from the GROUPS database that have at least a size of 60×60 pixels (13,051 out of a total of 28,231).

Our measure for face age regression error is the Mean Absolute Error $MAE = \frac{1}{N} \sum_{i=1}^{M} |y_i - \hat{y}_i|$ where y_i is the actual label of a face image and $\hat{y}_i$ is the estimated age by a given algorithm. This is a non robust measure. To highlight outlier's influence in MAE a cumulative score curve shows the percentage of testing data below a given age estimation error (see Fig. 1). We use cumulative score curve to compare two age estimation algorithms, the higher the curve the better the algorithm.

[4] `http://opencv.willowgarage.com`

3.1 Intra-database tests

The first set of experiments use one database for training and testing.

In the FG-Net database case we perform leave-one-person-out cross-validation. In this way we avoid the bias introduced in the evaluation when classifying images of the same person both in the training and testing sets. This means that we keep all the images of a subject for testing (around 12), training with the rest.

To quantify the influence of alignment on age regression we compare raw face detection with manual alignment in FG-NET (see table 1). The difference in MAE between global affine transformation (using eyes and mouth) and a global similarity transformation (using only the eyes) is lower than 0.3 years. When using soft aligned faces with raw face detection the MAE degrades by roughly 1.2 years.

We compare our results (see Table 1) on FG-NET with the two best published results [12, 5] using leave-one-person-out cross-validation with manual eye alignment. In terms of global MAE, our eye aligned results are one year worse than [12] and [5] while our face detection result is roughly 2 years worse. The cumulative score curves in Fig. 1 right, confirms that the RPK [12] or BIF [5] approaches are marginally better than our manually aligned algorithm. On the other hand, our algorithm is much simpler and with lower computational requirements. The BIF method relies on processing the image with a large bank of filters, while RPK relies on an a mixture model adaptation of a face image description based on the distribution of the DCT on all image patches.

The work of Jahanbekam et al. [6] uses also face detection alignment on FG-NET. Their MAE is 7.54, which is optimistically biased since they do not use a leave-one-subject-out evaluation, and consequently, the same subject can be in the training and testing sets. Even in this case we outperform their approach, since for our MAE in this experiment is 6.9 (see Table 1).

Experiment/Age Range	0-9	10-19	20-29	30-39	40-49	50-59	60+	Global
Affine Alignment	2.72	3.84	5.62	11.19	19.68	29.43	40.53	5.56
Similarity Alignment	2.85	3.76	5.6	11.58	19.65	27.67	42.11	5.7
Face Detection	4.68	4.39	6.57	13.62	19.84	29.68	38.12	6.9
RPK [12]	2.3	4.86	4.02	7.32	15.24	22.2	33.15	4.95
BIF [5]	2.99	3.39	4.3	8.24	14.98	20.49	31.62	4.77

Table 1. MAE on each age range in the FG-NET database with 25×25 pixels images

3.2 Cross-database tests

Most age estimation algorithms only perform single database tests. To evaluate the performance of an age estimation algorithm we are interested in the algorithm's generalisation capabilities. In this section we train our algorithm using

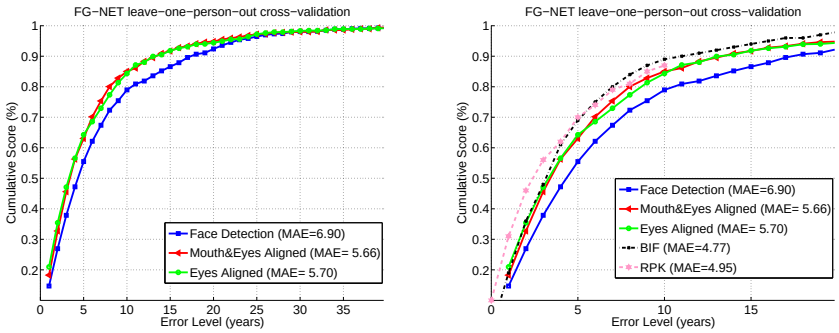

Fig. 1. Cumulative Score curves for FG-NET cross-validation experiments with 25×25 pixels sized images. Left: curves for different alignments using our method. Right: comparison with the two most competitive published methods.

one database and test it on a different database. In the case of GROUPS and PAL databases we train with 10-fold cross-validation. For training with FG-NET we perform leave-one-person-out cross-validation. In Fig. 2 we show the cumulative score curves and in Table 2 the MAE for our experiments.

We have made two groups of experiments. First train on a large database (GROUPS) and test on FG-NET and PAL. In this case we achieve a global MAE of about 15 years. In the second group of experiments we train with FG-NET, a small database, and test on GROUPS and PAL. The FG-NET/GROUPS experiment achieves also a MAE around 15 years. In the FG-NET/PAL case we achieve a much higher MAE because the age distribution in both databases is different. FG-NET has fewer people older than 40 whereas most of the subjects in PAL are above 40. This explains the differences on the results in Table 2.

Our results for GROUPS/FG-NET can be compared with others in the literature that use face detection and no further alignment [3]. In [3] a database from Internet with 219,892 samples is used for training. It is tested on FG-NET (see IAD/FG-NET in Table 2), being their MAE 9.49. Our result when training with a database with 13,051 samples is 12.62 for the GROUPS/FG-NET experiment in Table 2. We achieve a higher MAE because our database is one order of magnitude smaller and with a lower resolution age distribution. However, when looking at the per-age range MAEs, we get better results in 4 out of 7 age ranges (see columns IAD/FG-NET and GROUPS/FG-NET in Table 2).

