Q-STACK AGING MODEL FOR FACE VERIFICATION

Andrzej Drygajlo, Weifeng Li and Kewei Zhu

Swiss Federal Institute of Technology Lausanne (EPFL), CH-1015 Lausanne, Switzerland
phone: +41 21 693 43 28, fax: +41 21 693 76 00, email: andrzej.drygajlo@epfl.ch
web: scgwww.epfl.ch

ABSTRACT

Permanence of biometric features for face verification remains a largely open research problem. Actual and up-to-date at the time of their creation, extracted features and models relevant to a person's face may eventually become outdated, leading to a failure in the face verification task. If physical characteristics of the individual change over time, their classification model has to be updated. In this paper, we develop a Q-stack classifier that performs face verification across age progression. Originally, Q-stack classifier has been proposed to use class-independent signal quality measures and baseline classifier scores in order to improve classification. In this paper we demonstrate the application of Q-stack classifier on the task of biometric identity verification using face images and associated metadata quality measure - age. We show that the use of the proposed technique allows for reducing the error rates below those of baseline classifier created at the time of enrolment.

1. INTRODUCTION

Successful utilization of face information for person recognition depends to a large extent on the classification features pertinent to this biometric modality, and their models. It is a well-known fact that the individual physical characteristic features change with time. In particular, aging changes person's physique at a slow rate, albeit irreversibly. It is therefore likely that individual face models created at some point in time may become less relevant or even obsolete as the time passes. This expectation is confirmed by recent research [1]. For this reason, it is of prime importance to understand and quantify the temporal reliability of face biometric features and models, and consequently of the face verification systems.

The problem of time validity of biometric templates received only a marginal attention from researchers. The variation caused by face aging is often neglected compared with pose, lighting, and expression variations.

Nowadays, digital face images are becoming prevalent in government issued travel and identity documents (e.g., biometric e-passports and national identity cards). The non-intrusiveness characteristic of face biometric often compensates for its relatively low accuracy. As a result, a number of critical security and forensic applications require automatic verification capability based on facial images. Developing face verification systems that are robust to age progression would enable the successful deployment of face verification systems in those large-scale applications.

Since faces undergo gradual variations due to aging, periodically updating (e.g. every six months) large-scale-application face databases with more recent images of subjects might be necessary for the success of face verification systems. Since periodical updating such large databases would be a tedious and very costly task, a better alternative would be to develop aging-aware face verification methods. Only such methods will have the best prospects of success in longer stretches of time [2].

Aging is a complex process that affects both the shape of the face and its texture. Most of the reported work has been focused around the problem of visualizing the changes of appearance of face images as the time progresses [3]. Very limited evidence is available as to the impact of the changes of appearance on actual recognition performance.

The biometrics research community realized the importance of temporal changes in individual features and databases are created with predefined intervals between sessions of data collection. However, in commonly used benchmarking databases this period is within the range of weeks or months [4]. Such short intervals are unlikely to give a good understanding of the temporal dynamics of biometric traits, and the observed short-term face image and feature variability is more likely to be due to the environment factors rather than time flow-related changes. In order to examine the long-term reliability of biometric features the collection sessions must cover periods measured in years. In order to address this problem in this paper we use daily photo recordings over years, publicly available on the YouTube and the MORPH Database [5] collected for investigation of face age progression. An inherent limitation to the use of the YouTube recordings is the amount of subjects that appear on the photos over a long enough time stretch (e.g. three or more years). The advantage of such recordings is that they provide a large amount of face images of an individual, sampled at daily time intervals over a long period. In the experiments reported here, we have used recordings which covered 1200 days.

Such recordings allow us to model age progression in human faces and to build face verification systems robust to age progression. Certain amount of early recordings is used to extract features and build models, whose temporal score dynamics is further analyzed based on recordings with later time stamps.

