Analysis of Facial Features in Identical Twins

Brendan Klare, Alessandra A. Paulino, and Anil K. Jain*
Dept. of Computer Science and Engineering
Michigan State University
East Lansing, MI, U.S.A.
{klarebre,paulinoa,jain}@cse.msu.edu

Abstract

A study of the distinctiveness of different facial features (MLBP, SIFT, and facial marks) with respect to distinguishing identical twins is presented. The accuracy of distinguishing between identical twin pairs is measured using the entire face, as well as each facial component (eyes, eyebrows, nose, and mouth). The impact of discriminant learning methods on twin face recognition is investigated. Experimental results indicate that features that perform well in distinguishing identical twins are not always consistent with the features that best distinguish two non-twin faces.

1. Introduction

As the performance of face recognition systems in constrained environments continues to improve [6], focus is shifting from methods that improve face recognition performance in general, to methods that handle specific cases of failure. Until recently, a scenario that has been known to confound even human face recognition had not been studied for automated face recognition systems. This scenario is the ability to distinguish between identical (monozygotic) twins. Because identical twins are genetically equivalent, their facial appearance is also quite similar. Generally, differences in their appearance can only be attributed to exposure to different environmental factors (such as dietary habits and aging) and rare cases of genetic mutations.

Failing to properly identify a subject who has an identical twin has many negative economic and security implications [23]. As such, two recent studies have measured the performance of commercial off the shelf (COTS) face recognition systems (FRS) in distinguishing between identical twins [23, 18]. The ability to match identical twin fingerprints has been examined as well [8]. However, there has yet to be an analysis of which facial features are the most distinctive between identical twin pairs, and whether face recognition systems can be designed to improve the performance in this scenario. Thus, while previous studies have provided a better understanding of the vulnerabilities of face recognition systems when presented with a twin imposter, this paper offers some initial guidance for how face recognition systems can be designed to improve twin identification.

Klare and Jain organized the facial features for representing face images into three levels. Level 1 features (such as Eigenfaces) describe the general appearance of the face, and are useful for quickly discarding large portions of the population. Level 2 features (such as SIFT [13] and LBP [17] descriptors, and Gabor wavelets) embed precise geometrical and structural measurements of the face, and are necessary for discriminating between two similar looking faces. Finally, Level 3 features (such as scars and facial marks) consist of skin irregularities and micro features that are useful for discriminating between ambiguous faces (e.g. identical twins) or incomplete faces (e.g. occlusions).

Our research measures the distinctiveness of Level 2 (MLBP and SIFT) and Level 3 (facial marks) facial features with respect to identical twins. A framework for component-based face recognition is developed in order to analyze the distinctiveness of the different facial components (eyes, eyebrows, nose, and mouth). Finally, we examine the impact of linear discriminant analysis methods on twin recognition to help understand (i) whether facial feature subspaces better suited for distinguishing between twin pairs can be learned, and (ii) whether such subspaces impact face recognition performance in non-twin comparisons. This work does not consider the performance of commercial face recognition systems on identical twins because this was previously examined [18].

2. Related Work

Sun et al. were the first to measure face recognition performance on a database of identical twins [23]. The study measured the performance of a leading COTS FRS

---

*A.K. Jain is also with the Dept. of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-713, Korea.

(a)   (b)

Figure 1. Examples of identical twin pairs from the 2009 Twins Day collection [18]. Each pair of images in (a) and (b) are two different subjects who are identical twins. Distinguishing between pairs of identical twins is one of the more challenging problems in face recognition.

on a dataset containing 51 pairs of identical twins. Analysis of the match score distributions from twin imposters and genuine subjects indicated high levels of similarity. However, the dataset was collected under non-ideal conditions, including minor pose variations, non-uniform background, and uncontrolled illumination.

Phillips et al. [18] presented the largest study to date of face recognition performance on identical twins. A dataset of 126 twin pairs was analyzed using the three top COTS FRS participants in the Multi-Biometrics Evaluation [6]. The results demonstrated that COTS FRS perform admirably in distinguishing between twin pairs captured in controlled environments with minimal time lapse. However, a smaller data set of twin pairs captured with a time lapse of one year (24 pairs) indicated that even this relatively small time difference was more detrimental to recognition performance than other factors such as adverse lighting conditions.

