Utilization of Matching Score Vector Similarity Measures in Biometric Systems

Xi Cheng, Sergey Tulyakov, and Venu Govindaraju
Center for Unified Biometrics and Sensors
University at Buffalo, NY, USA
xicheng,tulyakov,govind@buffalo.edu

Abstract

In biometric systems, people may be asked to provide multiple scans for redundancy and quality control. In the case of fingerprint matching systems, repeat fingerprint probes of the same physical finger can be available and data from such multiple samples can be fused for reliable authentication of individuals. Since multiple samples are from the same instance of the finger, some relationships between them, e.g. diversity or similarity, could be observed. In this paper, we investigate such relationships and use them in fusion in order to improve the performance of biometric systems. The relationships between samples are derived by measuring the similarity between matching score vectors with Pearson's correlation and cosine similarity measures. We conduct experiments using the FVC2002 dataset consisting of four fingerprint databases and trainable combination methods, likelihood ratio and multilayer perceptron. The results show that utilization of similarity measures for matching scores can further improve the multi-sample biometric fusion in both combination methods.

1. Introduction

Biometric systems [9] consisting of methods for recognizing people based on physical or behavioral human traits are widely used in our current society. Since unimodal biometric systems have limitations, such as non-universality, multibiometric systems are more and more used to achieve superior performance. Multibiometric systems use more than one biometric measures such as different modalities, multiple instances or samples, multiple algorithms of the same template. Multi-modal fusion was extensively researched in previous literature [13, 18, 15]. It is generally believed that, because of diversity of different modalities, fusion of such systems leads to better recognition results compared to unimodal systems.

The primary focus of this paper is to consider the fusion of multiple samples rather than multiple modalities. This case is natural in face tracking and recognition where videos contain multiple frames. Fusion of multiple frames may provide more information than one frame especially when frame is erroneous. In fingerprint systems, similar situation arises when first authentication attempt does not succeed, and the user might be asked to provide another fingerprint scan; the fusion of the results of two matching attempts could be performed for better performance.

Most of the traditional algorithms [13] for fusing biometric matching information assume that only limited number of scores is available, for instance, one matching score for one modality, and the other matching score for another modality. Instead of using single matching score for acceptance decision, it is possible to combine matching scores for all attempts and possibly increase the performance of biometric systems. Thus, the goal of fusion algorithm is to combine these scores in order to further separate genuine attempts against impostor attempts.

Matching a test template to the enrolled template generates one matching score which is a measure of similarity of templates. A fusion approach which simply combines scores from multiple test samples by some pre-determined rule, e.g. by averaging, is suboptimal since it does not consider the possibility that certain samples can have poor quality and others bear good quality. Preferably, good quality samples should somehow influence combined score more than poor quality ones. As additional factor, the diversity between multiple templates should also be considered. Indeed, the bigger the difference of the additional template with previous templates is, the more complementary information, useful for combination, it might have. On the other hand, if a second test template is very similar to the first one, it probably should be omitted from combination, since it does not provide any additional information.

In order to judge either quality or diversity between test templates, we can utilize the sets of scores obtained by matching these templates with all templates enrolled in the biometric database (Fig. 1). In this paper, we consider to use statistical measures of matching scores to reflect quality or diversity between multiple samples. Using this auxiliary information with matching scores can make a better authen-

978-1-4673-1612-5/12/$31.00 ©2012 IEEE

tication decision in our experiments.

