Benchmarking Face Alignment

Hua Gao, Hazım Kemal Ekenel

Institute for Anthropomatics
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: {gao, ekenel}@kit.edu

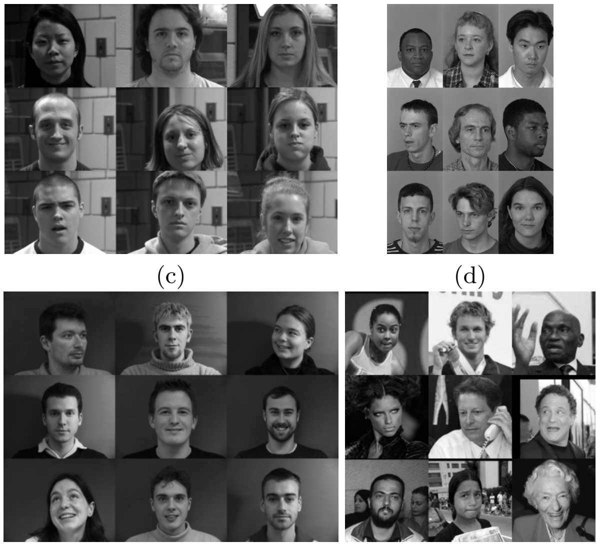
Abstract. Face alignment has been an active research area in computer vision community, since it provides an important basis for further analysis of facial properties such as identity, expression, gender, etc. for more than ten years. However, there exists no common evaluation benchmark so far which contains enough variations for evaluating both accuracy and robustness of alignment algorithms. We propose an experimental setup that includes multiple datasets containing different level of varieties. Annotation of landmarks are provided and two widely used evaluation metrics are suggested. We believe that this will provide a common platform for benchmarking face alignment.

1 Introduction

Numerous works have been conducted to solve the face alignment problem such as the active shape models[1], active appearance models [2], constrained local models [3] and many of their extensions [4–8], due to its importance in a wide range of applications, such as analysis of expression, pose direction, gender, age and identity of human faces. Collecting databases for benchmarking face alignment is an expensive task as labeling facial landmarks on face images is a tedious and time-consuming work. There are some databases available so far, such as the IMM database [9] and the XM2VTS [10] database, which are annotated with 58 and 68 landmarks respectively. However, the number of subject in the IMM database is very limited while the variation in the XM2VTS database is not sufficient.

The goal of this proposal is to provide an experimental setup and proper metrics for evaluating different face alignment systems. We focus on aligning faces in 2D still images which are captured with monocular cameras, since in real-world applications monocular cameras are most commonly used and their prices are cheaper than other configurations. We utilize four different publicly availabe face databases for the experiments. The collection includes different variations in pose, illumination, expression, occlusion, etc. These variations enable us to analyze different generalization capabilities on different level of variations.

The following two sections describe the proposed experimental setup and the suggested evaluation protocol in detail.



(c) (d)

(a) (b)

Fig. 1. Example of the face dataset: (a) FRGC v2.0 database, (b) FERET database, (c) IMM database, and (d) LFW database.

2 Data

To evaluate the accuracy and robustness of the face alignment algorithms, we collected a set of 1529 images from multiple publicly available databases, including the FRGC v2.0 database [11], the FERET database [12], the IMM database [9], and the Labeled Faces in the Wild (LFW) database [13]. Figure 1 shows sample images from these four databases. We partition all images into four distinct data sets. Table 1 lists the properties of each database and partition. Set 1 includes 400 images (one image per subject), where 200 images are from the FRGC database and the other 200 images are from the FERET database. Set 1 is used as the training set. Set 2 includes 389 images from the same subjects but different images as the FRGC database in Set 1. Set 3 includes 240 images from 40 subjects in the IMM database that were never used in the training. Set 4 includes randomly selected 500 images of 500 subjects from the LFW database. This partition ensures that we have two levels of generalization to be tested, i.e., Set 2 is tested as the unseen data of seen subjects; Set 3 and 4 are tested as the unseen data of unseen subjects. Set 4 is a particular challenging dataset since it is collected from the Internet. The images were captured under cluttered background and various real-world illumination environments using different types of cameras.

There are 58 manually labeled landmarks for each of the 1529 images. Figure 2 shows an example of annotated image. The annotation format follows the one proposed in [14], which also describes the contours of the face components using closed or open paths.