Klare and Jain have proposed a taxonomy of facial features that organizes the features into three levels (Level 1, Level 2, and Level 3) [10]. This taxonomy is used as a guideline for analyzing the distinctiveness of facial features in distinguishing between identical twin pairs. Level 1 features, such as appearance-based features, do not contain sufficient discriminatory information to improve the identification accuracy of identical twins and are not considered in this work. Level 2 facial features consist of representations that encode the shape and structure of the face and are localized with respect to the facial geometry. Examples of Level 2 features include densely sampled local binary patterns (LBP) [17] and SIFT [13] feature descriptors, Gabor wavelets, and anthropometric shape information. Level 2 features are critical in standard face recognition tasks, however their suitability for distinguishing between identical twins may be impacted by the strong facial similarities that pairs of identical twins exhibit.

In the facial feature taxonomy, Level 3 facial features consist of facial irregularities such as facial marks and scars, as well as micro texture information. One form of facial marks, a mole (or nevus), has been shown to be unique between identical twins [7]. While facial marks (and Level 3 features in general) are often insufficient for identification, they have been fused with a COTS FRS to improve the identification accuracy [19] and are believed to help resolve ambiguous matches such as identical twins [10].

Component-based face recognition involves explicitly measuring the similarity of two subjects using a specific facial component (e.g. eyes, nose, mouth). Elastic graph bunch matching (EGBM) [25] is similar to such an approach, where facial landmarks are used to compute Gabor representations at different facial components, but while the representation is measured per component, the matching is performed across all components taken together. The most advanced methods in component-based face recognition can be found in a host of algorithms for performing periocular recognition [20, 3, 16]. Periocular recognition makes use of only images of the eyes (and often eyebrows) to determine a subject's identity.

3. Facial Features

3.1. Level 2 Features

We examine two types of Level 2 facial features: multi-scale local binary patterns (MLBP) and scale invariant feature transform (SIFT) feature descriptors. The multi-scale local binary patterns are the concatenation of LBP feature descriptors with radii 1, 3, 5, and 7 [11]. Both the MLBP and SIFT feature descriptors are sampled across the face (or component) in 32x32 sized patches that overlap by 16 pixels. The feature descriptor from all image patches are concatenated into a final feature vector x representing the face or facial component. All face images are normalized

as described in Section 6.

3.2. Level 3 Features

In addition to SIFT and MLBP, we also examine the use of facial marks in distinguishing between identical twins. Facial marks are Level 3 features that can be used to distinguish between identical twin pairs. The facial marks used in this work were manually annotated[1] on all images by a single examiner. The order of the images in our database was randomized to prevent a bias from labeling multiple images of the same subject, or their twin, in succession. The facial mark matching results are considered optimistic due to the use of a single examiner to label the images.

The previous work on facial mark identification presented a matching module that was compatible with textual queries (e.g. "mole on right cheek") [19]. In this work, we have developed a mark matching algorithm that does not facilitate textual queries, but improves the matching performance.

The first step in the mark matching algorithm is to align the face images. This is achieved using the geometric normalization technique mentioned in Section 6. In the second step, the Euclidean distance between a pair of marks in the two faces being compared is measured, and the marks are labeled a match if the distance is smaller than the threshold τ. Next, we eliminate one-to-many matches from the previous step by choosing the mark pair with the smallest distance. The final step involves using any mark matches found (or not found) between two faces to generate a similarity score.

The face mark similarity score $s_{i,j}$ between two faces i and j is generated as

$$s_{i,j} = \frac{n_{\text{Matched}}}{\sqrt{n_i \cdot n_j}} \qquad (1)$$

where n_{Matched} is the number of matched marks between face i and face j, and n_i and n_j are the total number of marks detected in faces i and j, respectively. However, the formula in Eq. 1 is not used in the following special cases. If either face i or face j has no face marks, then $s_{i,j} = 0$. This means there is not enough information to say whether they match or do not match because the absence of marks might be due to makeup or some other factor. If $n_{\text{Matched}} = 0$, then $s_{i,j} = -1$, which indicates a strong evidence that the marks are not from the same person.

The scoring method presented above is designed for sum of score fusion with match scores from Level 2 features by (i) decreasing the match score when face marks provide evidence of non-matching, (ii) increasing the match score when face marks provide evidence of similarity, and (iii) not altering the match score when we do not have enough

evidence to make a decision using face marks. Because the face marks alone are not sufficient for identification, we fuse the scores using weighted sum of score fusion [21], with a small weight assigned to face marks matching scores. Although the dataset we used was collected on the same day or only a few days apart, this framework should be robust to the case when the dataset is collected over a larger time lapse.

