Facial Landmarking: Comparing Automatic Landmarking Methods with Applications in Soft Biometrics

Amrutha Sethuram, Karl Ricanek, Jason Saragih, and Chris Boehnen

Face Aging Group, UNCW, U.S.A.
{sethurama,ricanekk}@uncw.edu, jason.saragih@csiro.au, boehnencb@ornl.gov
http://www.faceaginggroup.com

Abstract. Registration is a critical step in computer-based image analysis. In this work we examine the effects of registration in face-based soft-biometrics. This form of soft-biometrics, better termed as facial analytics, takes an image containing a face and returns attributes of that face. In this work, the attributes of focus are gender and race. Automatic generation of facial analytics relies on accurate registration. Hence, this work evaluates three techniques for dense registration, namely AAM, Stacked ASM and CLM. Further, we evaluate the influence of facial landmark mis-localization, resulting from these techniques, on gender classification and race determination. To the best of our knowledge, such an evaluation of landmark mis-localization on soft biometrics, has not been conducted. We further demonstrate an effective system for gender and race classification based on dense landmarking and multi-factored principle components analysis. The system performs well against a multi-age face dataset for both gender and race classification.

Keywords: facial landmarking, auto-landmarking methods, gender classification, race determination.

1 Introduction

Automatic facial landmarking is a very important step that precedes any task involving face recognition or analysis. These landmarks, also referred to as fiducial points or anchor points, are used for accurate registration of faces and have a significant effect on the impending analysis. While some applications, such as, face recognition and tracking use a few landmarks like the eye and eyebrow corners, centers of the iris, corners of the mouth, tip of the nose and chin for registration, other applications like age estimation, expression analysis, detection of intent or facial aging require a greater number of landmarks for analysis. Further, it is important that the detection of these points be accurate and robust to environmental variables e.g. illumination, occlusion, expression, pose, etc. It has been shown that precise landmarks are essential for face-recognition performance and that more landmarks results in higher recognition performance. However, under various conditions of image acquisition, automatic facial landmarking becomes

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part II, LNCS 7584, pp. 290–299, 2012.

a challenging task. There exist many automatic registration algorithms which perform well in detecting the internal landmarks e.g., the eye corners, centers of the iris and the corners of the mouth. More often, performance of these algorithms are reported solely on a few such, well-defined points, which artificially inflate the performance of algorithms in a real-world application. However, in this paper, we consider an extremely dense scheme consisting of 252-points on the face that includes internal and boundary points. Such a dense scheme is helpful in applications in soft biometrics such as expression recognition, detection of micro gestures, age estimation, gender and race classification, facial aging etc. While one can argue that accurate detection of a few internal points is sufficient for many tasks, it is often necessary that features such as the boundary of the face must be accurately detected for applications that involve facial synthesis or off-pose face recognition. The main contributions of this paper are: (1) Provide a set of baseline algorithms and performance metrics for extremely dense registration. (2) Evaluate the influence of automatic detection of landmarks on gender classification and race determination. The remainder of this paper is organized as follows: A background on existing automatic landmarking methods and those considered for this work is discussed in section 2. Experiments and Results are presented in section 3. Conclusion and future work is discussed in section 4.

2 Background

Algorithms proposed for automatic facial landmarking can be broadly classified into two categories: image-based methods and structure-based methods. In image-based methods, faces are treated as vectors in high dimensional space, which are then modeled as a manifold. The variability in facial features is captured through popular transformations like the Principal Components Analysis, Independent Components Analysis, Gabor Wavelets, Discrete Cosine Transforms and Gaussian derivative filters. The appearance of each landmark is then learned through the use of machine learning approaches like support vector machines, boosted cascade detectors and multi-layer perceptrons [1] [2][3].

