Assessing Facial Age Similarity: A Framework for Evaluating the Robustness of Different Feature Sets

Andreas Lanitis[1] and Nicolas Tsapatsoulis[2]

[1]Visual Media Computing Lab, Dept. of Multimedia and Graphic Arts
[2]Dept. of Communication and Internet Studies
Cyprus University of Technology
{andreas.lanitis, nicolas.tsapatsoulis}@cut.ac.cy

Abstract: A framework that can be used for assessing the suitability of different feature vectors in the task of determining the age similarity between a pair of faces is introduced. This framework involves the use of a dataset containing images displaying compounded types of variation along with the use of an ideal dataset, containing pairs of age-separated face images captured under identical imaging conditions. The use of the ideal dataset in conjunction with deliberate introduction of controlled noise, allows the extraction of conclusions related to the robustness of different feature vectors to different types of noise effects. The ultimate aim of this work is the derivation of comprehensive and accurate set of metrics for evaluating the performance of age progression algorithms in order to support comparative age progression evaluations.

1 Introduction

The topic of facial aging received considerable attention in the literature mainly in relation with the applications of age estimation and age progression (see [RCB09, FGH10] for comprehensive surveys). In facial age estimation the aim is to estimate the age of a subject by analyzing information derived from face images. In age progression the aim is to deform the appearance of a face in order to predict how the subject will look like in the future. Within this context the aim is to retain identity-specific facial features while transforming the appearance of age-sensitive features.

An important issue related to the development of accurate age progression algorithms is the formulation of a set of metrics that can be used for assessing the age difference between a pair of face images enabling in that way the assessment of the ability of an algorithm to generate faces showing a subject at the target age. Moreover, since in some cases, age progression may involve images captured under uncontrolled imaging conditions (i.e., images captured by surveillance cameras), it is important to assess the sensitivity of different metrics in the presence of different types of noise.

In this paper a comprehensive framework that can be used for assessing the suitability of different feature vectors in the task of determining the age similarity between a pair of faces is proposed. The framework in question involves the use of an ideal dataset,

containing pairs of age-separated face images captured under identical imaging conditions. The use of an ideal dataset in conjunction with deliberate introduction of noise, allows the extraction of conclusions related to the robustness of different feature vectors to different types of noise. In addition, images from the FG-NET Aging database [Lan08] were also used in order to obtain conclusions related to the ability of different feature vectors to assess age similarity using real life images that contain a combination of noise effects. As part of the experimentation, eight different types of features were considered. The ultimate aim is the derivation of comprehensive and accurate metrics for evaluating the similarity of age separated faces that can be used for assessing the performance of age progression algorithms. The early results and conclusions presented in this work constitute an important step towards accomplishing the aforementioned task.

2 Methodology

The aim of this work, and of the corresponding experiments, is to investigate the ability of a feature vector derived from a face image, to reflect aging information so that the age difference between a pair of face images can successfully be determined. As part of this investigation, two main approaches were adopted. The first involves the assessment of different features with respect to their correlation with age while the second approach is involved with the ability of a set of features to order a set of age separated images into increasing/decreasing age order. An important issue related to our methodology is the type of features used for describing face images. Towards this end eight different types of features are used. A short description of these features is provided in Section 2.3.

2.1 Correlation-Based Method

Let S_k be a set of N_k images showing the same subject k at various ages. Let also $\vec{x}_{ki}$ be an n-dimensional feature vector derived from the $i-th$ face image of subject k and a_{ki} be a scalar indicating the age of the face in that image. For every pair of images I_i^k and I_j^k belonging to set S_k we estimate the magnitude of the difference of feature vectors (df_{ij}^k) and difference in age (da_{ij}^k) as follows:

$$df_{ij}^k = \|\vec{x}_{ki} - \vec{x}_{kj}\|, \quad da_{ij}^k = a_{ki} - a_{kj} \tag{1}$$

