A Study of Face Recognition of Identical Twins by Humans

Soma Biswas, Kevin W. Bowyer and Patrick J. Flynn
Dept. of Computer Science and Engineering
University of Notre Dame.
{sbiswas, kwb, flynn}@nd.edu

***Abstract*—Recent studies have shown that face recognition performance degrades considerably for images of identical twins. Human face matching capability is often considered as a benchmark for assessing and improving automatic face recognition algorithms. In this work, we investigate human capability to distinguish between identical twins. If humans are able to distinguish between facial images of identical twins, it would suggest that humans are capable of identifying discriminating facial traits that can potentially be useful to develop algorithms for this very challenging problem. Experiments with different viewing times and imaging conditions are conducted to determine if humans viewing a pair of facial images can perceive if the image pairs belong to the same person or to a pair of identical twins. The experiments are conducted on 186 twin subjects, making it the largest such study in the literature to date. We observe that humans can perform the task significantly better if they are given enough time and tend to make more mistakes when images differ in imaging conditions. Our analysis also suggests that humans look for facial marks like moles, scars, etc. to make their decision and do worse when presented with images lacking such marks. Experiments with automatic face recognition systems show that human observers outperform automatic matchers for this task.**

I. Introduction

Recognition of facial images of identical twin siblings poses a considerable challenge for any face recognition algorithm because of the strong similarity between the face images. Recent research has showed that the performance of automatic face recognition technology deteriorates drastically when the images belong to identical twin siblings as compared to when they correspond to unrelated persons [1]. The degradation is shown to be far more drastic for face than for other biometrics such as iris and fingerprint.

Humans are very good at identifying people from their images, and so human face recognition performance is often considered as a guideline for assessing face recognition algorithms [2]. To the best of our knowledge, no systematic human study has been performed that addresses the task of distinguishing between identical twins from their face images. Here, we perform experiments to determine if humans viewing a pair of facial images can perceive whether the images belong to the same person or to a pair of identical twin siblings. If humans are able to distinguish between facial images of identical twin siblings, it might mean that they are capable of observing discriminating traits which can potentially be used to improve the performance of face recognition technology.

In this investigation, human participants view pairs of facial images and respond according to their level of certainty whether they belong to the same person or to identical twins. First, we study the human performance when the participants view the images for a limited time of two seconds, which has been shown to be sufficient for matching images of unrelated persons [2]. We conduct another experiment to analyze whether humans can do better when the viewing time is increased. The variation in performance when the input images are taken under controlled indoor conditions or in an outdoor environment is also analyzed as part of the second experiment. We also study which facial features are most useful for humans to correctly distinguish between identical twins. The human performance is also compared against traditional and commercial automatic face matchers. The results of this investigation can be used to improve the performance of existing face recognition algorithms so that they are more suited to handle the challenges posed by facial images of identical twin siblings.

The rest of the paper is organized as follows. Section II discusses the related work. The dataset and the experimental setup of the first experiment is discussed in Section III and the analysis is given in Section IV. The follow up experiment is described in Section V. In section VI, human performance is compared against automatic face matching algorithms. The paper concludes with a summary and discussion.

II. Related Work

In this section, we discuss related work in the literature. Since identical twins cannot be distinguished by their DNA, there is increased interest in using different biometric traits for distinguishing between identical twins. Modeling facial expressions as isometries of the facial surface, Bronstein *et al.* [3] proposed an expression-invariant 3D face recognition approach which was successful in distinguishing one set of identical twin siblings. A hybrid feature by combining the traditional holistic facial appearance feature with a facial dynamics feature has also been shown to be successful in distinguishing between facial images of one pair of identical twins [4]. A face recognition system based on an optical recognition principle was also shown to be successful in distinguishing between identical twin siblings for a database

*This work is supported by the Federal Bureau of Investigation (FBI), the Biometrics Task Force and the Technical Support Working Group through US Army contract W91CRB-08-C-0093. The authors would like to thank Prof. Alice O'Toole for her valuable comments.

of ten pairs of subjects [5]. Recently, soft biometrics like facial marks have been used to differentiate identical twins [6] on a dataset which contained facial images from five pairs of identical twins. Since they were tested on very small number of twin pairs, the conclusions may not be statistically significant. Recently, Sun *et al.* [1] conducted unimodal and multimodal matching experiments on fingerprint, face and iris biometrics collected from 66 pairs of identical twins. They showed that it is much easier to distinguish between identical twin siblings using iris and fingerprint biometrics compared to using facial images. There has been some work on distinguishing between identical twins based on other biometrics like palmprint [7], fingerprint [8], iris [9], speaker identification [10], etc.