The paper specifically identifies a possible way of using the age information as a class-independent quality measure. Age is a factor that directly impacts the comparative quality of images recorded at different times. If the biometric samples being compared differ substantially in age, recognition accuracy is affected. Facial recognition is reputed to be more sensitive to data aging than iris or fingerprint. However, any biometric system will be sensitive to age if relevant physical changes occurred in the intervening period. More substantial physical degradation may become an issue as the difference in age increases. Aging can be considered a metadata [7] because the quality of the samples themselves is not the issue:

Figure 1: Sample face images of individuals with the age progression (about three years).

the age difference is the cause of the degradation of accuracy.

Using such interpretation, the recently developed framework of classification with quality measures, Q-stack [8], is deployed to create a new face verification system robust to aging of biometric templates. Q-stack is a general framework of classification with quality measures that is applicable to uni-, multi-classifier and multimodal biometric verification with one or more quality measures. This paper is focused on the Q-stack solution that allows for improved class separation using age as a metadata quality measure.

The paper is structured as follows. Section 2 provides correlation analysis between age metadata quality measure of example adult human faces and the scoring of PCA baseline classifiers. In Section 3 we propose a general framework - Q-stack aging model - a stacking classifier for the task of biometric identity verification using face images and associated metadata quality measure - age. Section 4 presents experimental results with their discussion and Section 5 concludes the paper.

2. AGING INFLUENCE ON THE FACE CLASSIFIERS

Figure 1 shows the daily image samples for the four people during 1200 days. We can see that the images are not taken under controlled conditions. The face images are not strictly frontal and subtle variations in head pose exist. Further, we observed that the face images were taken under nonuniform illumination conditions and with different backgrounds.

First, we analyze the influence of age progression on the baseline classifier scores. The data for each individual are divided into a model training data set (first 100-day images) and an evaluation data set (the images from 101th day to 1200th day). We perform a Principle Component Analysis (PCA) on the images. The PCA projection space is found using the model training data set and the test images are projected onto the subspace defined by eigenvectors associated with a set of 32 largest eigenvalues. During the classifier training phase, Gaussian Mixture Models (GMM) are built over the model training data set using the PCA features. During the testing phase, log likelihood of the genuine access class given the PCA features extracted from a testing image

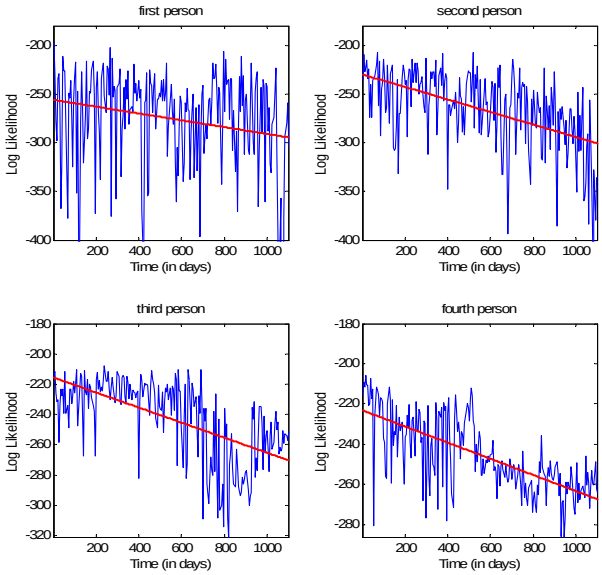

Figure 2: The influence of age progression on the scores. The bold lines in each sub-figure are linear fittings of the variations of log likelihood with the age progression. The four individuals correspond to the ones in Fig. 1.

(every five-day interval) in the evaluation data set is computed.

Figure 2 shows the effects of age progression (after 100 days) on the classifier scores for the four persons from Fig. 1. Table 1 presents the estimates of the pair-wise dependencies between log likelihood scores and the age progression, in terms of the Pearsons correlation coefficient (PCC) and the mutual information (MI). From Fig. 2 and Table 1 we can draw following observations:

Table 1: Pearsons correlation coefficient (PCC) and the mutual information (MI) between log likelihood and the age progression.