where $\|\vec{x}\|$ is the norm of vector $\vec{x}$

By concatenating the values of df_{ij}^k and da_{ij}^k for all distinct pairs of images I_i^k and I_j^k ($i = 1, 2, ..., N_k - 1$, $j = i + 1, ..., N_k$) we form the vectors $\vec{df}^k$ and $\vec{da}^k$:

$$\vec{df}^k = [df_{12}^k \; df_{13}^k \; ... \; df_{1N_k}^k \; df_{23}^k \; df_{24}^k \; ... \; df_{2N_k}^k \; ... \; df_{N_k-1N_k}^k] \tag{2}$$

$$\vec{da}^k = [da_{12}^k \; da_{13}^k \; ... \; da_{1N_k}^k \; da_{23}^k \; da_{24}^k \; ... \; da_{2N_k}^k \; ... \; da_{N_k-1N_k}^k] \tag{3}$$

The average of the correlation coefficient c^k of vectors $\vec{df}^k$ and $\vec{da}^k$ across all set S_k as indicated in eq. 4 provides quantitative evaluation of the ability of the set of features considered to reflect age differences.

$$c = \frac{1}{N_S}\sum_{k=1}^{N_S} c^k, \quad c^k = \frac{\vec{df}^k}{\|\vec{df}^k\|\|\vec{da}^k\|} \tag{4}$$

where N_S is the number of subjects (sets S_k) in the dataset.

Thus, for every different feature set the average correlation coefficient c is computed with the aid of eq. 1-4 allowing comparison of feature sets w.r.t their ability to capture age differences recorded in facial images.

2.2 Prediction of the order of age-separated images

Correlation based age estimation, as described in the previous section, assumes that the vector representation of each face image changes linearly with age. This is definitely non-true because age changes in humans are rather non-linear [LTM13] while the majority of feature representations do not linearly encode changes in face appearance. Thus, we could assume that finding feature vector representations that allow correct classification of age-separated images might also help. For instance, we can train (in a learning by example manner) classifiers that are fed with these features and provide age progression estimation. In this case the non-linearity, mentioned above, as well as the individual way that age-related changes appear in face images [LTM13] can be learned by the classifier.

Having the above in mind we have performed a second test on the various feature representations investigated in this work. Instead of trying to obtain an estimate of age difference, we considered the ability to predict the right order of a set of age-separated images based on the plain vector representations described in Section 2.3. For this purpose we have used the Mean Average Precision (MAP) metric, that is the mean value of Average Precision, in sets of age-separated images depicting different subjects. Let as denote with $p_k(i)$ the precision of predicting the first i age-separated images in set S_k (in a given order, ascending or descending), thus:

$$p_k(i) = \frac{number\ of\ rightly\ predicted\ images\ in\ S_k\ (among\ the\ first\ i)}{i} \tag{5}$$

By averaging $p_k(i)$ for all i we get the Average Precision (for set S_k) p_k. The MAP is computed by averaging p_k across all subjects (sets S_k).

$$MAP = \frac{1}{N_S}\sum_{k=1}^{N_S} p_k, \quad p_k = \frac{1}{N_k}\sum_{i=1}^{N_k} p_k(i) \tag{6}$$

where N_k is the number of age-separated images in set S_k. In our experiments in the ideal dataset we used N_k = 10 for all sets S_k. In the FG-NET dataset S_k is not fixed. A value of

p_k=1 indicates perfect prediction of the order of all images in set S_k, while a MAP value of 1 indicates perfect prediction of the order of all images in all sets.

The ranking order is computed based on the correlation coefficient between the vector representations of the images ($\vec{x_{ki}}$). That is, given the vector $\vec{x}_{k1}$ corresponding to either the "youngest" or the "oldest" face image in set S_k, the remaining images in the set are ranked according to the correlation coefficient c^k_{1j} (the image with the highest correlation with the image in question is ranked first, that with the second highest correlation is ranked second, and so on):

$$c^k_{1j} = \frac{\vec{x}_{k1} \cdot \vec{x}_{kj}}{\|\vec{x}_{k1}\|\|\vec{x}_{kj}\|} \;\; j = 2, \; ...., \; N_k \tag{7}$$

MAP is our second indicator on how effectively a feature set can be used for assessing the performance of age progression algorithms.