Humans are naturally trained to recognize faces from birth and there is strong evidence that suggests face recognition activity in humans takes place in the fusiform face area of the cortex [11]. Thus there has been a lot of interest in developing algorithms which replicate the human visual processing for face recognition. For example, biologically inspired features in the form of Gabor wavelets have been successfully used for recognizing faces [12]. It has been also seen that the performance of automatic algorithms can be considerably improved by fusing it with human performance [13]. Though quite a few human studies have been conducted in the past to study various aspects of general face matching problem [2], to the best of our knowledge, there has been no systematic study of human ability to distinguish between identical twins from their face images.

III. Experimental Setup

In this section, we describe the dataset used, the participants who took part in the study and the experiment protocol.

Dataset: The twins data used in our study was obtained from data collection sessions at the Twins Days Festival in Twinsburg, Ohio in August 2009 [14]. The dataset consists of 186 subjects, of which 34 are male and the remaining 152 are female. The twins participating in the data collection self-reported themselves as identical twins. No DNA testing was performed to confirm the claims. All data collected at the festival followed a data collection protocol approved by the Human Subjects Institutional Review Board (HSIRB) at the University of Notre Dame.

Participants: A total of 23 volunteers (all of them were students or staff members of University of Notre Dame) were recruited to participate in the recognition experiment. They did not receive any prior practice or training for the task. The volunteers were offered ten dollars for participation, and an additional five dollars if they correctly classified 80% or more of the image pairs. The experiment was approved by HSIRB at the University of Notre Dame.

Experiment Protocol: In the experiment, the participants were given a brief verbal description of the study, and asked to read and sign an informed consent form. Then they were asked to start a computer program that presented instructions along with a few sample trials. This was followed by 180 trials out of which 90 trials corresponded to match (same person) pairs while the other 90 corresponded to non-match (identical twin) pairs. The match and non-match pairs were interspersed and presented in a random order to each participant. The images used in this experiment were captured in an indoor environment with controlled lighting, frontal pose and neutral expression. The images were cropped based on the eye locations so that only the face portion was visible. This ensured that the responses of the participants were not affected by external factors like clothing, hair style, etc.

In each trial, the computer program displayed a pair of facial images followed by a prompt for a decision on whether they correspond to the same person or to identical twin siblings. The images were displayed for two seconds. (This is guided by the study presented in [2] that indicated that human performance to distinguish between faces of unrelated persons does not show significant improvement if humans are allowed to observe image pairs for more than two seconds.) Then the participants were asked the following question: *Are the two images of the same person or of identical twin siblings?* They were required to select one out of five possible responses.

1) Sure they are the same person
2) Think they are the same person
3) Don't know
4) Think they are identical twin siblings
5) Sure they are identical twin siblings

Fig. 1 and Fig. 2 show examples of face images displayed. The two images in Fig. 1 belong to the same person whereas the two images in Fig. 2 belong to identical twin siblings.

Fig. 1. An example of a pair of displayed images. Here the images are of the same person.

IV. Experimental Results and Analysis

In this section, we describe in detail the results obtained from the experiment described above.

A. Are humans able to distinguish between identical twins?

To find the overall accuracy, we count the number of times each participant correctly classified the pair of images to be of the same person or of identical twin siblings. For example, if the two images are of the same person, we consider the responses *Sure they are the same person* and *Think they are the*

Fig. 2. An example of a pair of displayed images. Here the images are of identical twin siblings.

same person as correct responses. Similarly, if the two images are of identical twin siblings, we consider the responses *Sure they are identical twin siblings* and *Think they are identical twin siblings* as correct responses. Across the 23 participants, the maximum accuracy attained is 90.56% and the minimum accuracy is 60.56%. The average accuracy is 78.82% (standard deviation 8.9%). We use a one-tailed t-test to evaluate the null hypothesis that humans did not perform better on this task than random guessing. The resulting p-value is 1.4×10^{-13}. Thus, we have statistically significant evidence that the participants performed better than random.

The Receiver Operating Characteristic (ROC) of the performance is shown in Fig. 3 (blue dotted curve). The ROC provides a complete picture of human accuracy in distinguishing between identical twins at the assessed confidence threshold levels (1 through 5) and is drawn using the same procedure as in [2]. The verification rate or hit rate is computed as the proportion of matched pairs correctly judged to be of the same person. The false acceptance rate or false alarm rate is calculated as the proportion of non-match pairs judged incorrectly to be of the same person.