Person	first	second	third	fourth
PCC	-0.21	-0.52	-0.61	-0.69
MI	0.09	0.12	0.23	0.27

- The plots in Fig. 2 as well as PCC and MI values in Table 1 show that there is an evident conditional dependency between age progression and baseline classifier scores. This motivates us to use age as class-independent quality measure in the Q-stack classifier in order to improve long-term classification performance of face verification systems.
- As shown in Fig. 2, there exists general tendency of the classifier scores distribution dependent on the age progression for all the individuals. As the age increases, the log likelihood values generally decrease with some exceptions.
- In Fig. 2, the variations of the log likelihood values do not always decrease as the age increases. For example, it is noticed that at the end part of sub-plot of the third person, there is a steep decrease and then an increase of the

log likelihood value, and that the steep decrease regions appear in the cases where the third person did not wear the glasses with black brims as usually. Since the images are not taken under controlled recording conditions, there are other factors such as illumination and hair-style which influence the classifier scores.

- The plots in Fig. 2 represent distributions of classification scores given the age progression. The Pearsons correlation coefficient (PCC) and the mutual information (MI) estimates in Table 1 only show the general tendency of the dependency between the classification scores and age progression.

3. Q-STACK AGING MODEL

Figure 3 shows a diagram of the $Q-stack$ classifier [6]. Identity-related information is composed of a biometric signal S, classified by a baseline classifier, resulting in a score x. At the same time, the signals undergo quality measurements, resulting in m quality signals $\mathbf{qm} = [qm_1, qm_2, ...qm_m]$. Multiple quality measures can be used to characterize one signal. The score x is concatenated with the quality measure vector to form an evidence vector $\mathbf{e} = [x, \mathbf{qm}]$. The evidence vector $\mathbf{e}$ becomes a feature vector for the stacked classifier. If no quality measures are present, the scheme shown in Figure 3 is equivalent to classical Wolpert's stacking [9].

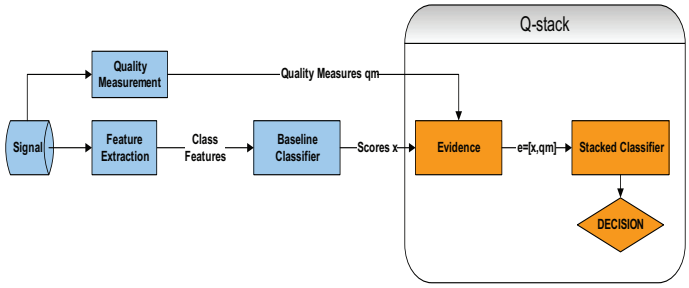

Figure 3: Q-stack architecture, in which baseline classifier scores and quality measures jointly serve as features to a second-level, stacked classifier.

The proposed method of Q-stack is a generalized framework which encompasses previously reported methods of using quality measures in biometric verification [6]. In particular, those methods have been shown to be case- and data-specific, largely heuristic approximations of an optimal decision boundary in the evidence space. As opposed to the ad-hoc methods which can hardly be generalized to new data sets and classifier architectures, $Q-stack$ attempts to approach the optimal decision function by learning the causal dependencies between quality information and baseline classifier scores from available data.

Age information can be used as one of the quality features in the framework of Q-stack. As class-independent feature, age information does not provide information about the identity of the individual whose face appears in the image. However, if one face model (template) is continuously used, time difference (age difference) has an obvious influence on the face matching scores as shown in Section 2. This influence translates into a statistical dependence between the scores and age information. This dependence is consequently modeled and exploited for greater classification accuracy in the Q-stack scheme.

4. EXPERIMENTS AND RESULTS IN AGING FACE VERIFICATION

In order to verify the claim that an inclusion of age progression in the evidence vector allows for more accurate classification in the Q-stack framework than using baseline classifiers, we conducted a series of experiments with various configurations of available evidence. In particular, the experiments aimed at showing that those configurations of evidence which include age progression as a classification feature to the stacked classifier give better classification results than baseline systems without using age progression information.

