Comparative Analysis of Decision-level Fusion Algorithms for 3D Face Recognition

Berk Gökberk, Lale Akarun
Boğaziçi University, Computer Engineering Dept.
Istanbul, TURKEY
{gokberk,akarun}@boun.edu.tr

Abstract

3D shape-based face recognition algorithms can be improved by using decision-level fusion algorithms. In this work, we present a comparative analysis of various fusion algorithms, and also propose novel ones. The contributions of the paper can be summarized as: i) a through analysis of several decision-level fusion algorithms, ii) a dynamically estimated reliability-assisted fusion schemes, and iii) a novel implementation of LDA-based cascaded serial fusion algorithm. Experiments conducted on the `3D_RMA` *dataset confirm that serial fusion offers the best solution, and dynamic calculation of reliability estimates improves the accuracy of the standard fusion schemes.*

1. Introduction

Human face is the most popular non-intrusive biometrics. However, in uncontrolled environments where illumination, pose and expression variations present, identification performance degrades quickly. To offer feasible solutions in such cases, researchers have focused on different modalities other than 2D image-based representations. Among these modalities, 3D facial shape is considered to be the most promising one in high-security scenarios.

The growing trend in 3D facial biometrics is surveyed in a recent article [3] where the systems are analyzed according to the information they use. The first group uses only 3D facial shape information whereas the second group uses texture information additionally. In this paper, we focus on the methods of the first approach. There are two main trends in this approach: the use of point clouds, and the use depth images. Point clouds are obtained by the raw 3D measurements from a 3D sensor whereas depth images are formed by projecting the 3D point cloud data to a 2D image according to z-depths. In each category, prior to matching, face normalization and registration should be carried out.

In *point cloud-based systems*, faces are generally registered via the Iterative Closest Point (ICP) [1] technique. An example is explained in [10] where coarse alignment is accomplished by automatically located three facial landmarks, and fine registration is done by the ICP. The surface matching error produced by the ICP is used as a dissimilarity value. Extensions of the ICP-based recognition systems for non-rigid deformations can be found in [4] and [7]. In *depth image-based approaches*, popular trend is to use statistical feature extraction methods such as Principal Components Analysis (PCA) [5]. The advantage of depth images is the ease of fusion with texture. Although fusion of 2D texture and 3D shape modalities are commonly used, their contribution in solely shape-based approaches is often not addressed. Examples of shape-based fusion algorithms can found in [11] and [6]. In [11], ICP-based surface matcher is fused with a profile-based classifier. A more elaborate analysis is conducted in [6] where five shape-based classifiers are combined at the rank-level. Their results confirm the validity of fusion in the shape domain.

In this work, we extend the analysis proposed in [6] with more advanced fusion schemes by comparing the benefits of abstract-level, rank-level and score-level fusion methods. We also propose new fusion schemes with the help of a novel estimation of classifier reliability at the identification phase in an online manner. The paper is organized as follows: Section 2 explains the individual 3D face classifiers. In Section 3, fusion methodologies are presented. Experimental results are provided in Section 4, and we conclude in Section 5. [1]

2. 3D Face Classifiers

We have developed four different 3D shape-based face classifiers. All of them requires a dense point-to-point correspondence between gallery and probe images. To obtain correspondences, we use a variant of the ICP technique,

---

[1] This work is supported by TÜBİTAK grant 104E080.

referred to as RegICP. RegICP finds a mapping from a given face to an average face model (AFM). This mapping can later be used to estimate dense correspondence between any two face. In the coarse alignment step, we locate the nose tip, and then use the ICP method to register *F* to the AFM. For each point in the AFM, we search for its nearest neighbor in *F*. After RegICP, each face *F* in the gallery and probe sets is densely registered to the AFM.

In *point cloud-based classifier* (PC), let Φ_i be a 3D face of the i^{th} individual. We can represent Φ_i as $\Phi_i^P = \{p_1^i, p_2^i, ...p_m^i\}$, where p^is are the coordinates of each point in the face and m is the number of points. Since the correspondence algorithm produces an ordering of facial points, we define the distance between two faces as: $D(\Phi_i^P, \Phi_j^P) = \sum_{k=1}^{m} ||p_k^i - p_k^j||$, where $||.||$ is the Euclidean norm. This distance function can be viewed as a discrete approximation of the volumetric difference between two facial surfaces.

In *depth image-based classifier* (LDA-DI), 3D points are projected to 2D gray-scale images. We have applied both PCA and Linear Discriminant Analysis (LDA) to the depth images. However, in the rest of the paper, we exclude the PCA results since it obtains significantly worse accuracy than LDA-based approach.