4. Component-based Representation

In this section a framework for analyzing two faces on a per-component basis is presented, which is illustrated in Figure 2. The motivation to analyze the discriminative information contained in each facial component with respect to identical twins is based on the need to isolate the different sources of information present in the face. Through such isolation we intend to better understand how face recognition systems can be modified to improve face recognition for identical twins.

The first step in the component-based analysis is to detect a set of facial landmarks on the face. This was performed using the open source Active Shape Model [4] software Stasm [15]. In order to improve the accuracy of landmark detection, the PittPatt Face Recognition SDK [1] was used to first detect the center of the two eyes and the nose. These landmarks were in turn used to initialize the active shape model by performing an affine transformation to align the mean shape with the three detected landmarks.

Once the facial landmarks are detected, the next step is to align each facial component. To align each facial component, we performed Procrustes analysis/alignment [5] on each facial component separately. While Procrustes alignment is often performed on the entire collection of facial landmarks [22], aligning each component separately results in a more accurate alignment. This per-component alignment does lose some information regarding the spatial relation between each component; however, our objective is to determine the information content in each component. Indeed, our analysis of the entire face incorporates information regarding the spatial relation among the components (e.g. how far the mouth is from the nose). The method of aligning each component separately has been used previously to cluster face images [12]. Figure 3 shows the improved alignment from aligning each component separately.

With the landmarks from each facial component aligned, we perform an affine transformation on the image using the rotation, translation, and scaling parameters from the Procrustes alignment. Each component is then cropped using the minimum and maximum horizontal and vertical landmark location from all images begin aligned. Figure 2 shows an example of the components after alignment and cropping.

Once each component is aligned, we are able to repre-

---

[1]While automated mark detection works reasonably well, we did not want inaccuracies from this process to degrade the feature analysis.
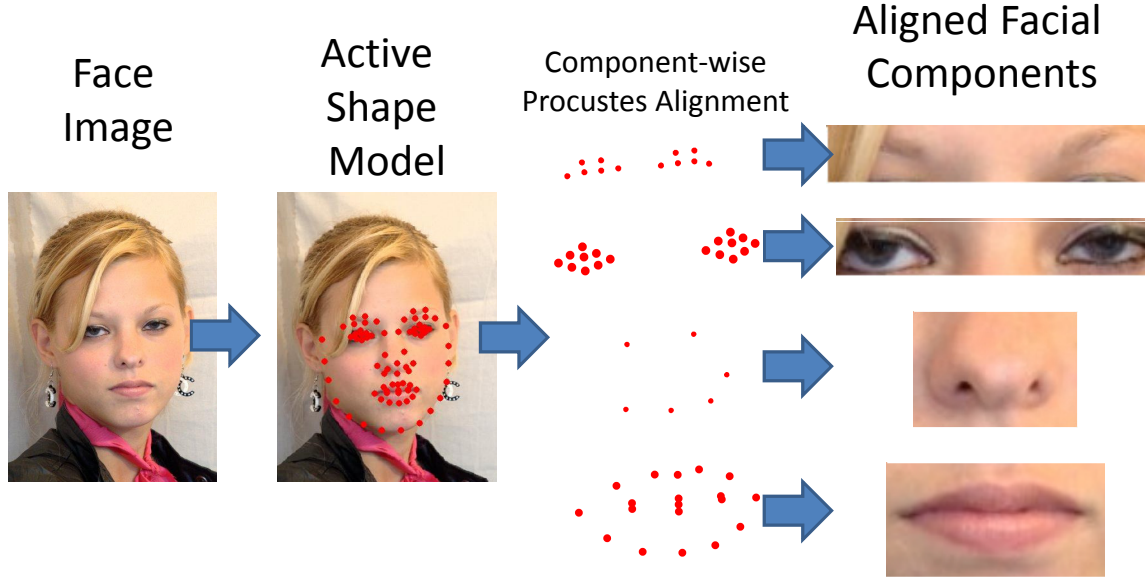
Figure 2. The process for extracting facial components. After detecting facial landmarks using an Active Shape Model, Procrustes analysis is performed separately on each facial component.

sent the component images using Level 2 features in the same manner as with the entire face (MLBP and SIFT). As mentioned earlier, the SIFT and LBP features were extracted using 32x32 patches with a 16 pixel overlap. As such, we scaled each component to a size that is a factor of 16. Though this scaling does not maintain the aspect ratio, it is performed consistently for each subject and thus does not have any negative impact. The final component image sizes are:

- Eyes = 64x160
- Eyebrows = 64x160
- Nose = 96x96
- Mouth = 64x160

The components shown in Figure 2 and Figure 3 are before the component scaling is performed.