	FRGC	FERET	IMM	LFW
Images	589	200	240	500
Subjects	200	200	40	500
Variation	Frontal, expression	Pose	Pose, expression	All
Set 1	200	200		
Set 2	389			
Set 3			240	
Set 4				500

Table 1. Summary of the datasets.

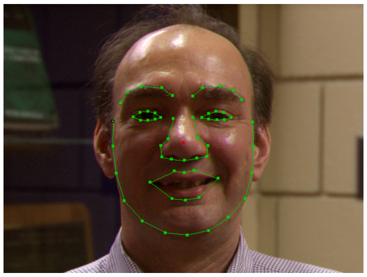


Fig. 2. Example of an annotated image.

3 Protocol

As stated in Section 2, Set 1 will be used as training data, and the other three sets will be used as testing sets. For propoer comparison, we do not suggest automatic initialization to avoid additional errors introduced by automatic face detection or eye localization. Instead, the randomly perturbed ground truth landmarks will be used as initial condition for starting an alignment. The procedure is repeated multiple times on each image of the testing set in order to perform a statistical evaluation of the result. The initial position of the landmarks is generated by perturbing the shape parameter with white Gaussian noise at different noise level. An alignment is claimed as converged if the Root Mean Square Error (RMSE) between the aligned landmarks and the ground truth is less than τ. The alignment robustness and accuracy is assessed by computing: (a) the Average Frequency of Convergence (AFC), given by the number of trials where the alignment converges divided by the total number of trials; and (b) the histogram of the RMSE (HRMSE) of the converged trials, which measures how close the aligned landmarks are to the ground truth. The evaluation metrics were widely adopted in this research domain [4–7]. As the face images in different sets might have various scales, the RMSE is normalized by dividing the eye distance.

The runtime complexity of the evaluated alignment system and hardware on which the experiments are conducted should also be provided.

4 Conclusion and Future work

We have presented an experimental setup for evaluating different level of generalization for face alignment. Four publicly available face databases are utilized for the experiments. Annotation of the landmarks will be provided as well. We suggested two evaluation metrics to evaluate both accuracy and robustness of the alignment algorithm. We believe that this experimental setup will provide a useful platform for comparing and analyzing different alignment approaches.

References

1. Cootes, T., Taylor, C., Cooper, D., Graham., J.: Training models of shape from sets of examples. In: Proc. of BMVC. (1992) 9–18
2. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Proc. of 5^{th} European Conference on Computer Vision. Volume 2. (1998) 484–498
3. Cristinacce, D., Cootes, T.: Feature detection and tracking with constrained local models. In: 17^{th} British Machine Vision Conference, Edinburgh, UK. (2006) 929–938
4. Liu, X.: Discriminative face alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence **31** (2009) 1941–1954
5. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision **60**(2) (2004) 135–164
6. Liang, L., Xiao, R., Wen, F., Sun, J.: Face alignment via component-based discriminative search. In: Proc. of the 10th European Conference on Computer Vision. (2008) 72–85
7. Saragih, J., Goecke, R.: A nonlinear discriminative approach to AAM fitting. In: Proc. of ICCV. (2008) 1–8
8. Zhou, Y., Gu, L., Zhang, H.: Bayesian tangent shape model: Estimating shape and pose parameters via bayesian inference. In: Proc. of IEEE CVPR. Volume 1. (2003) 109–116
9. Stegmann, M., Ersboll, B., Larsen, R.: FAME - a flexible appearance modeling enviroment. IEEE Trans. Medical Imaging **22**(10) (2003) 1319–1331
10. Messer, K., Matas, J., Kittler, J., J.Luettin, Maitre, G.: XM2VTS: The extended M2VTS database. In: AVPBA. (1999)
11. Phillips, P., Flynn, P., Scruggs, T., Bowyer, K., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: Proc. of CVPR. (2005) 947–954
12. Phillips, P., Moon, H., Rauss, P., Rizvi, S.: The feret evaluation methodology for face recognition algorithms. IEEE Trans. on PAMI **22**(10) (2000) 1090–1104

13. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (October 2007)

14. Stegmann, M.B.: Analysis and segmentation of face images using point annotations and linear subspace techniques. Technical report, Informatics and Mathematical Modelling, Technical University of Denmark, DTU (aug 2002) IMM-REP-2002-22.