Structure-based methods use prior knowledge about facial landmark positions, and constrain the landmark search using heuristic rules that involve angles, distances and areas. Very popular methods in this category are Active Shape Models (ASM) [4], Active Appearance Models (AAM) [5] and Elastic Bunch Graphing Methods. ASMs model textures of small neighborhoods around landmarks and iteratively minimizes the differences between landmark points and their corresponding models. The AAM typically looks at the convex hull of landmarks, synthesizes a facial image from a joint appearance and shape model, and seeks to minimize similarity to the target face iteratively. Cristinacce et al. [6] proposed the Constrained Local Model (CLM) approach that uses a set of local feature templates for detection of landmarks. There have been many variants and improvements on these classic ASM, AAM and CLM approaches. By fitting more landmarks and stacking two ASMs in series, Milborrow and Nicolls [7] locate features in frontal views of faces. In [8], Saragih *et al.* propose a regularized mean-shift algorithm to the CLM approach. Due to their widespread

use in various automatic landmarking applications, the availability of their actual implementations and also flexibility to train the algorithms using our dense scheme, the following three automatic landmarking methods are considered for this work.

Active Appearance Models: Active appearance models (AAM), a group of flexible deformable models, have been widely used for automatic landmarking. First proposed by Cootes et. al [5], AAM decouples and models shape and pixel intensities of an object. As described in [5], the AAM model can be generated in three main steps: (1) A statistical shape model is constructed to model the shape variations of an object using a set of annotated training images. (2) A texture model is then built to model the texture variations, which is represented by intensities of the pixels. (3) A final appearance model is then built by combining the shape and the texture models. The AAM software used for this work was obtained from [9].

Constrained Local Models: Constrained Local Models are a derivative of Active Shape Models (ASM). These are methods that make independent predictions regarding locations of the model's landmarks, which are combined by enforcing *a prior* over their joint motion. Most CLM variants implement a two step fitting strategy, where an exhaustive local search is first performed to obtain a response map for each landmark. Optimization is then performed using strategies to maximize the responses of the landmarks. In this paper, one such strategy that uses a gaussian prior over the model's PCA parameters is implemented as an automatic method to be compared against. The algorithm is itself presented by Saragih et al. in [8]

STASM: STASM or Stacked Active Shape Model [7] is an extension of active shape model proposed by Cootes [4]. As described in [7], STASM extends the active shape model by fitting more landmarks than actually needed, by selectively using two-instead of one-dimensional landmark templates and stacking two active shape models in series. The C++ software library to train and test the models for this work was obtained from [10]

3 Experiments and Results

3.1 Design of Experiments

Databases Used The automatic landmarking algorithms were trained using images obtained from MORPH [11] (a publicly available mugshot database), the PAL database [12] and the Pinellas County database (mugshot database with limited distribution). In addition, the data was formulated with distribution of images over four ethno-gender groups: African American Male (AAM), African American Female (AAF), Caucasian American Male (CAM) and Caucasian American Females (CAF), and age ranges as shown in Table 1 and Table 2 for training the general model. A total of 1155 images were used for training.

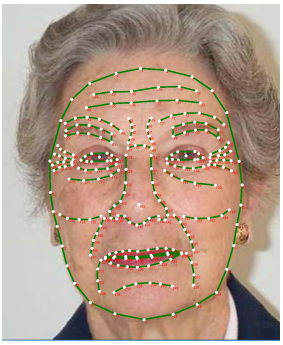

Fig. 1. Example of dense annotation map

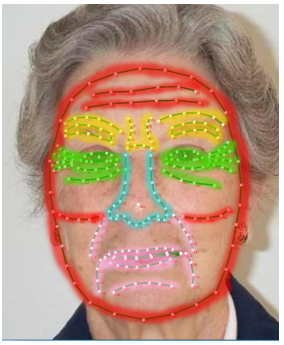

Fig. 2. Component scheme adopted for landmarking. Each expert was trained on annotating a color coded feature

Table 1. Training Data: Distribution based on ethno-gender groups and age ranges

Age Range	AAM	AAF	CAM	CAF
18-30	50	50	50	50
31-40	50	50	50	50
41-50	50	50	50	50
51-60	50	50	49	50
61-70	50	50	48	50
71+	38	21	49	50

Table 2. Training Data: Distribution based on databases

Database	AAM	AAF	CAM	CAF
Pinellas	88	62	104	180
MORPH	200	199	149	67
PAL	0	10	43	53

