Evaluating the Performance of Face-Aging Algorithms

Andreas Lanitis
Department of Multimedia and Graphic Arts, Cyprus University of Technology,
P.O. Box 50329, 3036, Lemesos, Cyprus
andreas.lanitis@cut.ac.cy

Abstract

An important aspect related to the development of face-aging algorithms is the evaluation of the ability of such algorithms to produce accurate age-progressed faces. In most studies reported in the literature, the performance of face-aging systems is established based either on the judgment of human observers or by using machine-based evaluation methods. In this paper we perform an experimental evaluation that aims to assess the applicability of human-based against typical machine-based performance evaluation methods. The results of our experiments indicate that machines can be more accurate in determining the performance of face-aging algorithms. Our work aims towards the development of a complete evaluation framework for age progression methodologies.

1. Introduction

The topic of face-aging received increased attention by the computer vision community during the recent years. This interest is motivated by the important real life applications where accurate age progression algorithms can be used. Such applications include the development of person identification systems robust to aging variation and the generation of accurate predictions of the current appearance of missing persons.

A major consideration pertaining to the development of improved face-aging algorithms is the formulation of performance evaluation methodologies that can be used for obtaining accurate performance evaluation results for different algorithms reported in the literature. In other face interpretation applications the establishment of objective performance evaluation metrics [10] coupled with the development of dedicated test sets has played a significant role in the development of the field.

However, in the case of age progression the definition of dedicated performance evaluation metrics is not straightforward and can be regarded as a research direction in the field, which can be as important as the development of face-aging algorithms. The process of assessing the performance of face-aging algorithms should focus on assessing the two important aspects related to face-aging:

- The ability to produce aged images that display aging characteristics of the target age group.
- The ability to retain the identity of subjects in aged faces.

In various occasions researchers who presented face-aging algorithms utilized different approaches towards the evaluation of the performance of their algorithms. In general such approaches are divided into human-based and machine-based evaluation methods. In this paper we provide a description of the most typical face-aging performance evaluation methods reported in the literature and perform comparative experiments in order to compare the suitability, effectiveness and accuracy of those measures. The ultimate aim of our work is to define and use the most suitable performance evaluation metrics that can be used for obtaining concrete conclusions related to the performance of face-aging algorithms.

The remainder of the paper is organized as follows: In section 2 we provide a brief literature review of the topic of face-aging followed by a presentation of the most commonly used face-aging performance evaluation metrics. In section 4 we describe an experimental evaluation process that aims to assess the suitability of different performance evaluation methods and in section 5 we present concluding comments.

2. Literature Review

Work related to face-aging could be classified into different categories according to the methods used, the range of ages considered, the type of data used, and the type of facial deformations inflicted. In this section we present few key approaches for each category.

Most researchers use gray scale images with hairstyles removed for training face-aging algorithms [4, 6, 9, 16]. The main reason for dealing with gray-scale images is the availability of data that usually comes in the form of gray scale images. Most researchers avoid dealing with hairstyles since the variability encountered in hairstyles needs special treatment. Initial attempts for dealing with hairstyles in aging algorithms were reported in the literature [18]. A number of researchers developed face

978-1-4244-2154-1/08/$25.00 ©2008 IE

aging methods that utilize 3D face data [5, 8, 17] instead of 2D images. The use of 3D faces eliminates variation due to face orientation within a training set. However, the main problem when dealing with 3D data is the availability of training data.

Age progression methodologies can be divided into the ones that operate on image-data directly [15,19] and to approaches that operate on a low dimensional coded face representation [4, 5, 6, 7, 9, 16, 17, 18] which is often derived based on the Active Appearance Model (AAM) methodology [2]. In general methods that operate on parametric face representations tend to be more tolerant to non-aging types of facial variation.

The range of ages considered in different studies is also an important factor. For example Ramathan and Chellappa consider only age progression in young faces [12] where most variation is associated with shape deformations; whereas in [3] only faces of adults are considered.

Age progression methodologies can be divided to the ones that deal only with shape deformations [21], texture deformations [3] or approaches that deal simultaneously with both shape and texture deformations [4, 6, 7, 8, 9, 15, 16, 17, 18] . Since the aging process affects both facial shape and texture, approaches that deal with both modalities are more promising. Due to the unique characteristics of the appearance of wrinkles in aged faces, the topic of wrinkle generation can be regarded as a special case of texture deformation [1, 3, 18, 19].

