Audio-Visual Twins Database

Jing Li[1], Li Zhang[1], Dong Guo[2], Shaojie Zhuo[3], Terence Sim[1]
[1] School of Computing, National University of Singapore, [2] Facebook, Menlo Park, [3] Qualcomm, Toronto
{lijing, lizhang, tsim}@comp.nus.edu.sg, dnguo@fb.com, jayzhuo@gmail.com

Abstract

Identical twins pose an interesting challenge for recognition systems due to their similar appearance. Although various biometrics have been proposed for the problem, existing works are quite limited due to the difficulty of obtaining a twins database. To encourage the methods for twins recognition and make a fair comparison of them by using the same database, we collected an audio-visual twins database at the Sixth Mojiang International Twins Festival held on 1 May 2010,China. Our database contains 39 *pairs of twins in total, including Chinese, American and Russian subjects. This database contains several face images, facial motion videos and audio records for each subject. In this paper, we describe the collection procedure, organization of the database, and usage method of the database. We also show our experiments on face verification, facial motion verification and speaker verification for twins to provide usage examples of the database.*

1. Introduction

Twins population has been growing in recent decades. As per statistical data [13], twins birth rate has risen from 17.8 to 32.2 per 1000 birth with an average 3% growth per year since 1990. Even though currently identical twins still only represent a minority (0.2% of the worlds population), it is worth noting that the total number of identical twins is equal to the whole population of countries like Portugal or Greece. Since identical twins share most of the genetic code, they may look alike in many physiological traits and sometimes even share similar behavioral characteristics. This poses a great challenge for recognition systems.

Different biometrics have been proposed to distinguish between identical twins, including physiological biometrics (such as 2D face [10, 15, 17], 3D face [5], iris [17], fingerprint [9, 17], palmprint [11] *etc.*) and behavioral biometrics (such as voice [3, 8] as well as handwriting [16]). Physiological biometrics are often greatly influenced by genes, therefore they become less effective for identical twins who share identical genes compared to general non-twins population. On the other hand, behavioral biometrics are more susceptible to individual's environment and may perform better in distinguishing between identical twins. The main drawback of behavioral biometrics is that they are not consistent over time and change as an individual grows older. In other related works, multimodal biometrics are used to distinguish between twin siblings. For example, Sun *et al.* [17] implemented a multimodal biometrics system using both 2D face and fingerprints. Their experimental results showed that fusion of 2D face and fingerprint (EER 7.65) performed worse than unimodal fingerprint recognition system(EER 6.79) for identical twins. CN *et al.* [6] proposed a multimodal biometrics system for twins recognition based on 2D face, fingerprints and iris pattern.

Compared to general non-twins population, works for twins recognition are rather limited due to the difficulty of obtaining a twins database. To the best of our knowledge, there is no twins database that contains audio records and face motion videos. Hence we came up with the idea to collect such a database, which should be able to contribute to moving forward the techniques for twins recognition.

2. Related Work

So far, public twins databases available to researchers include CASIA-Iris-Twins [17], 3D Twins Expression Challenge (3D TEC) Dataset [1], and ND-TWINS-2009-2010 [15]. All these databases are collected at twins festivals, where large number of twins are reachable. CASIA-Iris-Twins is a twin iris image dataset composed of 100 pairs of twins. It was collected in 2007 at the Annual Twins Festival in Beijing and the most of the subjects were children. 3D TEC Dataset contains 3D face scans of 107 pairs of twins with neural and smiling expressions by a Minolta Vivid 910 in a controlled light setting. ND-TWINS-2009-2010 was collected at the Twins Days festival in Twinsburg, Ohio in August 2009 and August 2010. The dataset contains 24050 face images of 435 attendees in both indoor and outdoor light settings. We list these databases in the Table 1.