5. Discriminative Learning

One of the significant contributions to automated face recognition has been the application of linear discriminant analysis (LDA). Belhumeur et al. [2] originally offered this solution on pixel representations of the face by solving the generalized eigenvalue problem $S_B \cdot \Psi = \Lambda \cdot S_W \cdot \Psi$, where S_B is the between-class scatter matrix and S_W is the within-class scatter matrix. The eigendecomposition has the effect of maximizing the Fisher criterion $||S_B'||/||S_W'||$, where S' is the scatter matrix from the subspace Ψ. Most practical face recognition scenarios require learning Ψ from a separate set of training subjects than the subjects the system will be tested on.

The original FisherFace method was applied on Level 1 appearance (pixel) features. Because appearance represen-

tations have high spatial redundancy and their dimensionality may be substantially reduced in size without losing discriminative information [10], the FisherFace approach has been quite successful in learning improved feature subspaces from the original representation. However, as faces are more commonly being represented with more verbose Level 2 features (such as LBP's), the richer information content impacts the ability to apply the standard discriminant analysis. Instead, methods such as random sampling LDA (RS-LDA) [24], direct LDA (D-LDA) [9], and regularized LDA (R-LDA) [14] are applied to handle degenerative scatter matrices.

In this work we adopt the random sampling LDA method for our discriminative subspace analysis. RS-LDA generates a set of B subspaces Ψ_b that are each learned on a subset of training instances (bagging) and a subset of features (random sampling). Thus, if $x \in \mathbb{R}^d$ is the initial representation of a face image (e.g. a concatenation of LBP feature vectors), for each subspace b, we learn Ψ_b on $x_b \in \mathbb{R}^{d_b}$, where the $d_b < d$ features in x_b are a randomly sampled subset of the original d features.

The following parameters are tunable in our implementation of RS-LDA: (i) $0 < s < 1$ is the percentage of the original set of d features to be sampled at each stage b, (ii) p is the number of face images to use at each stage, and (iii) $0 < v < 1$ is the percentage of variance to be retained in each stage. The variance is controlled by performing principal component analysis (PCA) on x_b at each stage.

We present a modification to RS-LDA that is designed to improve the performance on identical twins. Let x^{i_k} denote the feature vector of the k-th image for the i-th subject.
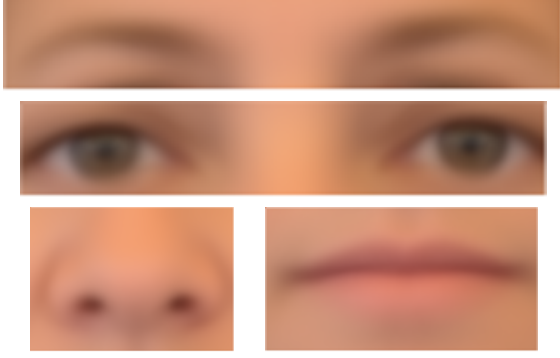
(a)

(b)

Figure 3. The mean shape for each facial component (eyebrows, eyes, nose, and mouth) from 523 images. (a) The mean shapes resulting from performing Procrustes alignment on each facial component separately. (b) The mean shapes from performing Procrustes alignment on the entire face. The mean components in (a) appear sharper than those in (b), indicating an improved alignment from the per-component alignment.

$\mu_i = \sum_{k=1}^{n_i} x^{i_k}/n_i$ is the mean vector for each subject i, where n_i is the number of instances for subject i. For all N subjects, the total mean vector is $\mu = \sum_{i=1}^{N} \mu_i \cdot n_i/n$, where $n = \sum_{i=1}^{N} n_i$. Traditionally, S_B is built as $S_B = \sum_{i=1}^{N} n_i \cdot (\mu_i - \mu) \cdot (\mu_i - \mu)^{\mathrm{T}}$ [2].

In order to learn a subspace designed to distinguish between pairs of identical twins, we introduce a method called Twin RS-LDA. Let μ_i^{Twin} be the average of the two mean vectors for the subject i, and his identical twin j, i.e. $\mu_i^{\mathrm{Twin}} = (n_i \cdot \mu_i + n_j \cdot \mu_j)/(n_i + n_j)$. Next we build the twin between-class scatter matrix $S_B^{\mathrm{Twin}} = \sum_{i=1}^{N}(\mu_i - \mu_i^{\mathrm{Twin}}) \cdot (\mu_i - \mu_i^{\mathrm{Twin}})^{\mathrm{T}}$. Finally, for each random sample stage b, we learn the matrix Ψ_b^{Twin} from an eigendecomposition of $S_B^{\mathrm{Twin}} \cdot \Psi^{\mathrm{Twin}} = \Lambda \cdot S_W \cdot \Psi^{\mathrm{Twin}}$.