3. Performance Evaluation Metrics

In this section we describe human-based and machine-based face aging performance evaluation metrics used by researchers in previous studies.

3.1. Human-Based Evaluation

A number of researchers [6, 15, 18, 19], rely on human observers for assessing the quality of aged faces. Usually two types of experiments are run: For the first type of experiments source and aged faces are presented to observers, who have to decide whether the perceived age is truly increased in age-progressed faces. Also in some cases observers are required to classify aged faces in age groups, in an attempt to verify whether age progressed faces truly display the age characteristics of the target age group.

The second type of experiments involves the assessment of the ability of an algorithm to produce faces that retain the id information of the source images. In this type of experiments volunteers are requested to view age progressed faces and decide whether they belong to certain subjects from the dataset.

Variations from the two main types of experiments stated above include the numbers of test cases, the number of observers, the age range considered and the type of face images presented (i.e whether hairstyles are shown and whether gray scale or color images are used).

3.2. Machine-Based Evaluation

Given an aged face and the corresponding target face, Scandrett et al [16] calculate the root mean square shape difference and the root mean square intensity difference between the two faces. Scandrett et al [16] also calculate the correlation coefficient between age progressed textures and the texture of all samples at the target age. They observed that faces aged using algorithms reported in [16] produce higher values of the correlation coefficient between the aged faces and the target face when compared to the correlation coefficient between aged faces and other faces at the target age group. Shape-based and texture-based metrics assess the general similarity of two faces and do not provide dedicated information related to age similarity and id similarity of two faces. Since both metrics are heavily depended on non-aging types of variations the use of these measures cannot be used in conjunction with face images that display significant non-aging related variability.

Geng et al [4] and Lanitis et al [6] assess the accuracy of face-aging based on two different metrics: The first metric is based on the Mahalanobis distance and the second is based on face recognition results. In the first case they estimate the Mahalanobis distance between the coded representation of aged and target faces. The use of Mahalanobis distance for evaluating face-aging depends heavily on the way that the covariance matrix is calculated as it is possible to calculate a covariance matrix that emphasizes either id or aging variation.

In the case of face recognition, Geng et al [4] and Lanitis et al [6] run face recognition experiments where they attempt to recognize faces in the cases where the age of faces in test images is significantly different than faces of the same subjects in the training set. Face recognition is run using either raw faces images or artificially aged faces. In such experiments improvements in face recognition rate when using age-progressed faces, proves the ability of an aging algorithm to produce faces that retain the id characteristics of the source image. However, testing the accuracy of face-aging based on face recognition results is an indirect way to test the performance of face-aging. Other parameters related to the design of the face classifier and the training sets used, influence the results.

In a more recent approach, Lanitis [7] used two Support Vector Machine(SVM)-based [20] metrics for assessing the accuracy of age progressed faces produced by different age progression algorithms. The first metric assesses the ability of a method to produce faces similar to the ones in the target age group. The second measure assesses whether aged faces retain the id information when

compared to the source image. However, this method requires the use of an aging database that contains multiple images per subject, so that it is feasible to train SVM id classifiers for different subjects.

4. Experimental Evaluation

In this section we describe experiments that aim to assess the potential of using the metrics described in section 3 for face aging performance evaluation. In our experimental evaluation we assess both the applicability of human-based and machine-based methods as applied to the problem of performance evaluation of age-progression methodologies. For the needs of our experiments instead of using age progressed images we use real images of subjects at the target age, since in that case we already know the true age and id of subjects in the test images. For the needs of our experiments we have used images from the FG-NET Aging database (http://fgnet.rsunit.com/) that contains 1002 images from 82 subjects. The database was divided in the following groups:

Group A and Group B: Each group contains half of the images of each of the 82 subjects. The separation of images was done in such way so that both groups contain similar distribution of ages for each subject. In this case group B contains 501 previously unseen faces images of subjects that also appear in group A.

Group C and Group D: Group C contains all images of subjects with ids 001-040 (498 images). Group D contains all images of the subjects with ids 041- 082 (504 images). In this case group D contains images of previously unseen subjects when compared to the faces in group C.