We collected an audio-visual twins database at the Sixth Mojiang International Twins Festival held on 1 May 2010, China. There are 39 pairs of twins in total with an age vari-

Database	Num. of pairs	Features	Total	Capture Apparatus	Reference
CASIA-Iris-Twins	100	Iris images	3183 images	OKI's IRISPASS-h camera	[17]
3D TEC	107	3D face scans with neutral and smiling expressions	424 scans	Minolta Vivid 910	[1]
ND-TWINS-2009-2010	217	face images with five pose, under indoor and outdoor settings	24050 images	Nikon D90 SLR camera	[15]
Our Database	39	frontal face and profile images, facial motion videos, audio records	234 images, 1950 videos, 239 audio records	Canon 350D DSLR, Cannon PowerShot S5 IS, Sony HD video camera, Audio-Technica Condenser Microphone	

Table 1. Available twin databases.

ation ranging from 7 to 52. Most of them are Chinese. Our database contains three subsets: high resolution face images, facial motion videos and audio records. To the best of our knowledge, this is the first twins database that contains both videos and audio records. In the remainder of this paper we describe the collection procedure, the organization of the database, usage protocols, as well as possible usage of the database via some examples.

3. Collection Procedure

The audio-visual twins database was collected at the Sixth Mojiang International Twins Festival held on 1 May 2010 in China. Our twins database collection involved a half day of setup and equipment testing, followed by a single day of data collection. The collection procedure includes three partitions: photography, facial motion video recording and audio recording.

3.1. Environmental Setup and Capture Apparatus

The collection procedure was performed in two studios inside tents. One studio was used for acquisitions of 2D face photographs, and the other one was used for acquisition of video and audio recordings. A green canvas was set up as background in both studios. Several LED lights were used to provide full illumination on subjects. A Canon 350D DSLR Camera and a Canon PowerShot S5 IS were used to capture the face images. A Sony HD video camera was used to record videos and an Audio Technica Condenser Microphone was used to amplify sounds.

3.2. Subjects

We collected data from 39 pairs of twins, among which 2 pairs were American, 3 pairs were Russian, and others were Chinese. The participants varied from 7 to 54 in age. Most of them were between 7 and 22. A histogram of the distribution of ages is shown in Figure 1. Twins suffering from severe myopia were excluded from our collection.

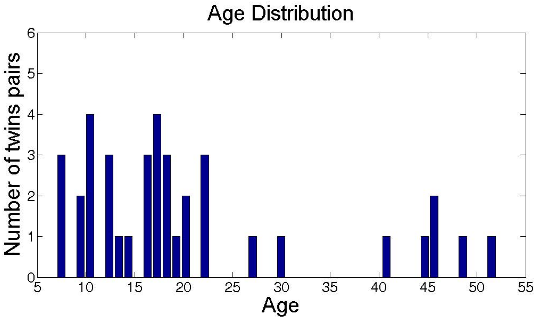


Figure 1. The age distribution of twins in our database.

3.3. Capture Procedure

The whole capture procedure contains 3 sessions taking about 20 minutes for each subject. In image session, the subject was required to remove their eyeglasses and face the camera with neutral expression. We then took photographs from the front as well as both sides of the subject.

In facial motion video session, we considered 6 expressions: smiling, anger, surprise, sadness, fear and disgust. We also included one record of free talking which captured subjects' facial motion in a more realistic situation. Subjects were verbally guided to show different expressions instead of imitating example expressions in pictures. We led each subject through the following steps.

1. **Smiling:** We verbally asked the subject to perform a slow smile from a neutral expression while we captured a video with camera. The procedure was repeated 3 times with an interval of $3-5$ seconds.

2. **Anger:** We repeated the previous step but asked the subject to perform anger expression.

3. **Surprise:** We repeated the previous step but asked the subject to perform surprise expression.

4. **Sadness:** We repeated the previous step but asked the subject to perform sad expression.

5. **Fear:** We repeated the previous step but asked the subject to perform fear expression.

6. **Disgust:** We repeated the previous step but asked the subject to act perform disgust expression.

7. **Free talking:** We asked the subject to answer a few easy questions in front of the camera while we captured a video. There were no standard answers and the subject was free to say anything for the questions. The capture procedure of free talking lasted about 2 minutes.

