Challenge: Face verification across age progression using real-world data

Gowri Somanath and Chandra Kambhamettu
Video and Image Modeling and Synthesis Lab, Department of Computer Science,
University of Delaware, Newark, DE. USA.
somanath,chandra@cis.udel.edu

1. Overview

Analysis of face images has been the topic of in-depth research with wide spread applications. Face recognition, verification, age progression studies are some of the topics under study. In order to facilitate comparison and benchmarking of different approaches, various datasets have been released. For the specific topics of face verification with age progression, aging pattern extraction and age estimation, only two public datasets are currently available. The FGNET and MORPH datasets contain a large number of subjects, but only a few images are available for each subject. We present a new dataset, VADANA, which complements them by providing a large number of high quality digital images for each subject within and across ages (depth vs. breadth). It provides the largest number of intra-personal pairs, essential for better training and testing. The images also offer a natural range of pose, expression and illumination variation. We demonstrate the difference and difficulty of VADANA by testing with state-of-the-art algorithms. Our findings from experiments show how VADANA can aid further research on different types of verification algorithms.

The following sections provide details for the proposed challenge. The dataset details, the need and motivation for its creation, comparison to existing benchmarks and the experiments performed on the same have been provided in the attached paper.

2. Problem definition and challenges

There are various problems in facial image analysis, such as face detection (finding faces in a given image), face recognition (matching new image to a known dataset), face verification (determine if a given unknown pair of face images belong to same person) and many others. In this work, we focus on face verification specifically in the case of age progression.

Problem definition: The input is a pair of facial images. The images are such that at least region from top of forehead till the chin is covered. Though in general, the images cover from top of head region and part of neck region also. The identity of the person(s) in the images is not known a priori. The system must determine if the two images belong to the same person (intra-personal pair or intra-pair) or to different persons (extra-personal pair or extra-pair). The two images are taken across a time period such that the age gap between the pair may range from 0 to 9 years. Also, the pose, expression and illumination is uncontrolled for both images.

Training setup: During the training phase, the system is provided with pair of images (both intra-pairs and extra-pairs). The age of the subject in a given image and thus the age gap between a pair is provided during training. A classifier is trained using the features from the images.

Testing setup: During the testing phase, the input is a pair of images. The subjects in these pairs are different from those in the training, i.e, the training and testing subjects are non-overlapping. There is no explicit age (or age-gap) information provided at this stage. The system must classify the pair as either intra-personal or extra-personal.

Applications: The above problem definition closely resembles various real-world application scenarios such as passport verification, security and surveillance matching in videos/image captured over a period of time, clustering of people in large datasets where identities are unknown and many others.

Challenges: The challenges stem from various aspects of the above problem definition: (1) The subject identities are not known, the system must therefore only rely on the information from the pair of images to determine the final classification. (2) The images are taken at different times, ranging from a gap of few months to up to 9 years (as in the case of passport verification). The effects due to aging thus contribute to shape and appearance changes even for an intra-pair (same person).

The shape changes are mostly seen in children (age below 18 years) and seniors, while appearance changes manifest from wrinkles, facial hair or accessories such as glasses. (3) Along with the above, the imaging conditions themselves lead to challenges from difference in pose, expression and illumination between the given pair. Thus, robust image processing techniques to account for the image-based challenges along with effective feature design to handle the first two are required in order to provide an effective solution.

3. Dataset summary and challenges

Details of the proposed dataset and its comparison to the existing benchmarks and protocols, is given in the attached paper. We will provide a summary of the need for a new dataset and its key features here. For the problem of face verification with age progression defined above, currently only two public datasets are available to the research community - FGNET [1] and MORPH [12, 11]. These datasets have provided the initial support and testing of the various algorithms devised for the problem [6, 3, 8, 2, 13, 12]. However, they lack in a few key respects:

1. Age distribution: The key aspect of this problem definition which distinguishes it from the general face recognition or verification problems is that of age progression. Also, it has been known that aging effects are different in children and adults [10]. In order to closely relate to the real-world, a benchmark must have large number of sample images for adults (age $>$ 18) and a good distribution of images in various age groups and age gaps. As shown in Table 2 in the paper, the widely used FGNET benchmark contains only 363 images of adults and Album 1 of MORPH has 1520. Through VADANA, we offer 1913 such images. The age group distribution is shown in Table 4 of the paper, and we can observe that VADANA offers orders of magnitude more images for the various groups. A sample of the age cross-section for a subject is shown in Figure 1(a).