Obtaining Ground-Truth Data Obtaining ground truth coordinates for points on an image is not only tedious but also a time-consuming task. Great effort must be taken to get the annotators to consistently locate and label the right features on the image. In addition, to account for inter-observer variability in obtaining gold standard points requires that the images be annotated by at least 3 or more trained annotators for each image [13]. Thus it has become customary for researchers to report the performance of their algorithms on datasets that have ground truth as part of their distribution. However, since this work is based on a dense 252 landmark scheme for the face, which is not available in any face datasets, it was necessary that we generated the ground truth in-house. Given the large number of training and testing data and the density of the annotation scheme itself, it was not practical, both in terms of time and resources, to obtain repeat measurements for the ground truth data by 4 or 5 well-trained annotators. Instead, a component scheme was developed, in which, experts were

trained to annotate a specific region or component of the face. Each expert annotated a specific feature/component of the face as color coded in Figure 2 across all the images in the training and testing data. This ensured that variations in annotating features of the face were kept at a minimum.

Training and Testing Methodology. For purposes of evaluation and comparison, each of the automatic landmarking methods were trained on the entire set of 1155 images, which will be henceforth termed as *general* model.

Similar to the collection of training data, testing data from the various ethno-gender groups i.e. African American Females (AAF), Caucasian American Males (CAM), African American Males (AAM) and Caucasian American Females (CAF) were formulated to evaluate the performance of the algorithms. In addition to the 1155 images that were manually annotated to train the models, ground truth for test images for the AAF, CAM, AAM and CAF ethno-gender groups were obtained using the same component scheme. The distribution of test images used in this work is as shown in Table 3 and Table 4.

For testing the performance of the landmarking algorithms, each of the algorithms were trained on the general model and automatic landmarks were obtained for the general test set. These detected landmarks were then compared to the ground truth that was generated for the testing data.

Table 3. Testing Data: Distribution based on ethno-gender groups and age ranges

Age Range	AAF	CAM	AAM	CAF
18-30	50	50	50	50
31-40	44	49	50	50
41-50	41	50	50	50
51-60	49	50	50	50
61-70	11	50	43	50
71+	0	50	2	26

Table 4. Testing Data: Distribution based on databases

Database	AAF	CAM	AAM	CAF
Pinellas County	103	240	140	244
MORPH	80	44	101	6
PAL	12	15	4	26

3.2 Performance Measures

In our experiments, we evaluate the efficiency of the AAM, CLM and STASM landmarking algorithms in a two step process. First, the actual error in point detection is evaluated on all of the 252 points,by comparing the points detected by the algorithms with the ground-truth available on the testing data. Next, the influence of these detected landmark points is quantified when applied to gender classification and race determination as shown in Figure 3

Error Analysis of Automatic Point Detection. The efficiency of the algorithms are evaluated by analyzing the errors associated with the detected points