In audio record session, the subject was asked to read three pieces of short texts in either Chinese or English. For Chinese subjects, they were asked to read these texts in Chinese: (1) “1, 2, 3, ..., 10”; (2) a paragraph from a Chinese essay; (3) a famous Chinese poem “Longing in the night” by Bai Li. For other subjects, they were asked to read these texts in English: (1)“1, 2, 3, ..., 10”; (2) translation of the paragraph from the Chinese essay; (3) lyrics of the song “Seasons in the sun” by Westlife. To make it more natural for English subjects, they were also asked to sing the song “Seasons in the sun”. Thus, there are four audio records for each English subject. All the texts were repeated 3 times in each record with an interval of 3 seconds. Each record lasted 30 seconds in average. While recording, the subject was required to speak as clearly as possible and avoid making mistakes.

4. Database Organization

The database contains 266 high resolution images, 1714 short facial motion videos, and 239 audio records. The total size of the whole database is around 17.8GB.

4.1. Twins Images

Our twins images dataset contains one frontal face image and two profile images for each subject in neutral expression. The resolution of images is 3456×2304 pixels. An example is shown in Figure 2. This subset of data would be useful for evaluating performance of face recognition algorithms for twins identification problem. As the profile images are in high resolution they are able to provide discriminative information for ear recognition of twins.

4.2. Facial Motion Videos

Facial motion videos dataset contains two parts: (1) six expressions motion videos which last for 3 seconds in average, including smiling, anger, surprise, sadness, fear, disgust, and (2) a free talking video which lasts for about 2 minute. The six expressions are illustrated in Figure 3. Each expression motion was recorded at least three times. The expression motion videos could be useful in two ways: (1) they can be used as the video samples for twins facial motion recognition; (2) images with different expressions (such as Figure 3) could be clipped from the videos and used to evaluate performance of twins face recognition methods robust to expressions. The free talking videos would be helpful to provide extra data support for text-independent facial motion recognitions. Their audio component could be also exploited with the audio records in the following part as additional data support for text-independent twins speaker recognition.

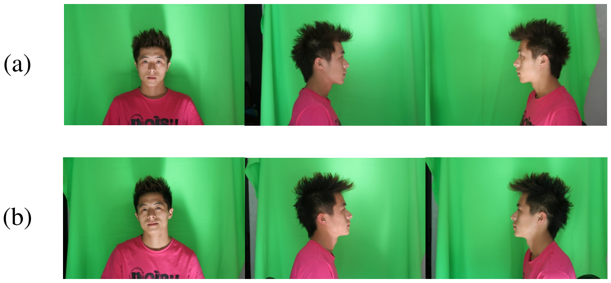


Figure 2. An example of twins images. A frontal face image and two profile images from both sides were captured for each subject. The elder twin is illustrated in (a), while the younger twin is illustrated in (b).

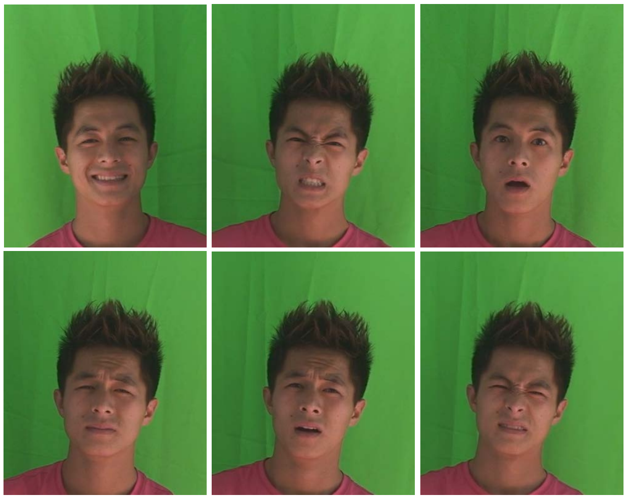


Figure 3. An illustration of six expressions. The subject is showing smile, anger, surprise, sadness, fear and disgust expressions in order.