For our experiments we perform the training and we carry out initial investigations using images from groups A and C and test the performance using images from groups B and D.

4.1. Machine-Based Methods

In this section we describe experiments that aim to quantify the applicability of different machine-based metrics to the problem of performance evaluation of face-aging. Performance evaluation metrics are used for assessing two aspects of facial similarity: The id and age similarity. The id similarity measure assesses the similarity of two faces with respect to inter-individual characteristics whereas the age similarity measure assesses the age similarity between two faces with respect to aging related characteristics. As part of the experiments we estimate the id and age similarity between pairs of images and assess the ability of each method to produce different values in the cases that we deal with faces belonging to the same subject or belonging to the same age group when compared to the case that we deal with faces not belonging to the same subject or age group.

4.1.1 Performance Evaluation Metrics

In our experimental evaluation we evaluate the following metrics for assessing the similarity between a pair of faces:

Normalized shape-based:
The shapes of two faces, as defined by 68 landmarks located on key characteristics, are aligned and the root mean square distance between corresponding points is evaluated.

Normalized texture-based:
The shape-free texture of the internal facial region is collected and normalized with respect to the variance at each pixel and the average intensity within the facial region. The root mean square difference between the textures of corresponding pixels within the facial region is calculated.

Mahalanobis:
Unlike the shape-based and texture-based measures that operate directly on images, the Mahalanobis distance is calculated based on a low dimensional Active Appearance Model-based representation [2] of the faces in the training and test sets. Details about coding faces into this representation appear elsewhere [2, 6]. For the needs of this experiment, all faces in the train and test sets are divided into age groups defined by 5-year age intervals between the minimum and the maximum age of faces in our database. Although it is possible to apply this method using the distributions of samples for each age instead for each age-group, it is advantageous to use age groups because (1) there are not enough samples in the database to model the distributions for each age and (2) in age groups of adults aging variation within an age group is negligible.

All images from the training set are used for training a shorter distance classifier for id classification and age-group classification. During the training procedure covariance matrices that describe the scatter of samples belonging to different subjects and age groups are estimated. We refer to the two covariance matrices as the age and id covariance ($\mathbf{C}_{age}$ and $\mathbf{C}_{id}$). Assuming that $\mathbf{X}_1$ and $\mathbf{X}_2$ are the model-based representation of two faces the Mahalanobis based id (d_{id}) and age (d_{age}) distances are calculated using:

$$d_{age} = (\mathbf{X}_1 - \mathbf{X}_2)\mathbf{C}_{age}^{-1}(\mathbf{X}_1 - \mathbf{X}_2)'$$
$$d_{id} = (\mathbf{X}_1 - \mathbf{X}_2)\mathbf{C}_{id}^{-1}(\mathbf{X}_1 - \mathbf{X}_2)' \quad (1)$$

As illustrated by different authors [5, 6, 16] the process of aging causes diverse effects for different subjects. In an

attempt to deal with this issue, for each age group we establish two distributions, corresponding to the two genders. During the calculation of the age distance (d_{age}) we use the appropriate covariance matrix according to the gender of the two samples.

SVM-based:

We use the AAM-based coded representation of faces from the training set to train age and id SVM classifiers. As an extension to the basic method we also train SVM age classifiers for different genders, in an attempt to deal with the diversity of aging variation. The parameters of the classifier (i.e kernel type and kernel parameters) are optimized by using part of the training set as a test bench. Given a pair of faces the age and id similarity measures are calculated by considering the SVM similarity between the coded representation of the first face and the actual age and id distributions that the second face belongs to. In this respect the SVM similarity between a feature vector and a distribution is calculated using

$$SVM_{sim}(\mathbf{X}) = \mathbf{w}_i \cdot k_i(\mathbf{X}) + \mathbf{b}_i \qquad (2)$$

Where $\mathbf{X}$ is a vector containing a set of face model parameters and $k_i(\mathbf{X})$ is a vector containing the kernel function evaluations of $\mathbf{X}$ with respect to each of the support vectors of the ith class. $\mathbf{w}_i$ is a vector containing the weights for each support vector of the ith class and $\mathbf{b}_i$ contains the bias value for the ith class. The support vectors, the vector with the weights ($\mathbf{w}$) and the vector with bias values ($\mathbf{b}$) are defined during the process of training an SVM classifier.