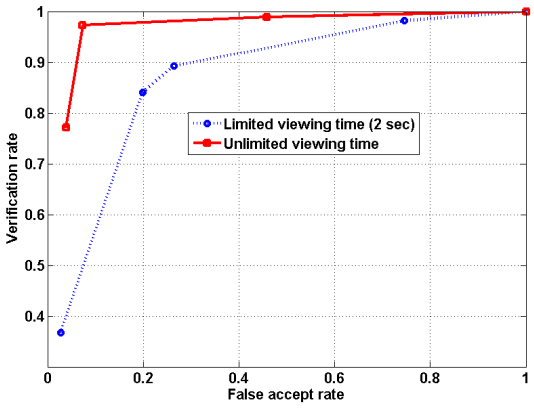

Fig. 3. Human performance when the image pairs are viewed for limited time (two seconds) vs. unlimited time.

B. Do humans perform better when they are certain?

As discussed earlier, the participants had the choice to respond according to their level of confidence. For example, if the participant felt that the two images belonged to the same person, they could choose either *Sure they are the same person* or *Think they are the same person*, depending on their confidence level. We observe that the confidence level varies significantly across the participants. On one hand, one of the participants was certain (i.e chose Options 1 or 5) for 118 out of 180 trials, while on the other hand, three participants were not certain of any of their responses (i.e. did not choose Options 1 or 5 even once). The average number of certain responses across all participants was 60 out of 180 trials.

Considering the trials for which the participants were certain about their response, the average accuracy of correct classification of the image pairs as belonging to the same person or to a pair of identical twins is 93.12%. Thus the performance is significantly better on the subset of trials where the participants were certain about their response.

C. Self-learning

The participants who volunteered for the study did not receive any prior training to classify images of identical twins and none of them had an identical twin sibling. We want to analyze whether the participants can learn by themselves the subtle differences between the facial images of identical twins. The average improvement in the performance in the second half of the trials as compared to the first half is 1.5%. Out of the 23 participants, 14 performed better in the second half, while only seven performed better in the first half.

This improvement might mean that as the participants viewed more images of twins, they trained themselves and performed better in the second half of the trials. A one-tailed t-test shows that the difference is not statistically significant (p-value 0.3065). In our study, the participants did not receive any feedback after each visual stimuli whether their response was correct or not. Providing feedback on their response could have helped the participants learn better and thus perform better in the second half as compared to the first half.

D. Are males more easy to classify than females?

Several researchers have studied the effect of gender on the face recognition performance, and though individual studies find men or women easier to recognize, there is no consistent gender effect [15]. Here, we investigate if such gender effects are present when humans are asked to distinguish between identical twins.

The number of male pairs in the twins data used for the experiment is considerably lower than that of female pairs. The total number of correct and incorrect responses for male pairs is 571 and 153 respectively with an accuracy of 78.87%. On the other hand, the total number of correct and incorrect responses for female pairs is 2693 and 723 respectively with an accuracy of 78.84%. So there is no significant difference in matching accuracy for male and female pairs. Similar results have been reported for facial images of unrelated persons [15].

V. Experiment with unlimited viewing time

We see from the above analysis that humans do not perform very well in distinguishing between images of identical twin siblings given just two seconds to view images. Studies have shown that for face images of unrelated persons taken under different illumination conditions, increasing the viewing time of the images beyond two seconds does not significantly increase the recognition performance of human participants [2]. So we design another experiment to test if humans can do a better job in distinguishing between identical twins when given sufficient time. We also want to test how external imaging factors affect the human performance and understand which facial features are most important for humans to distinguish between identical twin siblings.

The visual stimuli used for this experiment consisted of 100 pairs out of which 50 pairs were images captured under controlled indoor conditions while the remaining 50 pairs were captured in outdoor uncontrolled environment. For each subset of 50, 25 were match pairs (images of same person) and the other 25 were non-match pairs (images of identical twin siblings). The four subsets of image pairs were interspersed and presented in a random order to the participants and the order was different for each participant. In this experiment, the participants were given unlimited time to view the image pairs and make their decision. As in the first experiment, each trial consisted of the computer program displaying a pair of facial images with a prompt for a decision on whether they belong to the same person or identical twin siblings. The participants were also asked which features helped them make the decision. They were given the following choices: *Eyebrows, Eyes, Nose, Lips, Moles/Scars/Freckles, Skin color/Texture, Wrinkles, Facial hair* or *Make-up*. The participants were also asked to make note of any features which were not there in the list. Based on their decision, the participants were also asked to mark the facial features that helped them make the decision.