4.3. Audio Records

The audio records dataset contains audio records of three different texts for each subject. The audio records last for 30 seconds in average. An example is shown in Figure 4.

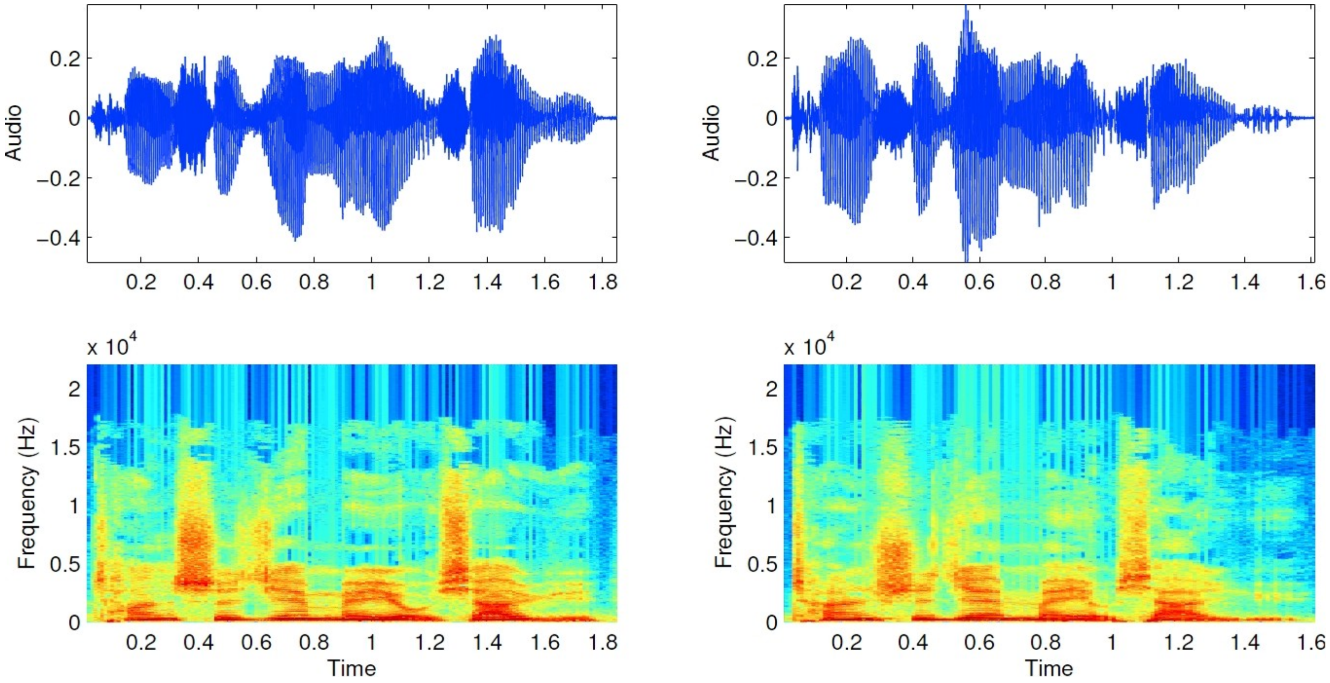


Figure 4. Example audio signals (first row) and corresponding spectrograms (second row) of identical twins reading the same text. Each column is from one sibling of the twins.

This subset would be useful for twin speaker recognition, in both text-dependent and text-independent settings. For text-independent speaker recognition, the free talking audio components in the facial motion videos dataset could offer additional data supports.

5. Protocol Specification

In order to provide a fair evaluation metric for twins recognition methods, we suggest protocols for using our database in three common tasks: twins verification, twins identification and pairwise twins similarity.

5.1. Twins Verification

In this task, a decision is made to accept or reject a given probe and its claimed identity. We divide the database into two parts, one for training ("MotionTraining.txt" and "AudioTraining.txt"), and one for testing ("MotionTesting.txt" and "AudioTesting.txt"). In the training set, there are 12 facial motion videos (2 for each expression) and 6 audio records (2 for each texts) for each subject. The probes in the testing set are tested against both of the corresponding twins.