2.3 Evaluation process

As part of the evaluation process, two main experiments are carried out: The first experiment involves an ideal set of images, whereas the second involves images captured under non-controlled imaging conditions.

Experiment 1: Ideal Dataset. For this experiment 300 face images, involving both real and synthetic ones, is used. In particular the database contained 30 pairs (thus, N_S=30) of real images showing the same person at two different ages (with an average age gap of 15 years). The dataset includes images from both genders and the age range is between between 1-75 years. Images in these pairs were captured under the same conditions in relation to lighting, face size, face orientation and facial expression hence the only distinct source of image differences in each pair is aging variation. For every pair of images a weighted morphing scheme was used for generating eight intermediate uniformly spaced images between the two images in each pair. Therefore, for each of the 30 subjects in the datasets we had available two real and eight synthetic images (thus, N_k=10 for every k). Figure 1 shows an indicative example of images belonging to the same subject, used in the experiment.

In order to assess the effect of different types of noise encountered in face images, in relation to the ability to assess age similarity, three different types of noise with increasing intensity, were added to one of the two images in each pair considered. In particular images were blurred using a Gaussian filter with standard deviations ranging from 0 to 50, salt and pepper noise covering 0% to 50% of the image and the correct positions of the landmarks were displaced by a random amount of 0 up to ± 10 pixels. It should be noted that the displacement of the landmarks, although it does not affect the image itself, it affects the overall process of feature extraction. Examples of noise-corrupted icons used in the experiments are shown in Figure 2.

Experiment 2: Real Life Dataset. In the second experiment the FG-NET dataset [Lan08] is used. The dataset contains 1002 images from 82 subjects (thus, N_S=82 while N_k varies

Figure 1: Indicative examples of images from the "ideal dataset". The leftmost and rightmost images are real images. The intermediate images are generated by interpolating the real images.

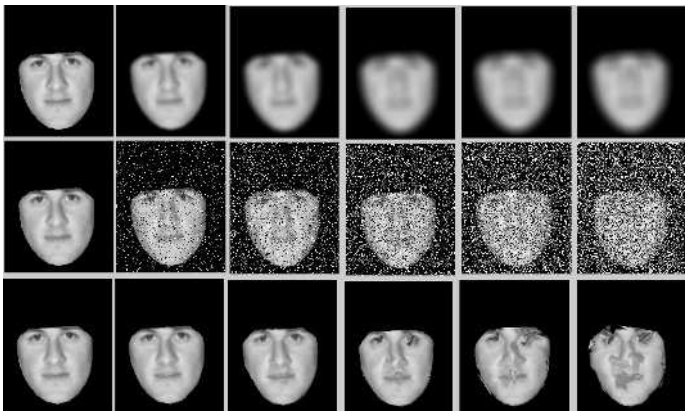

Figure 2: Examples of noise corrupted images. Images blurred with a Gaussian filter with standard deviation ranging from 0 to 50 (top row), images corrupted with salt and pepper noise with density ranging from 0% to 50% (middle row) and shape normalized facial patch extracted when the landmarks were displaced by a number of pixels ranging from 0 to 10 (bot-tom row) are shown.

with k). Apart from aging variation, images in the FG-NET aging database contain increased variation due to illumination, image quality, face orientation and expression, allowing the derivation of conclusions related to the ability of feature vectors to assess age similarity, in cases where images are captured under non-controlled conditions.