4.1.2 Experimental Results

During the training stage we use images from group A of the FG-NET aging database for training Mahalanobis and SVM age and id classifiers. During the testing phase we compare the appearance of all pairs of faces in group B (250500 cases) based on the shape-based, texture-based, Mahalanobis distance-based and SVM–based metrics defined in section 4.1.1. In the case of the age similarity measure we perform experiments using either a single or two distributions (one for male and one for female samples) for each age group. The results of the experiments are shown in Table 1. In order to improve the clarity of presentation of quantitative results and facilitate the comparison between different methods, we scale numerical results to values between zero and 100 where a value of 100 indicates maximum similarity and a value of 0 indicates minimum similarity.

According to the results, the shape-based and texture-based measurements do not display remarkable differences in similarity values when dealing with faces belonging to the same age group or same subject, when compared to the case that we deal with different age groups or different subjects. This is expected, since the FG-NET Aging database contains faces that display significant variations in illumination and pose. In [16] where shape and texture-based measures were used for assessing the accuracy of age-progressed faces, face images were preprocessed in order to suppress non-aging related variation. Mahalanobis-based measures produce improved results when compared to the shape and texture-based metrics. In this comparative test, SVM-based methods produce the best performance since SVM age and id similarity measures display noticeable differences when dealing with the same id or age group, when compared to cases that we deal with pairs of images of different persons or faces belonging to different age groups. The use of different age group distributions for male and female subjects results in slightly improved performance. Further work is required for establishing multi-modal age group distributions, in order to cover diverse aging and id effects for different groups of subjects so that metrics that display increased separation of id and age similarities are derived.

ID Similarity Metrics*				
	Shape	Texture	Mahalanobis	SVM
Same Age	76 (12)	83 (10)	82 (7)	40 (9)
Diff. Age	74 (12)	82 (10)	81 (8)	39 (11)
Same ID	79 (13)	84 (10)	87 (7)	63 (11)
Diff. ID	75 (12)	83 (10)	81 (8)	37 (9)

Age Similarity Metrics				
	Mahalanobis	Mahalanobis (With age distributions for each gender)	SVM	SVM (With age distributions for each gender)
Same Age	82 (8)	87 (6)	56 (11)	59 (10)
Diff. Age	79 (8)	83 (8)	38 (9)	39 (9)
Same ID	85 (8)	85 (8)	43 (11)	43 (11)
Diff. ID	81 (8)	81 (8)	42 (11)	42 (11)

** In the case of Shape and Texture the same metrics are used both for id and age similarity*

Table 1: Comparison of machine-based performance evaluation metrics.

4.2. Human-Based Methods

In most studies presented in the literature, the performance of face-aging algorithms is assessed based on human observations. In this experiment we attempt to quantify the performance of human observers in the task of evaluating age progressed faces. In our experiments 30 human observers were presented with 100 pairs of face images chosen randomly from Group D of the FG-NET Aging Database.

For each test pair, observers were requested to indicate whether the two faces belong to the same age group and whether the two faces belong to the same subject. The experiment was run twice where for the first run

volunteers were presented with gray-scale hairless faces, whereas for the second experiment raw face images were used. Typical samples of test pairs for the two experiments are shown in figure 1. Table 2 (2nd and 3rd rows) show the results of the experiments.

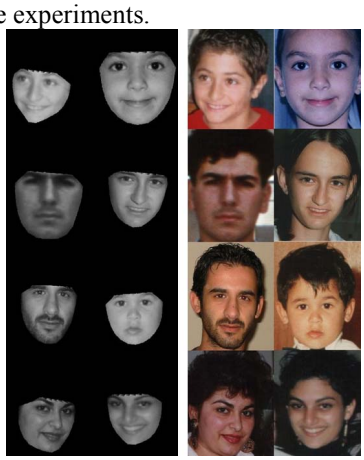

Figure 1: Samples of pairs of faces with hairstyles cropped (left column) and pairs of raw face images (right column).