By replacing S_B with S_B^{Twin}, Twin RS-LDA attempts to maximize the difference energy between twin pairs, as opposed to the difference energy between all pairs of subjects. The within-class scatter S_W remains the same because we still seek a subspace that minimizes the within-class varia-
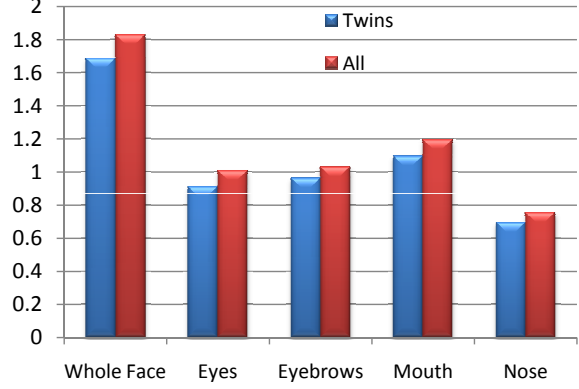

Figure 4. The Fisher separability values for each facial component and the entire face using MLBP feature descriptors. The "Twin" scores are computed from F^{Twin} from Eq. 2, and the "All" scores are computed from F from Eq. 3.

tion.

Both RS-LDA and Twin RS-LDA are applied to MLBP and SIFT feature descriptors. Further, they are both applied to the complete face image as well as each facial component separately.

Generating the two between-class scatter matrices also allows us to compare the inherent separability of different features and components with respect to distinguishing only twin pairs and distinguishing all pairs of subjects. For this purpose, we introduce the Fisher separability measure F to measure the separability amongst all comparisons, and F^{Twin} for only twin comparisons. Letting $\lambda_{B^{\mathrm{Twin}}} \in \mathbb{R}^d$ be the vector of eigenvalues from S_B^{Twin}, $\lambda_B \in \mathbb{R}^d$ be the vector of eigenvalues from S_B, and $\lambda_W \in \mathbb{R}^d$ be the vector of eigenvalues from S_W, we compute the Fisher values as

$$F^{\mathrm{Twin}} = \frac{\sum_{i=1}^{d} \lambda_{B^{\mathrm{Twin}}}(i)}{\sum_{i=1}^{d} \lambda_W(i)} \quad (2)$$

$$F = \frac{\sum_{i=1}^{d} \lambda_B(i)}{\sum_{i=1}^{d} \lambda_W(i)} \quad (3)$$

The relative values of these two measures will help understand if certain facial components are more distinct in identifying twins.

6. Experimental Results

Our experiments were conducted on the twin data collected in 2009 at the Twins Day festival, which has been described in detail in [18]. We did not have access to the pairs collected in 2010 [18]. The 2009 database contains 87 pairs of identical twins (174 subjects). Each subject contained either two or four images depending on whether they attended one or two days of data collection. The images used were all frontal pose, neutral expression, and the subjects did not
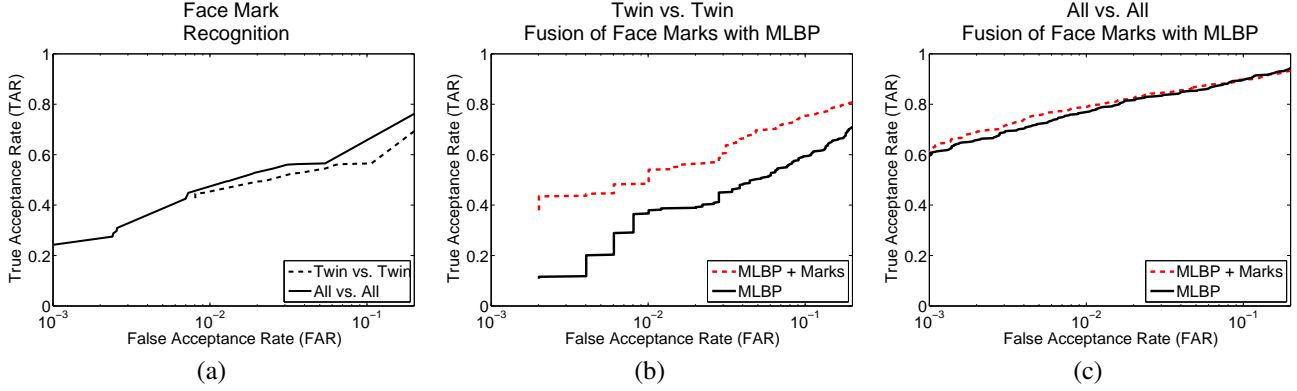
Figure 5. Recognition accuracies using facial marks features. Note that face marks in identical twins may not be fully distinct or may not even be present (a). Still, the amount of performance improvement from fusing face marks with MLBP features in Twin vs. Twin (b) over All vs. All (c) demonstrates the value of facial marks in distinguishing identical twins.