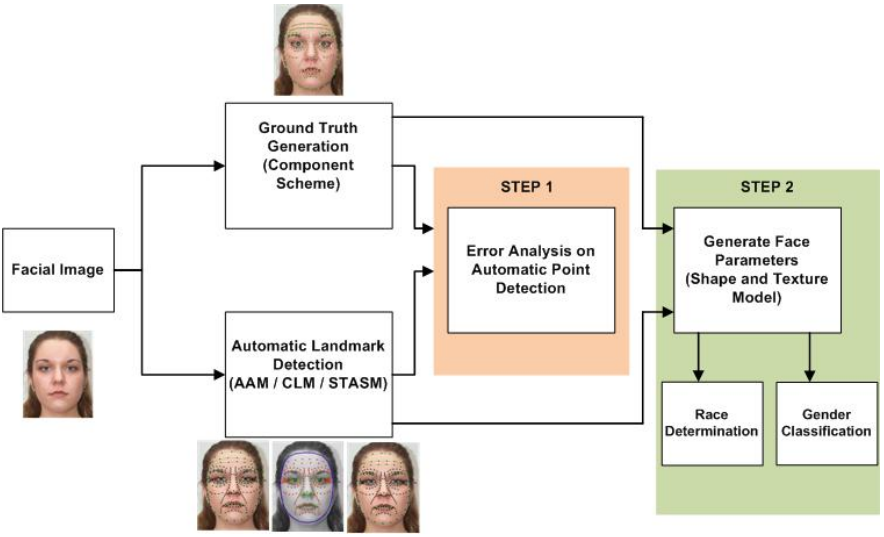

Fig. 3. Evaluation of automatic landmarking methods

when compared to the ground-truth.The interocular distance d_{io} is used as a normalization factor for computing error measures. Interocular distance is the distance between the centers of left and right eye and is often used in state-of-the art studies in 2-D facial landmarking. Since the distance error measure is scaled by the interocular distance, it is invariant to the variation in size of each individual face, which allows scaled comparison of point to point errors between images. The interocular distance varied between 48.8 and 130.11 pixels for the test data.

The detection error of a point i is defined as the Euclidean point to point distance between the ground-truth point T_i and detected point $\hat{T}_i$:

$$e_i = \frac{||T_i - \hat{T}_i||}{d_{io}} \tag{1}$$

An *average error per annotation point* p, across all the images in the test set was computed using the formula

$$m_p = \frac{1}{md_{io}} \sum_{i=1}^{m} d_i \tag{2}$$

where d_i are the Euclidean point to point errors for each individual annotation point and m is the number of images in the test data set. The average error for each of these individual points is as shown in Figure 4 for the test data.

The classification rate C_i can be defined as:

$$C_i = \frac{\sum_{j=1}^{m} e_i^j < 0.1}{m} \tag{3}$$

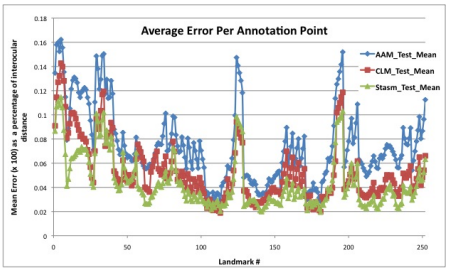

Fig. 4. Plot of average error per annotation point as a percentage of interocular distance vs. landmark

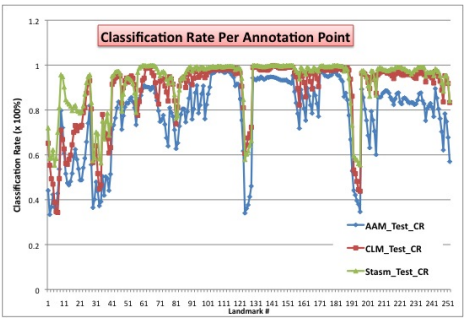

Fig. 5. Plot of average error per annotation point as a percentage of interocular distance vs. landmark

where j is the image number and m is the total number of images in the dataset. The classification rate for the test data is as shown in Figure 5.

The average image error on an image I can be defined as

$$m_I = \frac{1}{nd_{io}} \sum_{i=1}^{n} d_i \tag{4}$$

where n denotes the number of landmarks (252) and d_i values are the Euclidean point-to-point distances for each individual landmark location. The average image error for the different algorithms for the general model is as shown in Figure 6.

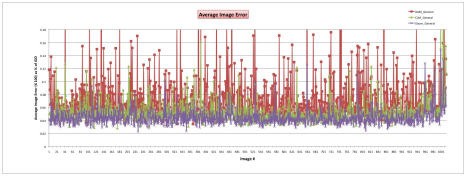

Fig. 6. General Test Data: Average Image Error

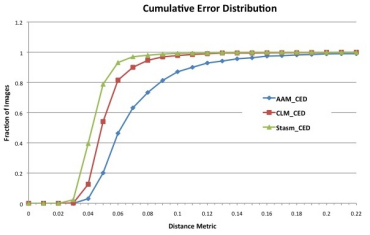

Fig. 7. Comparison of cumulative error distribution of point to point error

The cumulative error distribution of point to point error measured on the test set is as shown in Figure 7.

Also, to compare the errors associated with the individual physical features of the face, average errors were computed on the set of landmarks that make up the physical feature. Results are as shown in Table 5 for the general model.