2. Data quantity: As observed above, the training and testing input to such system is pairs of images (especially the number of intra-pairs). It is expected that learning and classification techniques can be made robust with the availability of large number of samples. It is also essential to test the scalability of the algorithms towards real-world applications. As detailed in Table 3 of the paper, FGNET only provides 1,164 intra-personal pairs for adults, while MORPH Album 1 offers 1,324. VADANA on the other hand offers 146,528 intra-pairs for adults. The further break down for the different age gaps is also provided to show the contrast with the current benchmarks.

3. Real-world variations: The MORPH dataset was composed of mugshot images, which mainly consist of frontal pose with some containing two profile views. This is useful to verify the algorithms for controlled setups but do not reflect real-world scenarios such as the security and surveillance applications. VADANA, from the nature of data collection, offers realistic variations within and across subjects and ages, as shown in Figure 1. The variations include pose changes (frontal to gradual profile views), expression variations (neutral, smile and others), difference in accessories (same subject with different glasses and without glasses) and illumination variations (indoor, outdoor, lighting direction and extent).

4. Image quality: With the increasing use of medium to high quality digital medium for recording videos and images in most environments, it is essential to test the algorithms on similar data quality. Both MORPH and FGNET are mainly composed of scanned and grayscale images. VADANA is composed of 2268 24-bit digital images and 30 scanned images.

5. Standard partitions and protocols: Though there is an increasing number of algorithms proposed, the source code is generally not available. Re-implementation can be difficult due to inadequate details in the paper and differences due to choice of various parameters. Standard partitions would make comparison fair and straightforward. FGNET and MORPH do not provide such fixed partitions, which lead some previous works to sample the database in weakly specified and non-repeatable fashion [6]. As detailed in the following sections, we therefore provide a set of fixed standard partitions to be used for comparing algorithm performances.

4. Protocols and experiments

Here we summarize the protocols and experiments on VADANA. The details are provided in Sections 3 and 4 of the paper.

4.1. Protocols

Following common practice, we divide the pairs into different age-gap bins ([0,2],[3,5] and [6,8]). For each age-gap bin, multiple 'sets' will be provided. Each 'set' is derived from different sampling of the available images in the dataset as detailed below (more details in paper):

For each set, we first fix the age gap bin. Then we fix a maximum limit on intra-pairs per subject. We then sample the dataset with the above conditions to obtain the intra-pairs. The pairs thus obtained are divided into folds. The folds are generated such that subjects are non-overlapping across folds and each fold has nearly equal number of intra-pairs. For each fold, corresponding extra-pairs are also generated. The number of extra-pairs equals the number of intra-pairs per fold. The details for the sets is provided in Table 1 (Table 6 in the paper). The different sets for the same age gap have been created to have less than 60% overlap (by random sampling of the available images and pairs per subject). For each age gap, the results are intended to be averaged across the sets. Performance is measured using the Equal Error Rate (EER) as done in most face analysis works. The following measures are computed to determine accuracy.

$$\text{Correct Acceptance Rate (CAR)} = \frac{\text{\# of correctly classified intra-pairs}}{\text{Total \# of intra-pairs}}$$

$$\text{Correct Rejection Rate (CRR)} = \frac{\text{\# of correctly classified extra-pairs}}{\text{Total \# of extra-pairs}}$$

Equal Error Rate (EER) is the error rate when CAR equals CRR, i.e, (1-CAR) or (1-CRR) when CAR=CRR. This CAR (or CRR) is the accuracy at EER and the accuracy for the algorithm. The experiment is run with different parameters of the algorithm and the CAR, CRR is obtained for each fold. For each parameter setting, the average CAR and CRR over the different folds is calculated. The ROC curve can be obtained by plotting these CAR against CRR for different parameter settings. The EER is obtained by using the point on ROC curve where CAR equals CRR.

Age gap	# of intra-pairs/fold X # folds	# of sets
[0,2]	$\approx$ 7000 X 5 folds	4
[3,5]	$\approx$ 240 X 3 folds	2
[6,8]	$\approx$ 145 X 2 folds	3

Table 1. Details of the experimental sets provided for uniform comparison. Note that each fold also has an equal number of extra-personal pairs, hence the total number of pairs per fold is double of that indicated above.