wear glasses. The images were collected in tents with studio lighting; however, a small directional illumination was detected in the images. Because of the negative effect the illumination has on the MLBP and SIFT representations, a difference of Gaussian (DoG) filter ($\sigma_1 = 2$ and $\sigma_2 = 4$) was used to discard this low frequency noise. The use of a DoG filter nearly doubled the true accept rate of the MLBP features.

We measured the face recognition performance using each of the four facial components (eyes, eyebrows, nose, and mouth) as well as the entire face. For the entire face, we geometrically normalized the face using the center of the eyes to (i) rotate the image so that the angle between the eyes was 0 degrees, (ii) scale the image so the inter-pupilary distance was 75 pixels, and (iii) crop the face to a height of 250 pixels and a width of 200 pixels.

The RS-LDA and Twin RS-LDA subspaces were trained using a training set of twin pairs, and was tested using a different set of twin pairs (i.e. no subject in training was used for testing). Due to the relatively small number of twin pairs, bootstrap sampling was performed to generate the training and testing splits. From the 87 twin pairs, we sampled with replacement 87 pairs for training. That is, we sampled 87 total pairs such that some twin pairs would occur multiple times within the 87 training pairs, and other twin pairs would be unused for training. The remaining pairs were used as an evaluation set for the split. The evaluation results are averaged over 500 bootstrap trials.

Table 1 lists the matching accuracies for the entire face as well as each facial component. Both SIFT and MLBP feature descriptors are shown. For each component/feature combination, we reported the accuracy of using only the feature descriptor and using the feature descriptor in conjunction with Twin RS-LDA and RS-LDA. The performance is measured across two scenarios: Twin vs. Twin and All vs. All. For Twin vs. Twin, all imposter comparisons are from pairs of identical twins. This measures how

often each feature and facial component cannot distinguish between the twin pairs. All vs. All is from the same twin data set, but also includes comparisons between different subjects who are not twins. The performance on All vs. All helps provide context for the Twin vs. Twin results.

One of the more surprising results from our experiments was the negligible difference in recognition accuracies noticed from RS-LDA and Twin RS-LDA. However, this may be due to the small number of training images impeding the ability to properly learn improved feature extractions. However, a general performance increase is noticed in most component/feature combinations, indicating some value in subspace learning. The fact that Twin RS-LDA did not decrease the recognition performance in the All vs. All scenario may suggest that face recognition vendors can incorporate discriminative subspaces trained using an objective function similar to Twin RS-LDA. Given a large number of training subjects, this may result in an improved ability to distinguish identical twins while not impacting the standard All vs. All face recognition scenario.

Figure 4 shows the Fisher separability for each facial component as well as the entire face using MLBP features. No facial component seems more useful in a twin scenario than a standard matching scenario. Of particular interest, though, is the strong Fisher value of the mouth component. While the mouth is not generally considered as informative as the eyes or eyebrows, the per-component alignment performed may have increased the discriminative value of this facial region.