4.3. Human-Based VS Machine-Based Methods

In this experiment we attempt to compare directly the performance of human-based and machine-based performance evaluation methods by running an experiment similar to the one described in section 4.2. However, in this case we use machine-based criteria to decide whether a pair of images contains faces of the same subject or the same age group.

For the needs of the experiment we use images from group C of the FG-NET Aging Database for training Mahalanobis and SVM classifiers that can be used for defining whether a pair of faces belongs to the same subject or the same age group. During the training stage we generate a training set that contains all pairs of faces within the training set and for each pair we calculate the absolute difference between the coded representations of the two samples. The difference vectors constitute the training set for training two-class classifiers for deciding whether two faces belong to the same subject or same age group. This framework bears similarities with work on face id verification reported by Ramathan et al [11, 13]. During the test phase we run two experiments. For the first one we test the system with the randomly selected 100 pairs from set D, that were used in conjunction with the human test described in section 4.2. These results give an indication of the comparative performance between machines and humans. For the second experiment we test the system using all pairs of faces from set D (253512 test cases). This experiment gives an indication of the results expected on a large scale evaluation experiment.

During the test phase the coded-based representation of two faces is obtained and the absolute difference between the two vectors is established and used as the feature vector for classification. Table 2 shows the results of the experiments.

Method	Number of Test Cases	Correct ID Rate	Correct AGE Rate
Humans(Color/Hair)	100	77.8	75.7
Humans(Cropped/Gray)	100	66.9	63.8
Mahalanobis	100	70.0	52.0
SVM	100	80.0	75.5
Mahalanobis	253512	69.8	57.8
SVM	253512	97.7	76.3

Table 2: Human-based VS Machine-based Evaluation

Experimental results demonstrate that in the case that humans are presented with gray-scale hairless faces, machine-based performance evaluation using an SVM classifier is superior to human-based evaluation. Even when human observers are presented with color faces with hair, SVM classifiers produce comparable performance to the performance achieved by humans. The Mahalanobis-based classifier produces inferior performance that is attributed to the fact that the distributions of face differences for faces belonging to the same age group or same subject are not normal. When the SVM classifier is tested on all pairs from the test set it achieves a noticeable improvement in performance in relation with the determination of id similarity. This happens because the percentage of image pairs containing images of the same subject in the whole test set (253512 cases) is much lower when compared to the percentage of image pairs of the same subjects in the reduced test set with 100 test cases. Since it is not feasible to request human volunteers to report results for all test pairs it is not possible to provide actual figures for the results of human observers when tested on the 253512 test cases.

5. Conclusions

The topic of accurate performance evaluation for age progression methodologies is of utmost important for further development of this field. Due to the nature of the problem, standard performance evaluation metrics similar to the ones used in other face interpretation applications are not applicable to this problem. For that reason most researchers working in the field rely on the judgment of human observers for evaluating their methods. However, human-based evaluation is a time-consuming process that produces subjective results. In this paper we demonstrate that it is possible to replace human expertise with appropriate machine-based evaluation methodologies, in the task of performance evaluation of face-aging algorithms. By doing so it is possible to get improved and more accurate performance evaluation metrics using a fast, low cost process that can be used for large scale evaluation experiments.

For the results presented in this paper we have used the

FG-NET Aging Database. In an attempt to verify the applicability of the proposed performance evaluation scheme we plan to report results when different publicly available aging databases are used. For example we plan to report results using images from the Morph [14] database. Since the FG-NET Aging database contains remarkable non-aging related types of variations, it is expected that improved results will be obtained when the same framework is applied on other aging databases.

We also plan to carry out work that aims towards the development of improved performance evaluation metrics. In this respect we plan to introduce additional metrics that aim to assess specific aspects of the aging process that are not described adequately by the existing id and age metrics. Also we plan to use person specific age distributions, so that individual aging trends are better explained. The early results using different age distributions for males and females suggest that this is a reasonable direction.

Our ultimate aim is to develop and use a complete framework for evaluating age progression methodologies. For this purpose a standardized aging database containing samples from different ethnic origins and uniform age distribution across age groups will be required. The introduction of a standardized performance evaluation framework will enable the direct comparison of face-aging algorithms reported in the literature so that the most promising technologies will evolve.

Acknowledgements

This work was supported by the Cyprus Research Promotion Foundation research grant (EPYAN/0205/08).