4 Conclusions

In this paper we have presented a contribution to the age regression problem with results roughly within the state-of-the-art. Following the Occam Razor's principle we attack the problem from a simplicity driven perspective and with a low computational requirements solution in mind. We have realised that some

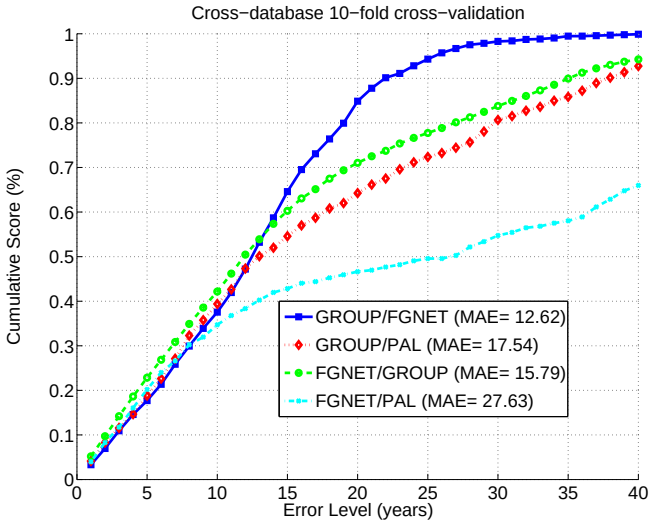

Fig. 2. Cumulative Score curves for cross-database experiments.

Experiment/Age Range	0-9	10-19	20-29	30-39	40-49	50-59	60+	Global
IAD/FG-NET[3]	10.98	8.15	6.05	7.92	13.42	22.75	29.96	9.49
GROUPS/FG-NET	15.55	12.98	6.88	5.65	12.20	19.66	22.64	12.62
GROUPS/PAL	—	10.42	7.59	6.69	9.30	17.27	28.90	17.54
FG-NET/GROUPS	9.56	5.77	9.41	—	—	29.55	53.52	15.79
FG-NET/PAL	—	5.56	5.84	14.27	23.62	32.85	49.10	27.63

Table 2. MAE on each age range in the cross-databases experiments

solutions in the literature are computationally complex getting in return low gain age estimation performance.

With manual eye alignment we get competitive results using a very simple and fast algorithm. When using soft aligned images, by means of face detection, the MAE estimation is only one year worse than the manual alignment. It is thus unclear whether full automatic alignment, which is computationally intensive, is worthy. A similar result was reported in the gender recognition problem [8, 2].

Moreover, we believe that the alignment problem can be solved by training, which would make the on-line computation much more efficient. By requiring no hard-alignment, our method is simple and fast both in training and in on-line classification. Given the low computational requirements, this method may be implemented in smart-phones or IP cameras.

The benchmark database for age estimation, FG-NET, has a very low number of images in some of the age ranges. This makes it difficult to train any learning algorithm and makes it difficult to get definitive conclusions by using only this

database. Therefore, cross-database experiments are a must in order to push the state-of-the-art in facial age estimation.

Acknowledgement

The authors gratefully acknowledge funding from the Spanish *Ministerio de Ciencia e Innovación* under contracts `TIN2010-19654` and the *Consolider Ingenio* program contract `CSD2007-00018`.

References

1. Andrew C. Gallagher, T.C.: Understanding images of groups of people. In: Proc. of CVPR. pp. 256–263 (2009)
2. Bekios-Calfa, J., Buenaposada, J.M., Baumela, L.: Revisiting linear discriminant techniques in gender recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(in press) (2011)
3. Bingbing, N., Zheng, S., Shuicheng, Y.: Web image mining towards universal age estimation. In: Proc. of ACM International Conference on Multimedia (October 2009)
4. Fukunaga, K.: Introduction to statistical pattern recognition. Academic Press (1990)
5. Guo, G., Mu, G., Fu, Y., Huang, T.S.: Human age estimation using bio-inspired features. In: Proc. of CVPR. pp. 112–119 (2009)
6. Jahanbekam, A., Bauckhage, C., Thurau, C.: Age recognition in the wild. In: Proc. of ICPR. pp. 392–395. IEEE (2010)
7. Johnson, R., Wichern, D.: Applied Multivariate Statistical Analysis. Prentice-Hall (1998)
8. Mäkinen, E., Raisamo, R.: Evaluation of gender classification methods with automatically detected and aligned faces. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(3), 541 – 547 (March 2008)
9. Minear, M., Park, D.C.: A lifespan database of adult facial stimuli. Behavior Research Methods, Instruments and Computers 36, 630–633 (2004)
10. Ralph Gross, I.M., Baker, S.: Generic vs. person specific active appearance models. Image and Vision Computing 23(11), 1080–1093 (2005)
11. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (May 2004)
12. Yan, S., Zhou, X., Liu, M., Hasegawa-Johnson, M., Huang, T.S.: Regression from patch-kernel. In: Proc. of CVPR (2008)
13. Yun Fu, G.G., Huang, T.S.: Age synthesis and estimation via faces: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(11), 1955–1976 (2010)