Figure 5 shows the recognition performance when using facial mark features. We observed a slight decline in accuracy when using facial marks on the Twins vs. Twins scenario when compared to the All vs. All scenario. However, the benefit of using facial marks for twin comparisons is seen by the strong performance increase when fused with a Level 2 feature. By fusing the facial mark and MLBP scores in the framework we presented in Section 3.2, there
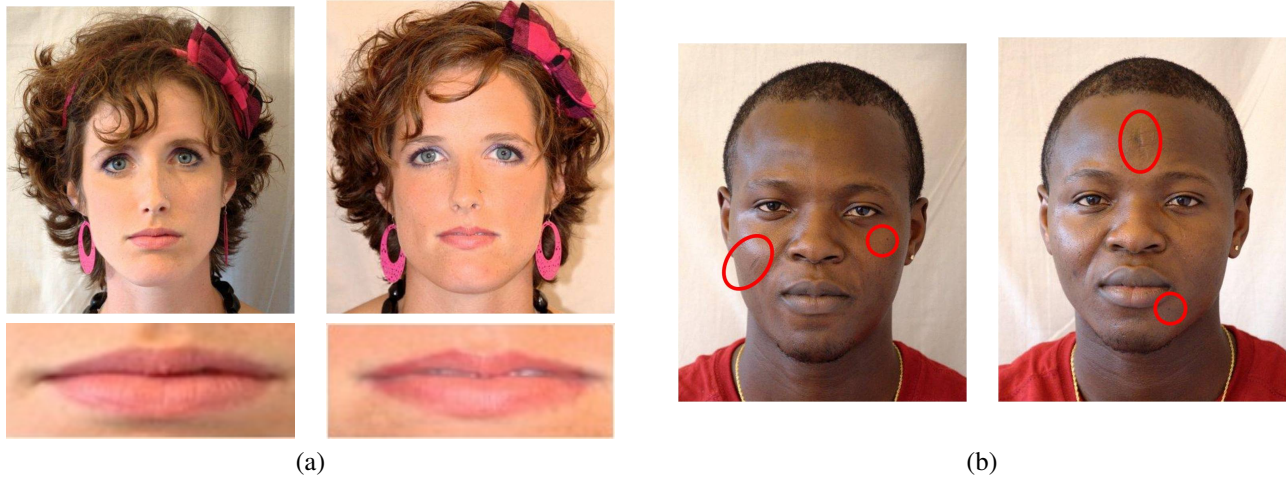
Figure 6. The two examples above each show a pair of identical twins that failed to match using MLBP and SIFT features across the entire face. (a) Using only the mouth component, the two twins were distinguishable. (b) Fusing face mark match scores enabled this twin pair to be distinguished.

is no significant difference in the verification rates of All vs. All. However, the verification performance for Twin vs. Twin significantly improves when using facial marks (at a FAR of 1.0%, the TAR is 37% without mark fusion and 50% with mark fusion). Therefore, with the correct framework, facial marks information can be used to help distinguish between identical twins without decreasing performance on other subjects.

7. Conclusions

This research analyzed the discriminability of Level 2 (MLBP and SIFT) and Level 3 (facial marks) features with respect to distinguishing identical twins. Additionally, the separability of different facial components (eyes, eyebrows, nose, and mouth) as examined. It was shown that the saliency of facial features changes from standard face recognition tasks to attempting to distinguish identical twins. This is illustrated in Figure 6, where only the mouth component was able to separate a twin pair, and face mark fusion separated another twin pair.

Future work should involve expanding the size of twin datasets to improve the training for twin identification. The results shown in this work should also be considered by commercial vendors in order to improve system performance in the presence of identical twin pairs.

Acknowledgement

References

[1] PittPatt Face Recognition SDK, Pittsburgh Pattern Recognition, http://www.pittpatt.com.

[2] P. Belhumeur, J. Hespanda, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 19(7):711–720, 1997.

[3] S. Bharadwaj, H. Bhatt, M. Vatsa, and R. Singh. Periocular biometrics: When iris recognition fails. In *Proc. of IEEE Conference on Biometrics: Theory, Applications and Systems*, 2010.

[4] T. F. Cootes et al. Active shape models—their training and application. *Computer Vision Image Understanding*, 61(1):38–59, 1995.

[5] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[6] P. J. Grother, G. W. Quinn, and P. J. Phillips. MBE 2010: Report on the evaluation of 2D still-image face recognition algorithms. *National Institute of Standards and Technology, NISTIR*, 7709, 2010.

[7] Gu Zhu et al. A major quantitative-trait locus for mole density is linked to the familial melanoma gene cdkn2a: A maximum-likelihood combined linkage and association analysis in twins and their sibs. *The American Journal of Human Genetics*, 65(2):483–492, 1999.

[8] A. K. Jain, S. Prabhakar, and S. Pankanti. On the similarity of identical twin fingerprints. *Pattern Recognition*, 35(11):2653–2663, 2002.

[9] H. Y. Jie, H. Yu, and J. Yang. A direct LDA algorithm for high-dimensional data – with application to face recognition. *Pattern Recognition*, 34:2067–2070, 2001.

Table 1. True accept rates (TAR) at a fixed false accept rate (FAR) or 1.0% are listed for the Level 2 features (MLBP and SIFT) at each facial component.