Table 5. General Model: Average Error Associated With Physical Features (as a percentage of interocular distance)

Algorithm	Face Outline	Eyebrow (L)	Eyebrow (R)	Eye (L)	Eye (R)	Nose	Mouth	Soft Tissue
AAM	12.1	7.9	7.0	3.6	3.3	4.4	6.3	8.2
CLM	9.2	5.5	**5.0**	**2.5**	**2.4**	3.2	3.9	5.7
STASM	**7.2**	**5.2**	5.3	2.6	2.6	**2.7**	**2.9**	**4.7**

In the analysis of the average image error on our data, a Kruskal-Wallis H test was performed. It was found that there was a statistically significant difference between the different algorithms ($H(2) = 980.88$, $P < 0.01$) with a mean rank of 2151.69 for AAM, 1480.72 for CLM and 932.09 for STASM. Since STASM and CLM have relatively lower mean errors, their performance is better than AAM. From Figure 4, it can be seen that STASM has the lowest average error per annotation point. From Figure 5, it is seen that the average error per annotation point is comparable in both CLM and STASM on all points other than the boundary points on the face. In addition, the average image error on each of the training images were lower for STASM and CLM as shown in Figure 6. Also, Figure 7 shows that STASM has a higher fraction of images on which the error was below 10% of the interocular distance. At 10% of the inter ocular distance, STASM and CLM perform better than the AAM. It is evident from Table 5 that, STASM and CLM have lower errors on individual facial features when compared to the AAM. It can be seen from Figure 4 and Table 5, that errors associated with points on the soft tissue are larger than the errors associated with points on an actual facial feature e.g. left eye, right eye, nose and the mouth. This may be also due to the fact that it is a challenging task for the algorithms to detect the points when the actual trait of the soft tissue e.g. creases on the forehead, nasiolabial lines or the crows feet, is not very well defined across all faces, i.e. these creases and lines are not apparent on young faces. Future experiments will delve into quantifying the errors on older faces. The primary errors are found on the face boundary. Further, the left face boundary–chin to the ear–generates the most errors. This is helpful for applications where automatic detection of the outline of the face is also important, in addition to the internal features. Finally, although the time taken to train the models and time to automatically detect these landmarks were not quantified directly from the experiments performed, it was observed that STASM and CLM have much lower training and detection times when compared to the AAM. This suggests that STASM and CLM are better methods for automatic landmarking on realtime application, e.g. face tracking on video or large-scale batch analysis of faces, which may be executed in cloud based systems.

Application to Soft Biometrics. In practice, automatic landmark detection leads to mis-localization of a few annotation points. An evaluation of the influence of this mis-localization was performed on gender classification and race

determination. For each of these classifiers, an active appearance model (AAM) was trained on the set of training images and the corresponding combined appearance parameters were used to train a Support Vector Machine (SVM). All face parameters, except the four, which explains rotation, translation, and scale, were used to train each (gender and race) classifier. The trained SVM classifiers were then used to determine the gender and race of the images in the hold out testing data. The results are as shown in Table 6 and Table 7 below. It can be seen that, performance of CLM and STASM on gender and race classification are comparable and better than AAM. Also, comparing the difference between the efficiencies of classifiers on ground truth and auto-detected points, gender classification is more sensitive to automatic landmarking errors than race determination. An example of an image in which race was falsely misclassified to be caucasian by all the three algorithms is as shown in Figure 8. Similarly, an example in which the gender was misclassified to be a female by all the three algorithms is as shown in Figure 9. In both the examples the classification was correct, when the face parameters were generated from the groundtruth points.