Twins verification can be performed under text-dependent and text-independent settings. In text-dependent verification, the speech contents (or expressions) of training samples and testing probes should be of the same. In text-independent verification, samples with different speech contents (or expressions) are used in the training and testing procedure. In other words, text-independent verification should be robust to difference in speech contents (or expressions) while text-dependent verification is not necessarily robust to difference in speech contents (or expressions).

Methods for verification are usually evaluated by false accept rate (FAR), false reject rate (FRR), equal error rate (EER). FAR measures the percent of unmatched probe-identity pairs which are incorrectly accepted. FRR measures the percent of matched probe-identity pairs which are incorrectly rejected. For convenience, EER, where FRR and FAR meet, is used to evaluate the methods for verification. In many cases, accuracy rate, computed as $1-\text{EER}$, is also used to describe the performance of a verification method.

5.2. Twins Identification

This task recognizes and estimates an identity for a given probe. The division of database for this task is the same with that for the task of twins verification: "MotionTraining.txt" and "AudioTraining.txt" for training, and "MotionTesting.txt" and "AudioTesting.txt" for testing. However, unlike twins verification, this task does not require claimed identity for the testing probe. Indeed, it should estimate the identity for the probe.

Twins identification methods are evaluated by accuracy, which measures the percent of correctly identified probes.

5.3. Pairwise Twins Similarity

In this task, a similarity score of a given pair of probes (any format of facial motion videos or audio records) is computed based on given training dataset. The pair of probes can be either from the same subject (positive pair) or from two of a twin siblings (negative pair). Similar to twins verification, this task can also be performed in either text-dependent or text-independent settings.

We use the same training set (“MotionTraining.txt” and “AudioTraining.txt”) as for the twins verification task. For text-dependent testing, we randomly produce 2000 pairs of probes with 900 positive pairs and 1100 negative pairs for facial motion videos, and 800 pairs of probes with 400 positive pairs and 400 negative pairs for audio records. The testing probe pairs are given in “DepenentMotionSimilarityTesting.txt” and “DependentAudioSimilarityTesting.txt”. For text-independent testing, we randomly produce 10000 pairs of probes with 5000 positive pairs and 5000 negative pairs for facial motion videos, and 2500 pairs of probes with 1000 positive pairs and 1500 negative pairs for audio records. The testing probe pairs are given in “IndepenentMotionSimilarityTesting.txt” and “IndependentAudioSimilarityTesting.txt”.

To evaluate the performance of methods for this task, a threshold θ can be designed to convert the task to a verification problem. A probe pair with a similarity score larger than θ is regarded as accepted, otherwise is regarded as rejected. Then FRR, FAR, EER can be used to evaluate the methods for this task.

6. Experiments on the Database

To provide usage examples of our database, we performed three experiments on face verification, facial motion verification and speaker verification for twins.

6.1. Twins Face Verification

Based on the database, we implemented experiments to solve the twins face verification problem, using three traditional approaches: Eigenface [19], Local Binary Pattern [2] and Gabor filters [7]. We randomly selected 8 images clipped from the motion videos for each subject. The images were then registered by eye positions detected by STASM [14] and resized to 160×128 pixels. For Eigenface, we vectorized gray intensity in each pixel as feature and perform PCA to reduce the dimension. For LBP, we divided the image into 80 blocks. For each block, we used normal lbp feature and extracted a 59-bins histogram. For Gabor, we used 40 Gabor (5 scales, 8 orientations) filters and set the kernel size for each Gabor filter to 17×17 pixels. A PCA was performed to reduce the feature dimension for LBP and Gabor. We used k-nearest neighbour algorithm to verify the twin identities and chose k with the best result for each method. In order to verify the challenge of face recognition for twins, we conducted an additional experiment for general population. The training procedure followed the steps of Gabor method, whereas, in testing stage, we randomly selected 2000 pairs of image and identity from the pre-processed data as probe and claimed identity.