An example of face verification system is defined as follows: Class A - genuine identity claim from the first person in Fig. 1; Class B - imposter attempt from the other three individuals in Fig. 1. The baseline GMM classifier is built from the training samples originating from the first 100 images (the first 100 days) of the first person. Figure 3 shows a diagram of the Q-stack framework for our face verification experiments, in which baseline GMM classifier scores and age progression information jointly serve as features to a second level stacked classifier. At this level, the following Support Vector Machine (SVM) based classifiers are employed:

SVM classifier with linear kernel (SVM-lin) The SVM-lin classifier is a linear classifier that maximizes the classification margin between the classes in the same classification space where the evidence vectors lie. Unlike the commonly-used linear discriminants [10], the SVM-lin does not make assumptions regarding the Gaussianity of the class-conditional joint distributions. Since it is a linear classifier in the evidence space, the SVM-lin classifier is able to capture linear dependencies between the components of the evidence vector.

SVM with radial basis functions kernel (SVM-rbf) The SVM-rbf classifier utilizes the kernel trick [11] to find a linear separating hyperplane in a transformed, arbitrarily high-dimensional space [12]. Projected back onto the original evidence space, the decision boundary generated by the SVM-rbf classifier may be therefore nonlinear and of arbitrary complexity. The SVM-rbf classifier is capable of capturing non-linear dependencies between the evidence components.

4.1 Experiments on YouTube data

Figure 4 shows the application of Q-stack to the baseline, SVM-lin and SVM-rbf classifiers with evidence vector $\mathbf{e} = [x, qm]$, where qm represents daily age progression, for the training data set (first 100-day images). Figure 5 shows the application of Q-stack for the same types of classifiers for the evaluation data set (with 5-day age progression from 100 to 1200 days). In both figures, the horizontal dashed line shows the decision boundary of the baseline classifier. The dash-dot and solid lines denote the Q-stack decision boundaries using SVM-lin and SVM-rbf stacked classifiers.

Table 2 shows the verification performance of short-term 100-day training data set (Table 2). Classification results of face verification are reported in terms of false acceptance rate (FAR), false rejection rate (FRR) and half total error rate

Table 3: Verification performance of different methods over three years in terms of false acceptance rate (FAR), false rejection rate (FRR) and half total error rate (HTER).

	0-0.5 year	0.5-1.0 year	1.0-1.5 years	1.5-2.0 years	2.0-2.5 years	2.5-3.0 years
Baseline						
FAR [%]	0	0	0	0	0	0
FRR [%]	94.44	100.00	91.67	100.00	97.22	100.00
HTER [%]	47.22	50.00	45.83	50.00	48.61	50.00
SVM-lin						
FAR [%]	3.70	10.19	10.19	11.11	2.78	2.78
FRR [%]	69.44	55.56	30.56	38.89	19.44	30.56
HTER [%]	36.57	32.87	20.37	25.00	11.11	16.67
SVM-rbf						
FAR [%]	0.93	5.56	5.56	9.26	2.78	1.85
FRR [%]	75.00	61.11	36.11	50.00	22.22	30.56
HTER [%]	37.96	33.33	20.83	29.63	12.50	16.20

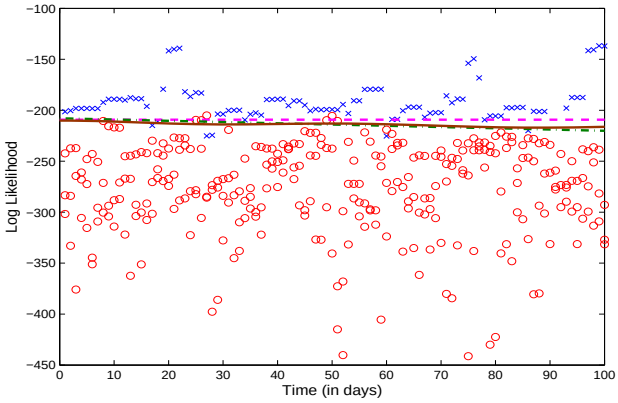

Figure 4: Class separation in the age-score space for the first 100 days. The '×' and '○' marks represent the scores of the genuine user (first person in Fig. 1) and the impostors (the other three persons in Fig. 1), respectively. The dashed line shows the decision boundary of the baseline classifier. The dash-dot and bold lines denote the Q-stack decision boundaries using SVM-lin and SVM-rbf stacked classifiers.