The experimental evaluation (in both experiments) involves the use of a variety of feature sets. In the *Shape-Normalized Texture (T)* set faces are represented by the texture in the shape-normalized facial region. In the *Raw Shape (S)* set each face is represented by the coordinates of 68 landmarks located on key positions. In the *Active Appearance Model Parameters (AAM)* set faces are coded into a number of Active Appearance Model [CET01] parameters. The *Local Binary Patterns (LBP)* set consists of vectors that include LBP [AHP06] patterns extracted form 33 local facial patches (these values are concatenated to form each vector). The *Histogram of Oriented Gradients (MHOG)* [DT05] are derived from four windows that constitute the bounding box of the facial region. In contrary the *Local Histograms of Oriented Gradients (PHOG)* are derived at windows lo-

Feature Type	T	S	AAM	LBP	MHOG	PHOG	AFD	SHIK
Ideal Dataset	0.47	0.52	0.59	0.53	0.59	0.61	0.42	0.54
FG-NET Dataset	0.22	0.29	0.37	0.30	0.36	0.40	0.29	0.38

Table 1: Correlation based evaluation for different features sets when tested on the "Ideal Dataset" and the "FG-NET Dataset"

cated on the 68 landmark points of each face under consideration. In the *Autocorrelation of Fourier Descriptors (AFD)* set 64 coefficients of the autocorrelation function of Fourier descriptors [CB84] are computed to represent each face. Finally, the *Spatial Histogram of Keypoints (SHIK)* is a SIFT based image representation obtained by accumulating the SIFT features on the points of a fractal grid [TT13].

3 Experimental Results

We evaluated the feature sets described above in terms of the proposed methodology using both the correlation-based method and the order prediction method. When no noise was added to the images from the "Ideal Dataset" (see Table 1), PHOG, AAM and MHOG achieved the best performance. Bearing in mind that the best performers are based both on local information (PHOG and MHOG) and global information (AAM) it can be concluded that both global and local information is important in reflecting age information. Also, it is interesting to note that on the "Ideal Dataset" even the shape representation achieves adequate performance as shape variations due to pose and expression variation are minimal in that case.

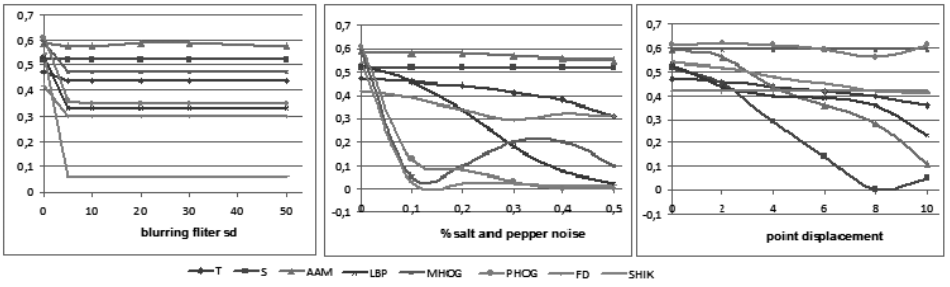

Figure 3: Correlation based evaluation for different features sets when images from the "Ideal Dataset" were blurred with the use of a Gaussian filter of increasing strength (x-axis) (left), when 'salt and pepper' noise of increasing strength (x-axis) was added (middle) and when the landmarks in the images were displaced by an increasing amount (x-axis)(right). Reported values (y-axis) refer to the average correlation coefficient.

When dealing with blurred images from the "Ideal Dataset" (Figure 3, left), AAM and MHOG features retain a reasonable performance with increasing amounts of blurring in-

Test Type	T	S	AAM	LBP	MHOG	PHOG	AFD	SHIK
Ideal (Progr.)	1	1	1	0.89	0.96	0.96	0.91	0.96
Ideal (Regr.)	1	1	1	0.92	0.96	0.96	0.89	0.96
FG-NET (Progr.)	0.78	0.78	0.80	0.72	0.77	0.80	0.74	0.75
FG-NET (Regr.)	0.76	0.77	0.77	0.75	0.75	0.78	0.74	0.77

Table 2: Results of the Progression and Regression tests on the "Ideal Dataset" and the "FG-NET Dataset": Mean Average Precision (MAP, see eq. 6) in predicting the right age order given a "young" image and predicting the right order of the "older" ones (Progression test) and given an "old" image predict the right order of the "younger" ones (Regression test)

dicating that these features could be used for assessing age similarity when dealing with non-ideal images (i.e., images captured by surveillance cameras). When images from the "Ideal Dataset" were corrupted with salt and pepper noise (Figure 3, middle) the performance of methods relying on local information exhibit significant performance deterioration as this type of noise disrupts the local texture. In contrast for this type of noise, global features (i.e., AAM) display the best overall performance.