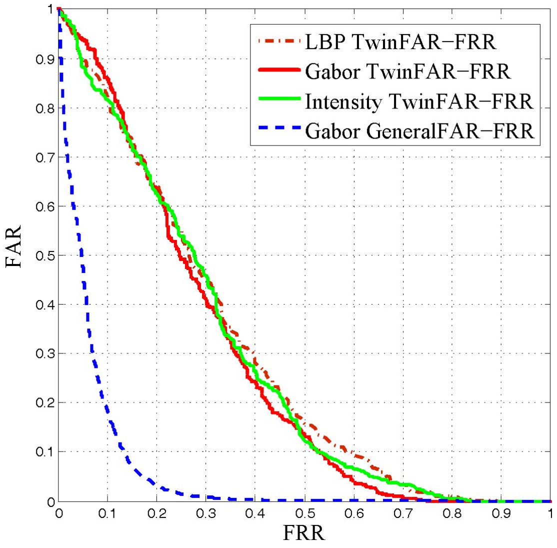


Figure 5. Accuracy of different face recognition approaches

Experimental results are shown in Figure 5. This figure shows that there is little difference between Intensity, LBP and Gabor for twin verification. The EERs for them are 0.352 (PCA), 0.340 (LBP) and 0.338 (Gabor). Through a comparison between Gabor method for twin siblings and general population, we can see that twins face recognition is much difficult than face recognition for general population. The EER of Gabor method for general population is around 0.122, while EER of Gabor method for twins is around 0.330.

6.2. Twins Facial Motion Verification

Based on the facial motion videos, we carried out experiments to distinguish between twins using a sparse displacement algorithm (SDA) [18] and a dense displacement algorithm (DDA) [20]. For each subject, we used twelve video clips, two for each of the six expressions. For the sparse displacement method, we tracked several key points at the neutral and apex of different expression face. We then calculate the displacement of these points and regarded the displacement vector as feature. The probe motion displacement was then compared with that of the same expression from claimed subject. For the dense sparse displacement method, we tracked the face along the video and used eye-center position for alignment. Then we warped the face to meanface and constructed the deformation feature as in

	Smile	Anger	Surp.	Sad	Fear	Disg.
SDA	0.833	0.889	0.907	0.778	0.889	0.800
DD	0.933	0.917	0.857	0.917	0.839	0.768

Table 2. Performance of SDA and DDA in each expression

[20]. We then carried out pairwise verification and computed the verification score as the weight summation of deformation feature similarity.

We conducted the experiments for all six expressions. The experimental results are shown in Table 2. On the whole, DDA performs better than SDA, since the overall accuracy is 0.864 in DDA and 0.850 in SDA. And in some particular expressions, smile, anger and sad, DDA performs dramatically better than SDA, because the average accuracy of SDA in these expression in 0.833, while the average accuracy of DDA in these three expressions is 0.922. However, DDA needs more computation and requires more stable displacement tracking, because DDA extracts dense displacement from each pixel rather than sparse displacement.

6.3. Twins Speaker Verification

Based on the audio records of our database, we performed experiments for twins speaker verification using Gaussian Mixture Model (GMM) method under both text-dependent and text-independent settings.

For each subject, we used three audio records of three repetitions of different texts. At first, the each record was cut into three clips, each of which contained one repetition of the text. Thus, there were 9 audio clips of 3 different texts for each subject. For text-dependent verification, the database was divided into 117 sets according to the texts and pairs of twins. Each set contains 3 repetitions of the same text by each pair of twins. We used two of the audio clips as gallery for training for each subject and the remaining one as a probe for testing. For text-independent verification, the database was divided into 39 sets according to pairs of twins. Each set contains 3 repetitions of 3 different texts by each pair of twins. We used the clips of 2 texts as gallery for training for each subject and the clips of remaining one text as probes.