In *profile-based classifier* (PRO), we automatically locate seven equally spaced vertical profile curves. First the central profile of the AFM is found by performing principal direction analysis at the points around the nose tip. After locating central profile, three left and right lateral profiles are found on the AFM. By using correspondence, these profile curves can be easily found for each face. The distance between faces Φ_i and Φ_j is defined as the sum of the z-depth distances between each corresponding profile curve.

Shape index-based classifier (SI) is based on the curvature analysis at each facial point in the 3D space. Shape index is defined as by: $S = \frac{1}{2} - \frac{1}{\pi}\arctan(\frac{\kappa_1+\kappa_2}{\kappa_1-\kappa_2})$ where $\kappa_1 > \kappa_2$. κ_1 and κ_2 are the *principal curvatures*. S has a range of $[0, 1]$. 3D faces can be represented by the shape index values at each vertex: $\Phi_i^S = \{s_1^i, s_2^i, ..., s_m^i\}$ where s^is are shape index values. The distance between two faces in the shape-index representation is calculated by $D(\Phi_i^S, \Phi_j^S) = \sum_{k=1}^{m} ||s_k^i - s_k^j||$.

3. Fusion Techniques

Information fusion techniques are commonly used in pattern classification if there are multiple ways to solve a given problem [9]. Here, we consider only decision level fusion techniques: i) *abstract-level fusion*: classifiers output just the nearest class label, ii) *rank-level fusion*: classifiers produce a rank list of class labels in decreasing order of similarity, and iii) *measurement-level fusion*: classifiers generate similarity scores. The fusion methods that are studied in this work are as follows:

Consensus Voting (`CV`): Assume that each pattern classifier outputs the nearest class label. In *consensus voting*, the class having the highest vote is declared as the final opinion. Ties are broken arbitrarily to arrive at a decision.

Borda Count (`BC`): As a rank-level fusion technique, *Borda Count* method calculates the combined ranking by calculating the sum of the ranks of each class assigned by the individual classifiers. As a final opinion, the class having the smallest total rank is reported.

`SUM` and `PRODUCT` Rules: Sum rule computes the combined class scores by: $y_j = \sum_{k=1}^{K} s_{kj}, j = 1, \ldots, M$ for K classifiers for an M-class classification task where s_{kj}'s are scores. Similarly, product rule obtains the final score by multiplying individual scores. Generally, classifiers give measurements with diverse scales and *score normalization* should be performed. We use the min-max normalization [8] in this work.

Weighted Sum Rule (`WS`): In original sum rule, individual pattern classifiers are considered to be equally powerful. However, if some of them are more accurate, then it is useful to weight them in the fusion process. Let w_j's be the weights of classifiers computed on a separate validation set, then weighted sum rule is written as: $y_j = \sum_{k=1}^{K} w_j \times s_{kj}$.

Improved Consensus Voting: Consensus voting can be improved with the help of score measurements. Suppose that each classifier outputs both the nearest class label ω_i and its associated confidence or reliability value α_i. We will explain a specific method of determining α_i's in the next section. In *improved consensus voting*, if there are ties, then the class having the highest total confidence is selected as the final decision. We call this technique *Improved Consensus Voting with Maximum Reliability (`CV-M`)*. Another variant is to select the class among top-ranked classes having the maximum average reliability. This is a consensus voting in a reliability sense. We call this method as *Improved Consensus Voting with Average Reliability (`CV-A`)*.

Weighted Consensus Voting (`CV-W`): Similar to the weighted sum rule, individual classifier's strengths can be embedded into a consensus voting scheme where the counts of the top-ranked classes are increased by w_j's (in the original `CV`, we simply count each vote as one).

Highest Confidence (`HC`): *Highest confidence fusion* method simply selects the class among the top-ranks having the highest confidence. This technique essentially selects the class output of the most reliable pattern classifier.

Serial Fusion (`SF`): We have developed a serial two-stage cascaded fusion architecture. In the `SF`, the first pattern classifier, C_f, produces a nearest class list that is forwarded to a more complex second classifier C_s. Then, the second classifier determines the final class using the training patterns of the classes forwarded by C_f. The crucial point here is that the second classifier uses LDA to better separate similar classes in the transformed LDA subspace.