References

[1] L.Boissieux, G. Kiss, N. Magnenat-Thalmann and P. Kalra. Simulation of Skin Aging and Wrinkles with Cosmetics Insight. Computer Animation and Simulation, pp.15-27, 2000.

[2] T.F.Cootes, G.J. Edwards and C.J. Taylor. "Active Appearance Models". IEEE Transactions of Pattern Analysis and Machine Intelligence, 23, pp 681-685, 2001.

[3] M. Gandhi, A Method for Automatic Synthesis of Aged Human Facial Images. Msc Thesis, Dept. of Electrical and Computer Engineering, McGill University, 2004.

[4] X. Geng, ZH. Zhou and K. Smith-Miles. Automatic Age Estimation Based on Facial Aging Patterns. IEEE Transactions of Pattern Analysis and Machine Intelligence, 29(12), 2234-2240, 2007.

[5] T.J Hutton, B.F. Buxton, P. Hammond and H.W Potts. "Estimating Average Growth Trajectories in Shape-Space Using Kernel Smoothing", IEEE Transactions on Medical Imaging, 22(6), 2003.

[6] A. Lanitis, C.J. Taylor and T.F. Cootes, "Toward Automatic Simulation Of Aging Effects on Face Images". IEEE Transactions of Pattern Analysis and Machine Intelligence, Vol 24, no 4, pp 442-455, 2002.

[7] A. Lanitis, Comparative Evaluation of Automatic Age Progression Methodologies. EURASIP Journal on Advances in Signal Processing, Article ID 239480, 2008.

[8] W. Lee, Y.Wu and N. Magnenat-Thalmann. Cloning and Aging in a VR Family. Proceedings of IEEE VR'99 (Virtual Reality), 23-34, 1999.

[9] E. Patterson, K.Ricanek, M. Albert and E.Boone. Automatic Representation of Adult Aging in Facial Images. Procs. of the 6th IASTED International Conference on Visualization, Imaging, and Image Processing, 2006.

[10] P.J. Phillips et al. The FERET Evaluation Methodology for Face-Recognition Algorithms, IEEE Transactions of Pattern Analysis and Machine Intelligence, 22(10), pp 1090--1104, 2000.

[11] N. Ramanathan and R. Chellappa. "Face Verification Across Age Progression". IEEE Transactions on Image Processing, Vol 15, No 11, pp 3349-3361, 2006.

[12] N. Ramanathan and R. Chellappa. "Modeling Age Progression in Young Faces". IEEE Conference of Computer Vision and Pattern Recognition, 2006.

[13] L. Haibin, S. Stefano, N. Ramanathan, D. Jacobs. "A Study of Face Recognition as People Age", IEEE International Conference on Computer Vision, 2007.

[14] K. Ricanek and T. Tesafaye. MORPH A Longitudinal Image Database of Normal Adult Age-Progression. Procs. of the 7th IEEE International Conference on Automatic Face and Gesture Recognition, 2006.

[15] D.A.Rowland and D.I. Perrett, "Manipulating Facial Appearance through Shape and Color". IEEE Computer Graphics and Applications, Vol. 15, no 5, pp 70-76, 1995.

[16] C.M. Scandrett (née Hill), C.J. Solomon and S.J. Gibson. A person-specific, rigorous aging model of the human face. Pattern Recognition Letters, 27 (15),1776-1787, 2006.

[17] K. Scherbaum, M. Sunkel, H.-P. Seidel, V. Blanz , Prediction of Individual Non-Linear Aging Trajectories of Faces . Computer Graphics Forum 26 (3), 285–294, 2007.

[18] J. Suo, F. Min, S. Zhu, S. Shan and X. Chen. A Multi-Resolution Dynamic Model for Face Aging Simulation. Proceedings of the IEEE Conference of Computer Vision and Pattern Recognition, 2007.

[19] B.Tiddeman, M. Burt and D. Perrett, Prototyping and transforming facial textures for perception research, Computer Graphics and Applications, 21 (5), 42-50, 2001.

[20] V.N. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, New York, 1995.

[21] J. Wang and C.X. Ling. Artificial Aging of Faces by Support Vector Machines. Advances in Artificial Intelligence, LNCS, Springer, Vol 3060, pp 499-503, 2004.