Before training, we first framed all the audio clips and extracted 10 features from each frame. These features includes: Voice probability, Pitch [21], Energy, LPC [4], Spectral Centroid, Spectral Rolloff, Spectral Flux, MFCC [12] and other spectral statistics, such as spectral variance, skewness, kurtosis, slope, and the positions of spectral maximum and minimum. The voice probability characterizes the probability of frame to be speech contents of people. The voice probability is low during the breaks of the texts. By controlling the threshold for the voice probability, we filtered out unreliable frames that contain noises instead of human voice. After the filtering, we learned a Gaussian Mixture Model of each subject in all sets. The covariance matrix of all GMMs is assumed to be diagonal, whereas the number of components for GMMs changes with the feature types. In the testing phase, each probe was tested on the two GMMs trained on the corresponding pair. The number of components for GMMs was optimized on the test sets for better performance.

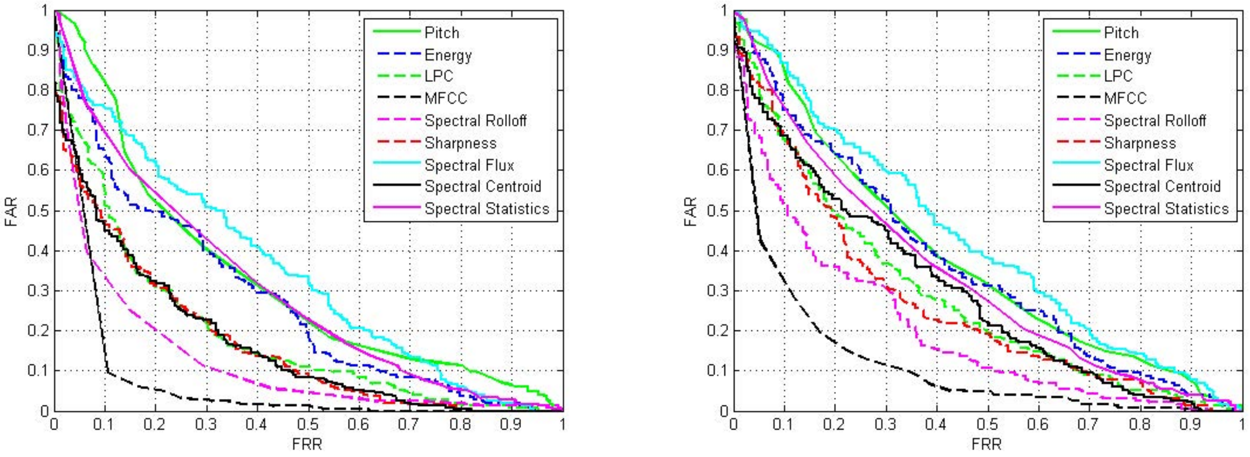
Experimental results are shown in Figure 6. As can be seen from, these 9 features are all effective in both text-dependent and text-independent verifications. For both text-dependent and text-independent verification, the performance of 9 features generally follows the order as: MFCC, Spectral Rolloff, Sharpness, LPC, other Spectral statistics, Energy and Pitch. MFCC has turned out to be the feature with the best performance for the voice verification. This is because MFCC contains vectorial data of the frequency over one frame, while other features such as Pitch, Sharpness and Energy contains only scalar data learned from the frames. Spectral Rolloff, being as the threshold frequency to measure the spectral shape of the subject, also performs well for the verification. It is more discriminative than other features such as Pitch, Energy *etc*. due to that it contains statistics information for the whole frame. Meanwhile, the results agree with the common sense that performance of text-dependent verification is better than the performance of text-independent verification. By comparing Figure 6(a) and Figure 6(b), we can see that EER rates for the chosen 9 features in text-dependent test are all less than those in text-independent test. This is because these features are often affected by the specific texts to some extent. Features of two audio clips concerning different short texts by the same subject may change a bit.