4.2. Experiments performed

For the experiments described in the paper, we have used the four sets for age gap [0,2]. This age gap is considered easiest for verification algorithms and hence helps get an approximate upper-bound on performance, since accuracy in general decreases with presence of larger aging effects [7]. It also establishes the difficulty level of VADANA with respect to other datasets. Also, we have used the aligned version of the dataset. We tested using three algorithms: (1) A baseline algorithm using eigenfaces approach with Random Forest based classifier, (2) state-of-the-art algorithm for face verification with age progression by Ling *et al.* [6], and (3) pair-matching using the algorithm by Nowak et.al [9]. The details for each algorithm, parameters used and the results are given the paper (Section 4: Experiment 1-3). As discussed in Section 5 of the paper, the above experiments provide insight into various aspects such as scalability of different schemes, contrasting trends of image difference and patch-based approaches and difficulty level of VADANA with respect to other datasets. A summary of the results is provided in 2 (Table 7 of paper).

	VADANA	FGNET [1]	LFW [4]	Jain [5]
Eigenfaces (PCA+RF)	52.33	99.33	-	-
Nowak pair matching (SIFT + ERFC) [9]	61.52	67 $\pm$ 2.2	73.0 $\pm$ 0.6	84.2 $\pm$ 0.31
Ling et. al (SVM+GOP) [6]	57.43	73.0	-	-

Table 2. Results from Experiments to compare VADANA with other datasets. The table shows accuracies at EER (see text).

5. Proposed benchmarking

5.1. Dissemination

Institutional Review Board (IRB) approval and subject consent has been taken for distribution and use of data for research purposes. The data will be distributed free of cost over the Internet after verification of the requesting group. The complete package will include the images and metadata (Table 5 in paper) and all the standard partitions (Table 6 in the paper). Code will also be provided to calculate the performance using the CAR, CRR and EER measures.

5.2. Future results

We wish to create a benchmark in the style of LFW face verification dataset and Middlebury stereo dataset, where authors of new algorithms can submit their results for the standard partitions. Results must be provided for at least one age-gap bin and all sets for that age-gap. The EER and the ROC curves provided will then be merged with currently available results. The performance of various algorithms can thus be easily compared in a fair and direct manner. The benchmark web page with the results will be maintained by our group.

References

[1] Fgnet aging dataset. face and gesture recognition working group 2000. 2, 3

[2] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(12):2234 –2240, dec. 2007. 2

[3] G. Guo, G. Mu, Y. Fu, and T. Huang. Human age estimation using bio-inspired features. pages 112 –119, jun. 2009. 2

[4] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *Faces in Real-Life Images Workshop in European Conference on Computer Vision (ECCV)*, 2008. 3

[5] V. Jain, A. Ferencz, and E. Learned-miller. Discriminative training of hyper-feature models for object identification. In *In British Machine Vision Conference*, volume 1, pages 357–366, 2006. 3

[6] H. Ling, S. Soatto, N. Ramanathan, and D. Jacobs. Face verification across age progression using discriminative methods. *Information Forensics and Security, IEEE Transactions on*, 5(1):82 –91, March 2010. 2, 3

[7] Y. M. Lui, B. D., D. B.A., J. Beveridg, G. Given, and P. Phillips. A meta-analysis of face recognition covariates. *IEEE 3rd International Conference on Biometrics Theory, Applications, and Systems*, pages 1 –8, 2009. 3

[8] G. Mahalingam and C. Kambhamettu. Face verification across age progression using adaboost and local binary patterns. *The 6th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, 2010. 2

[9] E. Nowak and F. Jurie. Learning visual similarity measures for comparing never seen objects. In *Conference on Computer Vision & Pattern Recognition*, jun 2007. 3

[10] J. B. Pittenger and R. E. Shaw. Aging faces as viscal-elastic events: Implications for a theory of nonrigid shape perception. *Journal of Experimental Psychology: Human Perception and Performance*, 1(4):374 – 382, 1975. 2

[11] A. Rawls and K. Ricanek. Morph: Development and optimization of a longitudinal age progression database, 2009. 2

[12] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression, 2006. 2

[13] J. Suo, S.-C. Zhu, S. Shan, and X. Chen. A compositional and dynamic model for face aging. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):385 –401, March 2010. 2