Table 6. Gender Determination Classification Accuracy (%)

Ground Truth	AAM	CLM	STASM
96.8	83.3	85.9	85.8

Table 7. Race Classification Accuracy (%)

Ground Truth	AAM	CLM	STASM
99.2	96.5	97.9	98

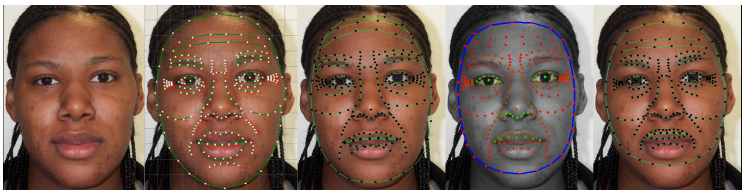

Fig. 8. Race Classification: Source Image, Ground-truth Annotations (✓), AAM detected(X), CLM detected(X), STASM detected points(X)

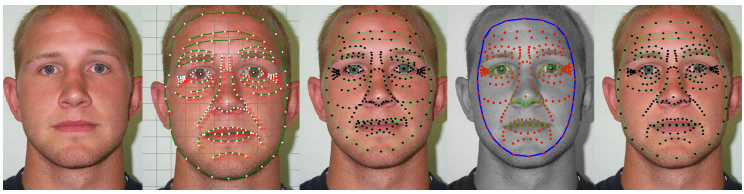

Fig. 9. Gender Determination: Source Image, Ground-Truth Annotations (✓), AAM detected(X), CLM detected(X), STASM detected points(X)

3.3 Conclusion and Future Work

In this paper, three state-of-the art algorithms - AAM, CLM and STASM, are compared for automatic landmark detection on a set of dense landmarks on the

face. The performance of each of the algorithms are evaluated both in terms of actual errors in landmark detection and consequently by their ability in gender classification and race determination. A particular strength of the CLM algorithm is its performance in the presence of occlusions. Future evaluations will include performance of these registration techniques in the presence of noise and occlusions. From the experiments that were conducted, it can be concluded from this paper that: (1) CLM and STASM are better algorithms to be used for automatic landmarking on a dense landmarking scheme in terms of accuracy, time taken to train the model and detection of landmarks. (2) The influence of mis-localization of annotation points, resulting from automatic landmarking algorithms, is pronounced on race determination and gender classification. The efficiency of the classifiers decreases with the degree of inaccuracy in the landmarks detected. However, while manual landmarks are still the best for gender and race classification, automatic detection algorithms such as STASM and CLM are viable surrogates.

Acknowledgments. This work was partially funded by ongoing efforts with National Institute of Justice, Oakridge National Labs, and Federal Bureau of Investigations Biometric Center of Excellence.

References

1. Viola, P., Jones, M.: Robust real-time object detection. International Journal of Computer Vision (2001)
2. Vukadinovic, D., Pantic, M.: Fully automatic facial feature point detection using gabor feature based boosted classifiers. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 2, pp. 1692–1698 (2005)
3. Dibeklioglu, H., Salah, A., Gevers, T.: A statistical method for 2-d facial landmarking. IEEE Transactions on Image Processing 21, 844–858 (2012)
4. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Computer Vis. and Image Under. 61, 38–59 (1995)
5. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Proc. Eur. Conf. Comput. Vis., vol. 2, pp. 484–498 (1998)
6. Cristinacce, D., Cootes, T.: Feature detection and tracking with constrained local models. In: BMVC, pp. 929–938 (2006)
7. Milborrow, S., Nicolls, F.: Locating Facial Features with an Extended Active Shape Model. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 504–513. Springer, Heidelberg (2008)
8. Saragih, J., Lucey, S., Cohn, J.: Deformable model fitting by regularized landmark mean-shift. International Journal of Computer Vision 91, 200–215 (2011)
9. AAM, `http://sourceforge.net/projects/asmlibrary/files/`
10. Stasm, `http://www.milbo.users.sonic.net/stasm/download.html`
11. Ricanek, K., Tesafaye, T.: Morph: A longitudinal image database of normal adult age-progression. In: 7th Int. Conf. on Auto. Face and Gesture Recog., pp. 341–345 (2006)
12. Minear, M., Park, D.: A lifespan database of adult facial stimuli. Behavior Research Methods, Instruments and Computers: A Journal of the Psychonomic Society, Inc. 36, 630–633 (2004)
13. Vučinić, P., Trpovski, Ž., Šćepan, I.: Automatic landmarking of cephalograms using active appearance models. European Journal of Orthodontics 32, 233–241 (2010)