A. Do humans perform better when given unlimited time?

One of the goals of this experiment is to explore if increasing the viewing time makes it easier for humans to distinguish between identical twins. Across the 25 participants, the maximum accuracy attained is 100% and the minimum accuracy is 78%. The mean recognition accuracy for this experiment is 92.88% and the median is 95%. Fig. 3 (red solid curve) shows the ROC obtained.

For generating the ROC, we consider only the indoor image trials from the second experiment for fair comparison. We observe that increasing the viewing time significantly increases the matching accuracy. The significant improvement in performance with increase in viewing time that we observe in our experiment can be attributed to the fact that facial images of twins are very similar with only subtle differences which can be better perceived given sufficient time. Our interpretation is that the added time is used by the participants to consider local features in making their decision.

B. Do humans perform better on controlled image pairs than on uncontrolled pairs?

Fig. 4 shows a pair of uncontrolled images used in the experiment. Although the images are of the same person, they appear very different due to illumination effects.

Fig. 4. Example of a pair of images of the same person taken outdoors with uncontrolled illumination.

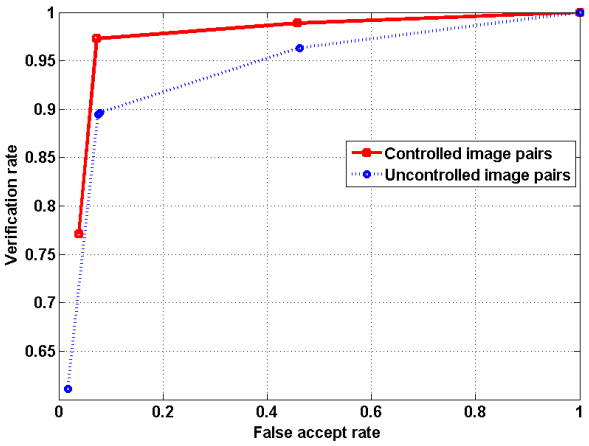

Fig. 5. Human performance for image pairs taken in controlled and uncontrolled settings.

As described earlier, out of the 100 image pairs used in the second experiment, 50 were controlled image pairs and the remaining 50 were uncontrolled image pairs with equal number of match and non-match pairs of each type. The mean accuracy obtained for the controlled image pairs is 94.96% and that for the uncontrolled image pairs is 90.80%. Fig. 5 shows the ROCs corresponding to the controlled and uncontrolled pairs. We observe that humans find it harder to match the uncontrolled image pairs as compared to the controlled pairs. We perform a one-tailed t-test to evaluate the null hypothesis that the performance for the controlled image pairs and the uncontrolled image pairs come from distributions with equal mean. The resulting p-value is 0.0035. Thus, we have statistically significant evidence that the participants did

better on the controlled pairs than on the uncontrolled pairs. So presence of external factors like illumination, etc. tend to make the already challenging problem of distinguishing identical twin siblings even more difficult.

C. What are the types of features that humans consider important to distinguish between identical twins?

In this study, given a pair of facial images, the participants were asked to choose the facial features which helped them in making their decision. Fig. 6 shows which feature types were chosen as important for the correct as well as incorrect responses. From the figure, it is evident that for the correct responses, the most important feature type chosen is moles/scars/freckles and it is significantly more important than any of the other types. For the incorrect responses, none of the features seem significantly more important than the others. This observation suggests that humans are more likely to be incorrect in their decision if they cannot find moles/scars/freckles in the images.

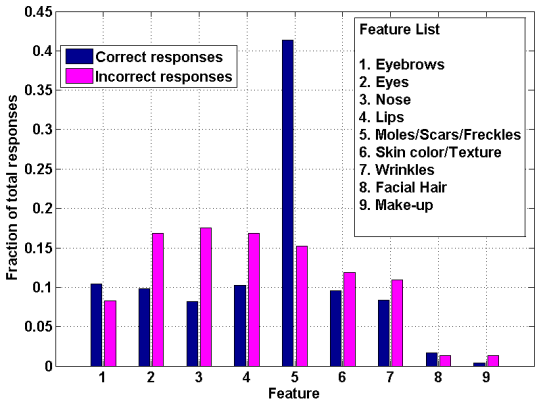

Fig. 6. Useful features for distinguishing images of identical twin siblings.