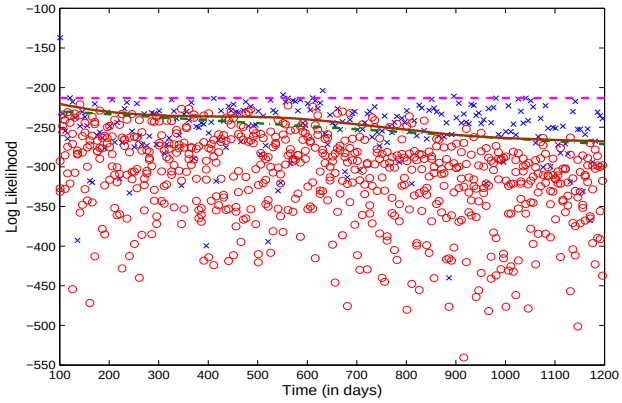

Figure 5: Class separation in the age-score space for the separate three years. The '×' and '○' marks represent the scores of the genuine user (first person in Fig. 1) and the impostors (the other three persons in Fig. 1), respectively. The dashed line shows the decision boundary of the baseline classifier. The dash-dot and bold lines denote the Q-stack decision boundaries using SVM-lin and SVM-rbf stacked classifiers.

(HTER) [13]. The decision boundary of the baseline classifier is decided with the minimum HTER. Indeed, as it is shown in Figure 4, the $Q-stack$ decision boundary does not deviate much from the baseline classification boundary given by the horizontal dot line.

Table 3 shows the verification performance of long-term 1100-day testing data set with successive periods of 180 days (0.5 year). The decision boundary of the baseline classifier is same as used in the training data set in Fig. 4. The decision boundaries of SVM-lin and SVM-rbf classifiers are obtained by training SVM-lin and SVM-rbf classifiers over first 100-day images of the training data set. It is found that the decision boundary (horizontal dashed line) of the baseline classifier is not effective any more and has to be adjusted in time (hard to be optimized with progressing age). On the other hand, the decision boundaries of SVM-lin and SVM-rbf classifiers in which the aging information is integrated are able to track the tendency between the scores and age progression, and provide significant improvements of classification accuracy in the evidence (score,age) space as opposed to the baseline classifier. In summary of the pilot research experiments using YouTube data, we can conclude that the use of age progression information as a metadata quality measure in the Q-stack classification scenario allows for improved classification in respect to the baseline classification results.

Table 2: Verification performance of different methods for the first 100 days in terms of false acceptance rate (FAR), false rejection rate (FRR) and half total error rate (HTER).

	FAR [%]	FRR [%]	HTER [%]
Baseline	2.33	5.00	3.67
SVM-lin	2.00	6.00	4.00
SVM-rbf	2.33	5.00	3.67

4.2 Experiments on MORPH data

The MORPH Database [5] is a publicly available database developed for investigating age progression. The images represent a diverse population with respect to age, gender, eth-

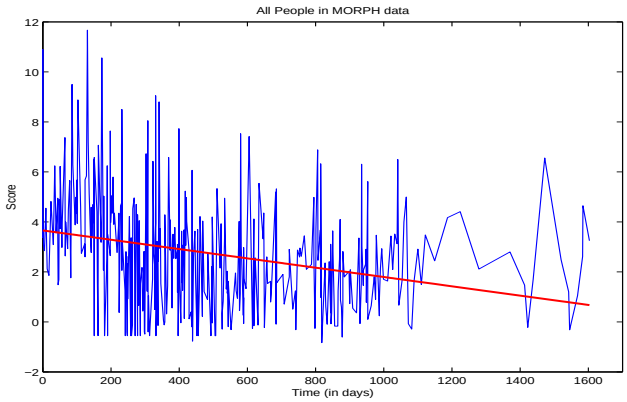

Figure 6: The influence of age progression on the scores for all the individuals in MORPH database. The bold line represents linear fitting of the variations of log likelihood with the age progression.

nicity, etc. These photos were taken between 1962 and 1998 but not taken daily. Therefore, there is different number of photos for each individual. The face images are not taken under controlled conditions. There are a lot of face images which are not frontal and in which the variations in head pose and facial expressions exist. For our studies on the age progression, we selected 14 individuals with more than 20 images for each individual and without the significant head pose and facial expression variations. The baseline GMM classifier is built over the first five images from all the 14 individuals using PCA features. Then we apply SVM-lin and SVM-rbf to the baseline GMM classifier.