7. Conclusions

Through this paper, we have described the collection procedure and organization of our visual-audio twins database. We also have pointed out the possible usage of the database, as summarized as follows:

- Evaluation of twins face recognition algorithms robust to expressions.
- Evaluation of twins ear recognition algorithms.
- Evaluation of twins facial motion recognition algorithms.
- Evaluation of twins speaker recognition algorithms.
- Evaluation of multi modal approaches for twin recognition based on face, ear, expression motion and voice.

We have suggested some protocols for three tasks. We conducted three experiments to show the usage of images, facial motion videos and audio records of the database.



(a) Text-Dependent Accuracy

(b) Text-Independent Accuracy

Figure 6. Performance comparison between text-dependent and text-independent speaker verification

8. Obtaining the Database

Anyone interested in receiving the database could visit the web site at http://www.comp.nus.edu.sg/∼face. In the near future, we will upload the database to server of Computer Science Department of National University of Singapore (NUS) and new download address will be updated in the database web site.

9. Acknowledgments

We would like to thank Yuan Cheng and Vlad Hosu for the discussion about this paper. We would also like to thank staff and all the participants of the Sixth Mojiang Internationl twins Festival for their contribution.

References

[1] Cvrl dataset.

[2] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):2037–2041, 2006.

[3] A. Ariyaeeinia, C. Morrison, A. Malegaonkar, and S. Black. A test of the effectiveness of speaker verification for differentiating between identical twins. *Science & Justice*, 48(4):182–186, 2008.

[4] B. S. Atal and S. L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, 50(2B):637–655, 1971.

[5] A. Bronstein, M. Bronstein, and R. Kimmel. Three dimensional face recognition. In *Numerical Geometry of Non-Rigid Shapes*, pages 1–15. Springer, 2009.

[6] D. CN, S. P. Sankar, and N. George. Multimodal identification system in monozygotic twins. *International Journal of Image Processing (IJIP)*, 7(1):72, 2013.

[7] I. Fogel and D. Sagi. Gabor filters as texture discriminator. *Biological cybernetics*, 61(2):103–113, 1989.

[8] IEEE. *Effectiveness of LP derived features and DCTC in twins identification-Iterative speaker clustering approach*, volume 1, 2007.

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

[10] B. Klare, A. A. Paulino, and A. K. Jain. Analysis of facial features in identical twins. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–8. IEEE, 2011.

[11] A. W.-K. Kong, D. Zhang, and G. Lu. A study of identical twins palmprints for personal verification. *Pattern Recognition*, 39(11):2149–2156, 2006.

[12] B. Logan et al. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, 2000.

[13] J. A. Martin, H.-C. Kung, T. Mathews, D. L. Hoyert, D. M. Strobino, B. Guyer, and S. R. Sutton. Annual summary of vital statistics: 2006. *Pediatrics*, 121(4):788–801, 2008.

[14] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. *ECCV*, pages 504–513, 2008.

[15] P. J. Phillips, P. J. Flynn, K. W. Bowyer, R. W. V. Bruegge, P. J. Grother, G. W. Quinn, and M. Pruitt. Distinguishing identical twins by face recognition. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 185–192. IEEE, 2011.

[16] S. Srihari, C. Huang, and H. Srinivasan. On the discriminability of the handwriting of twins. *Journal of Forensic Sciences*, 53(2):430–446, 2008.

[17] Z. Sun, A. A. Paulino, J. Feng, Z. Chai, T. Tan, and A. K. Jain. A study of multibiometric traits of identical twins. In

SPIE Defense, Security, and Sensing, pages 76670T–76670T. International Society for Optics and Photonics, 2010.

[18] S. Tulyakov, T. Slowe, Z. Zhang, and V. Govindaraju. Facial expression biometrics using tracker displacement features. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–5. IEEE, 2007.

[19] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[20] N. Ye and T. Sim. Towards general motion-based face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2598–2605. IEEE, 2010.

[21] R. J. Zatorre, A. C. Evans, E. Meyer, and A. Gjedde. Lateralization of phonetic and pitch discrimination in speech processing. *Science*, 256(5058):846–849, 1992.