This is also evident if we analyze which image pairs were correctly and incorrectly classified by most of the participants. Fig. 7 (left) shows an image pair (match pair) which were correctly classified by all the participants. This image pair has moles which makes it easier to identify the images as those of the same person. Fig. 7 (right) shows an image pair (non-match pair) which was most incorrectly classified (accuracy $= 60\%$). None of the images have any moles/scars/freckles which makes it difficult for humans to respond correctly.

VI. Automatic Face Matching Performance

Now we investigate the ability of automatic face matching algorithms to distinguish between identical twins. We experimented with two commercial matchers (Pittpatt [16] and Cognitec [17]) in addition to standard holistic face matching approaches based on Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). The algorithms were tested for the same 100 pairs used in the second experiment. Both the traditional matchers and Pittpatt performed very poorly in this task. This result is in agreement with recent research [1] that has shown that currently available face recognition algorithms perform poorly on facial images of identical twins. Only Cognitec matcher performed comparable to humans and Fig. 8 shows the ROC obtained. The average human performance (computed from all the 25 participants) is also shown for comparison. As can be seen from Fig. 8, human observers outperform the automatic matcher for almost the entire range of False Accept Rate (FAR).

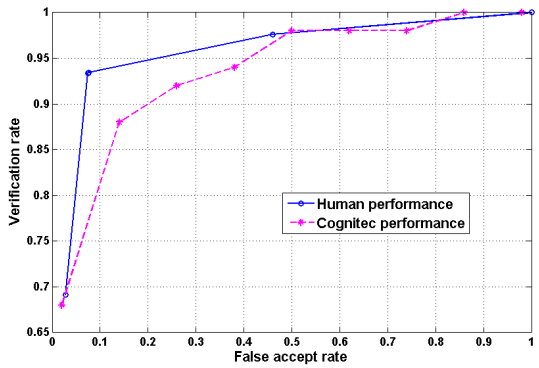

Fig. 8. Comparison of human performance against a commercial face recognition engine (Cognitec).

We further investigate if the automatic matcher and humans show similar behavior with regard to difficulty in distinguishing between identical twins. If algorithms and humans behave differently, human expertise can potentially be used to guide development of better computer algorithms . We analyze the non-match pairs that were found to be most difficult (corresponding to the highest similarity scores) by the automatic algorithm. Fig. 9 shows two non-match image pairs which got the highest similarity scores (greater than 0.9995) in the automatic experiment. But humans did reasonably well in distinguishing between these identical twin pairs (accuracy of 88% and 100% respectively). This indicates that human observers were able to capture facial characteristics different from the automatic algorithm that helped them do well on these pairs. We notice that these image pairs differ in moles/freckles distribution (a few are marked on the images in Fig. 9) and we conjecture that this may be a reason for the good performance of the humans. Therefore, one potential way to incorporate human knowledge to improve machine performance is to robustly detect facial marks and use them for facial characterization in addition to existing feature set used by automatic algorithms.

VII. Summary and Discussion

In this work, we performed a human experiment to determine if humans viewing a pair of facial images can successfully distinguish between images of the same person and of identical twin siblings. Given two seconds to view the image pairs, the average accuracy was found to be 78.82%. We observe that increasing the viewing time significantly improves

Fig. 7. Left: Example of an easy image pair (match pair) which were correctly classified by all the 25 participants; Right: Example of a difficult image pair (non-match pair) which were incorrectly classified most of the times with accuracy of 60%.