	Twins vs. Twins			All vs. All		
	No LDA	Twin-RSLDA	RSLDA	No LDA	Twin-RSLDA	RSLDA
All Patches:						
MLBP	31.5 ± 7.7	32.0 ± 7.5	31.7 ± 7.5	78.4 ± 5.3	78.2 ± 6.1	78.3 ± 6.3
SIFT	14.2 ± 5.6	22.7 ± 8.3	22.7 ± 8.0	73.2 ± 6.6	81.6 ± 6.2	81.6 ± 6.2
Eyes:						
MLBP	7.6 ± 3.7	11.3 ± 6.4	10.7 ± 6.0	41.1 ± 5.1	47.3 ± 5.9	47.3 ± 5.9
SIFT	1.4 ± 1.3	7.1 ± 3.8	6.8 ± 3.8	20.8 ± 3.4	41.9 ± 4.5	41.2 ± 4.4
Eyebrows:						
MLBP	15.6 ± 8.2	16.2 ± 9.4	16.0 ± 9.3	43.9 ± 4.4	45.1 ± 4.7	45.1 ± 4.8
SIFT	4.4 ± 4.0	7.9 ± 5.2	7.8 ± 5.0	21.2 ± 4.7	33.9 ± 6.3	33.1 ± 5.9
Mouth:						
MLBP	11.6 ± 7.4	12.2 ± 7.4	12.2 ± 7.3	43.6 ± 6.9	46.6 ± 6.2	46.5 ± 6.2
SIFT	9.0 ± 4.0	10.9 ± 6.3	10.9 ± 6.5	24.7 ± 4.1	35.4 ± 5.1	35.3 ± 5.1
Nose:						
MLBP	8.0 ± 4.0	7.7 ± 4.9	7.7 ± 4.8	25.9 ± 4.8	22.6 ± 8.5	23.1 ± 7.8
SIFT	6.0 ± 4.5	9.2 ± 5.6	8.9 ± 5.5	18.6 ± 3.7	22.8 ± 4.2	22.6 ± 4.3

 # Match Comparisons: (Twin vs. Twin)=109 ± 18, *(All vs. All)*=109 ± 18

 # Non-Match Comparisons: (Twin vs. Twin)=181 ± 24, *(All vs. All)*=10521 ± 2215

[10] B. Klare and A. K. Jain. On a taxonomy of facial features. In *Proc. of IEEE Conference on Biometrics: Theory, Applications and Systems*, 2010.

[11] B. Klare, Z. Li, and A. Jain. Matching forensic sketches to mugshot photos. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(3):639–646, 2011.

[12] B. Klare, P. Mallapragada, A. K. Jain, and K. Davis. Clustering face carving: Exploring the devatas of Angkor Wat. In *Proc. International Conference on Pattern Recognition*, 2010.

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[14] J. Lu, K. Plataniotis, and A. Venetsanopoulos. Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition. *Pattern Recognition Letters*, 26(2):181 – 191, 2005.

[15] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. *ECCV*, 2008.

[16] P. Miller, J. Lyle, S. Pundlik, and D. Woodard. Performance evaluation of local appearance based periocular recognition. In *Proc. of IEEE Conference on Biometrics: Theory, Applications and Systems*, 2010.

[17] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 24(7):971–987, 2002.

[18] P. J. Phillips et al. Distinguishing identical twins by face recognition. In *Proc. of IEEE Conference on Automatic Face and Gesture Recognition*, 2011.

[19] U. Park and A. K. Jain. Face matching and retrieval using soft biometrics. *IEEE Trans. Information Forensics and Security*, 5(3):406–415, 2010.

[20] U. Park, R. Jillela, A. Ross, and A. Jain. Periocular biometrics in the visible spectrum. *IEEE Trans. Information Forensics and Security*, 6(1):96 –106, 2011.

[21] A. Ross and A. Jain. Information fusion in biometrics. *Pattern Recognition Letters*, 24(13):2115–2125, 2003.

[22] J. Shi, A. Samal, and D. Marx. How effective are landmarks and their geometry for face recognition? *Computer Vision and Image Understanding*, 102(2):117 – 133, 2006.

[23] Z. Sun, A. Paulino, J. Feng, Z. Chai, T. Tan, and A. K. Jain. A study of multibiometric traits of identical twins. In *Proc of SPIE, Biometric Technology for Human Identification VII*, 2010.

[24] X. Wang and X. Tang. Random sampling for subspace face recognition. *Int. Journal of Computer Vision*, 70(1):91–104, 2006.

[25] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *Proc. of International Conference on Image Processing*, 1:129, 1997.