Table 4: Verification performance with MORPH database (averaged over all the 14 individuals).

	FAR [%]	FRR [%]	HTER [%]
Baseline	0.05	30.47	15.76
SVM-lin	0.29	10.93	5.61
SVM-rbf	0.10	6.40	3.25

Figure 6 shows the influence of age progression on the classifier scores for all the individuals in MORPH database. It is shown that there exists general tendency of the classifier scores on the age progression: As the age increases, the scores generally decrease. Table 4 summarizes the verification performance with MORPH database in terms of FAR, FRR, and HTER, which are averaged over all the 14 individuals[1]. It is found that the verification performance in terms of HTER is improved by using Q-stack method.

5. CONCLUSIONS

In this paper, we presented a novel theoretical approach to incorporating age, based on the concept of metadata quality measure, into the face verification process, based on the concept of classifier stacking (Q-stack). We noticed in preliminary experiments with difficult real-world data of the MORPH database and the YouTube data recorded every day during 1200 days and baseline PCA classifier, that while class-nonspecific, the age metadata quality measure is causally linked to the classifier scores, which allows for increased long-term class-separation in the score-quality measure space using a short-term enrolment model. Obtained results show that indeed Q-stack is a powerful method of combining scores and age as metadata quality measure for improved classification. This will allow us in the near future for exhaustive experiments with the whole MORPH database of 515 individuals using combination of age information with different quality measures of face image and multiple baseline classifiers.

[1] Since the number of genuine and impostor images is different for each time period, we have not analysed the classification performance for each successive time period as for the YouTube data.

REFERENCES

[1] Ramanathan, N., Chellappa, R.: Face Verification Across Age Progression. IEEE Trans. Image Processing. **15** (2006) 3349–3361

[2] Patterson, E., Sethuram, A., Albert, M.,Ricanek, K.,King, M.: Aspects of Age Variation in Facial Morphology Affecting Biometrics. IEEE Conference on Biometrics: Theory, Applications, and Systems (BTAS 2007). Washington, D.C., USA, 27-29 Sept. 2007

[3] Scandrett, C., Solomon, J., Gibson, S.J.: A person-specific, rigorous aging model of the human face. Pattern Recognition Letters. **27** (2006) 1776–1787

[4] Flynn, P.: Biometrics databases. Chapter 25 in Jain, A. et al. (eds).: Handbook of Biometrics. Springer, New York, 2008, 529–548

[5] Ricanek, K., Tesafaye, T.: MORPH: A longitudinal image database of normal adult age-progression. 7th International Conference on Automatic Face and Gesture Recognition (FGR'06). April 2006, 341–345

[6] Kryszczuk, K., Drygajlo, A.: Improving classification with class-independent quality measures: Q-stack in face verification. 2nd International Conference on Biometrics (ICB 2007). Seoul, Korea, 27-29 August 2007

[7] Hicklin, A., Khanna, R.: The Role of Data Quality in Biometric Systems. White Paper. Mitretek Systems, February 2006

[8] Kryszczuk, K., Drygajlo, A.: Q-stack: uni- and multimodal classifier stacking with quality measures. International Workshop on Multiple Classifier Systems. Prague, Czech Republic, May 2007

[9] Wolpert, D.: Stacked Generalization. Neural Networks, **5** (1992), pp. 241-259

[10] Fukunaga K.: Introduction to Statistical Pattern Recognition (Second Edition), Academic Press, New York, 1990

[11] Schölkopf, B, Burges C. J. C., and Smola A. J.: Advances in Kernel Methods, Support Vector Learning MIT Press, Cambridge, 1999

[12] Theodoridis, S., Koutroumbas K.: Pattern recognition, Second Edition, Elsevier, 2003

[13] Bengio, S., Marcel, C., Marcel, S., Mariethoz, J.: Confidence measures for multimodal identity verification, Information Fusion, Volume 3, Number 4, December 2002 , pp. 267-276