In real applications involving the extraction of age similarity metrics, it is essential to locate facial landmarks automatically; a process that in some cases may not produce accurate depictions of the landmark positions. It is therefore important to use features that are not sensitive to the accuracy of landmark location. The results of the experiment where landmarks were displaced by a number of pixels (Figure 3, right) indicate that local features (i.e. MHOG, PHOG and SHIK) can retain a reasonable performance despite landmark displacements.

When dealing with images from the FG-NET Aging dataset (see Table 1) the best performing features in decreasing order proved to be the PHOG, SHIK, AAM and MHOG features. The fact that top performers contain both local-based and holistic methods, indicate that ideally both global and local evidence should be taken into account when assessing age similarity in the presence of compounded sources of variation.

Table 2 summarizes the results of the age-order prediction test (see Section 2.2) for the "Ideal Dataset" and the FG-NET Aging Dataset. Reported values refer to MAP (see eq. 6). The performance of all feature sets in the ideal dataset is close to perfect but the real challenge is the performance on the FG-NET dataset in which age-separated images present a high degree of variability of the faces shown. Among the eight compared feature sets the PHOG and AAM present the highest scores in both age progression and age regression tests. This is in partial disagreement with the noise tests on the "Ideal Dataset" (see Figure 3, left & middle) showing that noise corruption is not so common in practical situations as we considered. On the other hand, the landmark displacement test is of high importance because inaccurate detection of facial protuberant points is quite frequent. As Figure 3 (right) shows PHOG is more robust than the AAM method in this perspective. Nevertheless, the overall conclusion is that feature sets that combine local and global information seem to be able to capture age-related information remaining quite robust to landmark detection distortions.

4 Conclusion

In this paper we have introduced an experimental framework for evaluating facial age similarity, through face images, algorithms and feature sets. This framework includes the use of a specially designed dataset, the FG-NET dataset and two different tests to evaluate performance in conjunction with eight different types of features. We are currently working to expand the dataset and to precisely define a set of comprehensive facial features that could be used for assessing facial age difference and facial similarity, based on the proposed framework. It is anticipated that the completion of the aforementioned task will have direct applications in assessing the performance of age progression algorithms. However, it is important to note that a metric based on facial age similarity is not enough for assessing age progression algorithms as metrics based on ID similarity between source and age progressed images, are also required.

References

[AHP06] T. Ahonen, A. Hadid, and M. Pietikainen. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, December 2006.

[CB84] R. Chellappa and R. Bagdazian. Fourier coding of image boundaries. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, (1):102–105, 1984.

[CET01] T. F. Cootes, G. J. Edwards, and C.J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[DT05] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *Proc. of the 18th IEEE International Conference on Computer Vision and Pattern Recognition*, pages 886–893. IEEE Computer Society, 2005.

[FGH10] Y. Fu, G. Guo, and T. S. Huang. Age Synthesis and Estimation via Faces: A Survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(11):1955–1976, November 2010.

[Lan08] A. Lanitis. Evaluating the Performance of Face-Aging Algorithms. In *Proc. of the 2008 IEEE Intl. Conf. on Face and Gesture Recognition*, pages 1–6. IEEE, 2008.

[LTM13] A. Lanitis, N. Tsapatsoulis, and A. Maronidis. Review of ageing with respect to biometrics and diverse modalities. In M. Fairhurst, editor, *Age Factors in Biometric Processing*. IET, UK, 2013.

[RCB09] N. Ramanathan, R. Chellappa, and S. Biswas. Computational Methods for Modeling Facial Aging: A Survey. *Journal of Visual Languages and Computing*, 20(3):131–144, June 2009.

[TT13] Z. Theodosiou and N. Tsapatsoulis. Spatial Histogram of Keypoints. In *Proc. of the 20th IEEE International Conference on Image Processing*, pages 2924–2928, 2013.