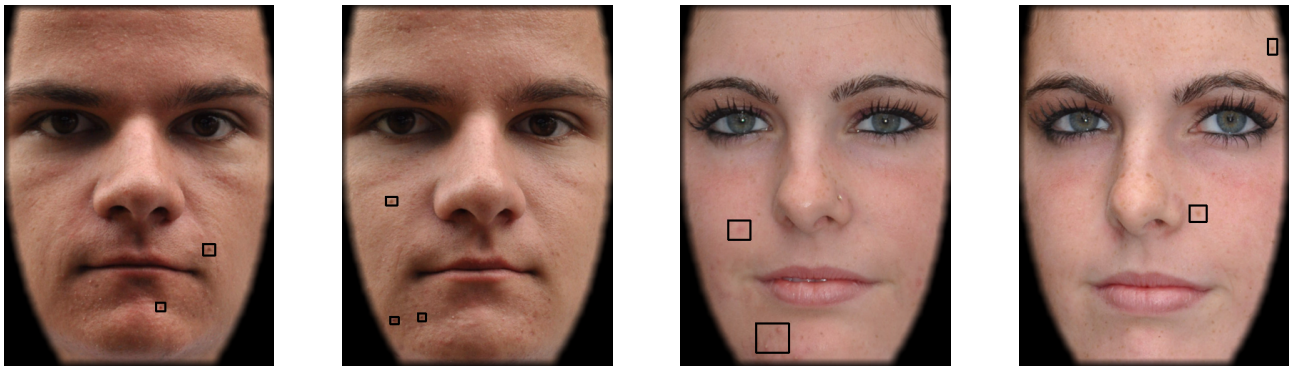

Fig. 9. Example of two non-match pairs which were given very high similarity score by Cognitec.

the matching accuracy. This can be attributed to the fact that facial images of identical twins are very similar with subtle differences which can be better perceived given sufficient time. We also observed that the performance was lower for the uncontrolled images as compared to the controlled images implying that the presence of external factors like illumination, etc. tend to make the already challenging problem of recognizing images of identical twin siblings even harder. For the correct responses the most important feature chosen was moles/scars/freckles while for the incorrect responses, none of the selected features were significantly more important than the others. We observed that humans perform much better than commercial face recognition algorithms. One potential way to improve machine performance is to robustly detect local facial marks and use them for facial characterization in addition to existing feature set used.

REFERENCES

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

[2] A. J. O'Toole, P. J. Phillips, F. Jiang, J. Ayyad, N. Pnard, and H. Abdi, "Face recognition algorithms surpass humans matching faces across changes in illumination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1642–1646, September 2007.

[3] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, "Three-dimensional face recognition," *International Journal of Computer Vision*, vol. 64, no. 1, pp. 5–30, August 2005.

[4] N. Ye and T. Sim, "Combining facial appearance and dynamics for face recognition," *LNCS*, vol. 5702, pp. 133–140, 2009.

[5] K. Kodate, R. Inaba, E. Watanabe, and T. Watanabe, "Facial recognition by a compact parallel optical correlator," *Measurement Science and Technology*, vol. 13, pp. 1756–1766, 2002.

[6] U. Park and A. Jain, "Face matching and retrieval using soft biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 406–415, September 2010.

[7] A. W. Kong, D. Zhang, and G. Lu, "A study of identical twins palmprints for personal verification," *Pattern Recognition*, vol. 39, no. 11, pp. 2149–2156, 2006.

[8] A. K. Jain, S. Prabhakar, and S. Pankanti, "On the similarity of identical twin fingerprints," *Pattern Recognition*, vol. 35, no. 11, pp. 2653–2663, 2002.

[9] K. Hollingsworth, K. Bowyer, and P. Flynn, "Similarity of iris texture between identical twins," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2010, pp. 22–29.

[10] A. Ariyaeeiniaa, C. Morrison, A. Malegaonkara, and B. Black, "A test of the effectiveness of speaker verification for differentiating between identical twins," *Science and Justice*, vol. 48, no. 4, pp. 182–186, December 2008.

[11] N. Kanwisher, J. McDermott, and M. M. Chun, "The fusiform face area: A module in human extrastriate cortex specialized for face perception," *The Journal of Neuroscience*, vol. 17, no. 11, pp. 4302–4311, 1997.

[12] E. Meyers and L. Wolf, "Using biologically inspired features for face processing," *International Journal of Computer Vision*, vol. 76, no. 1, pp. 93–104, January 2008.

[13] A. J. O'Toole, H. Abdi, F. Jiang, and P. Phillips, "Fusing face recognition algorithms and humans," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 35, no. 5, pp. 1149–1155, 2007.

[14] "Twins days festival official website," *http://www.twinsdays.org/*.

[15] Y. M. Lui, D. Bolme, B. A. Draper, J. R. Beveridge, G. Givens, and P. Phillips, "A meta-analysis of face recognition covariates," in *IEEE International Conf. On Biometrics: Theory, Applications And Systems*, 2009, pp. 1–8.

[16] "Pittsburgh pattern recognition," *http://www.pittpatt.com/*.

[17] "Cognitec: The face recognition company," *http://www.cognitec-systems.de/*.