4. Experimental Results

In our experiments, we have used the `3D_RMA` dataset [2]. There are 106 subjects each having five or six shots. The data is obtained with a stereo vision structured light system. There are slight pose and expression variations. In our simulations, we put three images per subject into the training set, one image to validation set, and remaining images into the test set. We have formed four experimental folds where different combinations are put into the training and validation sets. In each fold there are a total of 318 gallery images, and 193 probe images. Table 1 presents the correct classification accuracies of the four individual 3D face classifiers for four experimental folds. Point cloud-based (PC) classifier attains the best performance with 91.19 per cent average accuracy. Facial profile (PRO) and depth image-based (LDA-DI) have similar rates, whereas shape index-based classifier (SI) performed worst.

Table 1. 3D face classifier performances.

	PC	SI	PRO	LDA-DI
E_1	86.01	78.76	83.42	88.08
E_2	92.75	88.08	92.23	91.71
E_3	92.75	86.53	93.26	89.64
E_4	93.26	89.12	94.30	92.23
Mean	91.19	85.62	90.80	90.41

Table 2 presents the classification performances for each fusion method. As confirmed in [6], rank-level fusion improves the overall accuracy. We have observed that Borda count method obtains 92.75 per cent accuracy and improves the best individual classifier (namely, PC) by 1.56 per cent. However, when compared to other fusion methods, `BC` is found to be the worst one. Especially, when we look at the performance of consensus voting (`CV`), we see 94.17 per cent accuracy which is better than `BC`. This finding reveals the fact that the most discriminating information in the fusion is at the top-ranks of the classifiers. The idea of incorporating reliability is also found to be beneficial. Now, lets briefly summarize how the reliability is estimated.

Reliability Estimation: In measurement-level fusion, each classifier outputs dissimilarity scores for each class in the training set. The scores generated by a classifier are related to the reliability. However scores can not be directly used as confidence measures since score normalization methods do not produce exactly uniform score ranges. To illustrate the point, score values of three different classifiers that are sorted in ascending order are shown in Figure 1.a. Notice that the scores (distances) for the nearest class changes significantly for different classifiers. So, using only score values as confidence is wrong. We have developed a technique to estimate the confidence of a classifier using a relative distance between the rank-1 and rank-2 class. However, this relative distance also depends on the score range. For this purpose we re-normalize the scores again by the min-max algorithm. For a given sorted score values, re-normalized scores are calculated by $s^{''} = \frac{s^{'} - min}{max - min}$ where *min* is the score of the rank-1 class, and *max* is the score of the rank-K class. K is found dynamically by locating the highest derivative along the tails of the sorted score values. This point is illustrated in Figure 1.a. Figure 1.b shows the re-normalized $s^{''}$ scores for the nearest 20 classes. Now we can correctly estimate the reliability of a classifier by computing the slope of the line between rank-1 and rank-2 scores dynamically for a test face.

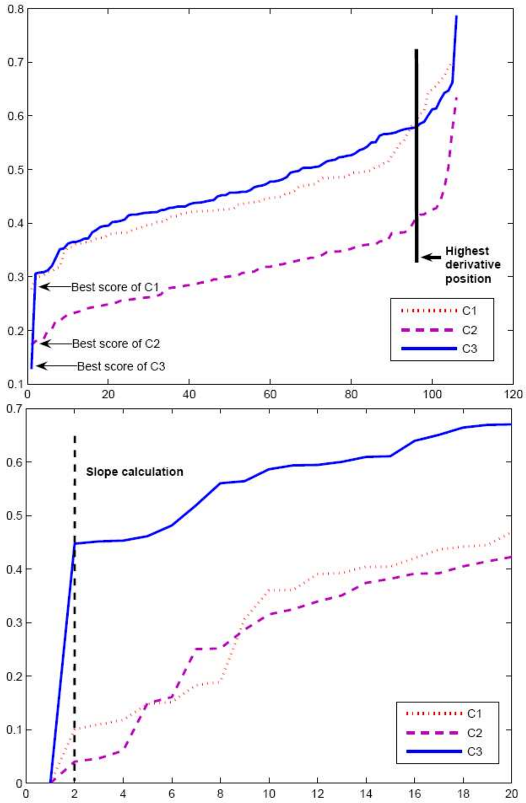

Figure 1. Top: Scores of three classifiers in ascending order, bottom: reliability estimation.

Reliability-assisted consensus voting schemes, `CV-M` and `CV-A` offers better identification rates when compared to `CV` by obtaining 94.82 and 94.30 per cent accuracies, respectively. When there are ties in `CV`, selecting a class having the maximum reliability (`CV-M`) is better than averaging

Table 2. Fusion performances on the `3D_RMA` database.

	`SUM`	`WS`	`CV`	`CV-M`	`CV-W`	`CV-A`	`BC`	`HC`	`PRODUCT`	`SF`
E_1	92.23	91.71	88.60	89.64	87.05	91.19	86.53	91.19	92.23	94.30
E_2	96.37	96.89	95.34	96.37	94.82	95.86	94.82	96.37	96.89	98.96
E_3	96.89	96.37	95.86	96.37	95.86	94.82	94.82	96.37	96.37	99.48
E_4	96.37	94.82	96.89	96.89	95.86	95.34	94.82	96.37	95.86	98.96
MEAN	95.47	94.95	94.17	94.82	93.39	94.30	92.75	95.08	95.34	97.93

reliability of top-ranked classes(`CV-A`). This fact is visible by observing the performance of highest confidence (`HC`) algorithm. Selecting the class having the highest reliability attains 95.08 per cent rate which is better than doing voting. This proves the usefulness of dynamic reliability estimation by the relative distance between rank-1 and rank-2 classes.

Arithmetic rules perform better than both abstract- and rank-level schemes, as expected. `SUM` and `PRODUCT` rules attain 95.47 and 95.34 per cent average identification rates, respectively. In all of the parallel fusion architectures, these are the best ones. However, we have found that employing a weighting scheme in the sum rule (`WS`) degrades the classification rate. Let P_i be the accuracy of classifier C_i on the validation set, then weights w_i are computed by $\frac{P_i}{\sum_{k=1}^{L} P_k}$. We have also tried other techniques (found in [9]) for weight estimation, but this approach gave the best performance. The failure of weighting-based fusion can be explained by the insufficient amount of validation data. However, in practical systems, we generally do not have sufficient amount of training samples per subject. So, this situation will always be the limiting factor. The same degradation is also present in weighted consensus voting (`CV-W`).

Serial fusion (`SF`), is the best performer among all of the fusion algorithms. In `SF`, point cloud-based classifier (PC) is used as a first classifier. It forwards the top 20 classes in the ranked-list to the second classifier. The second classifier also uses the point cloud features, but applies LDA to the training samples of the forwarded classes. For each probe image, LDA-based classifier dynamically constructs a subspace of dimensionality 19 which is the maximum LDA dimensionality. By employing this cascaded fusion, the performance of the system can be improved from 91.19 per cent (PC) to 97.93 per cent. This improvement validates the superiority of the serial architecture when compared to the parallel ones.

5. Conclusion

In this paper, we review and compare various decision-level fusion architectures, and also propose novel ones for 3D face recognition. We have developed four different face classifiers which are frequently used in the literature: ICP-based surface matcher, depth image-based LDA classifier, facial profile-based matcher, and shape index-based matcher. We have shown that sum and product rules, Borda count, and consensus voting have potential to improve identification rate. However, we have shown that using a novel reliability estimation technique we can obtain better fusion schemes such as improved consensus voting and highest confidence rule. It is also noted that weighted fusion rules can not reach optimal accuracies if you have insufficient validation data. We also prove that by a careful design of a serial scheme which embodies an LDA analysis, significant performance improvement is possible.

References

[1] P. Besl and N. McKay. A method for registration of 3D shapes. *IEEE Tran. on PAMI*, 14:239–256, 1992.

[2] C. Beumier and M. Acheroy. Automatic 3D face authentication. *Image and Vision Computing*, 18(4):315–321, 2000.

[3] K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition. *Computer Vision and Image Understanding*, 101(1):1–15, 2006.

[4] K. Chang, K. Bowyer, and P. Flynn. Adaptive rigid multi-region selection for handling expression variation in 3D face recognition. In *IEEE Workshop on FRGC*, 2005.

[5] K. I. Chang, K. W. Bowyer, and P. J. Flynn. An evaluation of multi-modal 2D+3D face biometrics. *IEEE Tran. on PAMI*, 27(4):619–624, 2005.

[6] B. Gökberk, A. A. Salah, and L. Akarun. Rank-based decision fusion for 3D shape-based face recognition. In T. Kanade, A. Jain, and N. K. Ratha, editors, *Proceedings of AVBPA, (LNCS)*, volume 3456, pages 1019–1029, 2005.

[7] M. O. İrfanoğlu, B. Gökberk, and L. Akarun. 3D shape based face recognition using automatically registered facial surfaces. In *ICPR*, pages 183–186, 2004.

[8] A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270–2285, 2005.

[9] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, 2004.

[10] X. Lu, A. Jain, and D. Colbry. Matching 2.5D face scans to 3D models. *IEEE Tran. on PAMI*, 28(1):31–43, 2006.

[11] G. Pan and Z. Wu. 3D face recognition from range data. *International Journal of Image and Graphics*, 5(3):573–593, 2005.