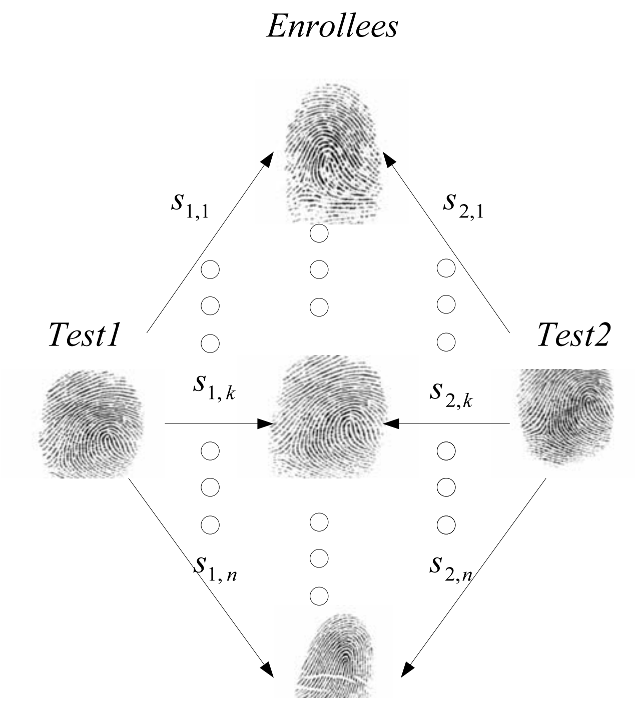


Figure 1. Sets of matching scores available for combinations of two test templates. In addition to using raw matching scores, $s_{1,k}$ and $s_{2,k}$, between test templates and enrolled template for person k, we want to utilize statistical measures for two sets of matching scores $\{s_{1,j}\}$ and $\{s_{2,j}\}$ between two test templates and all enrolled templates.

2. Previous Work

Fusion of multibiometric systems includes several different levels such as feature level, matching score level and decision level. Feature level fusion is extensively used in multi-sample literature where a more reliable template is constructed from several samples. Such approaches are used to generate a so called *super-template*.

Jain *et al*. [8] used a modified iterative closest point algorithm to get a transformation matrix defining the spatial relationship between two templates. After constructing a composite image of two fingerprint images, augmented minutiae sets were extracted. Experiments showed that mosaicking templates could improve the system performance compared to utilizing only one template.

Ryu *et al*. [11] proposed the concept *super-template* where highly credible minutiae from multiple fingerprints were incorporated. Update of likely true minutiae was achieved by a successive Bayesian estimation approach on a sequence of templates. Experimental results showed that better accuracy would be achieved if more impressions of the same finger were fused.

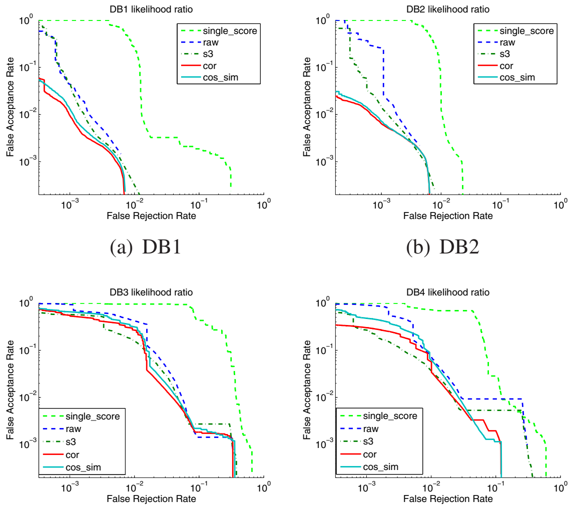
Literature [12] showed that, if extraction of reliable template was required, feature level fusion could be an important step. Exact minutiae positions were required in fuzzy fingerprint vault method. By estimating such positions using average of minutiae positions in few samples of finger, performance of fuzzy vault matching was improved.

In face recognition system, sequence of video frames could be used to generate a more precise model of face in [14]. A set of signatures could also be used to construct a template as a trained HMM in handwritten signature verification system [17].

Feature level fusion seems to be more effective than fusion at score level since features contain more information about biometric data. For example, feature's distribution could be better estimated from multiple samples than a single sample. But feature level fusion is more costly because we should access raw data to find correspondence between them. Generally, commercial matchers do not provide us such details. In addition, feature fusion algorithms are specific for each modality so that face and fingerprint systems should have different algorithms. In contrast, score level fusion can be widely applied to any biometrics. More important, most of work in feature level fusion, instead of making comparison to the result using score level fusion with same number of samples, only shows that their performance is better than utilization of one sample. The reality is, with good pre-determined rules, score level fusion which is more simple and flexible can get better results than using only one sample.

In [5], each frame extracted from a clip of video was matched to enrolled face templates resulting of a sequence of matching scores. Then evolving uncertainty of identity variables was captured by conditional entropy used to update combined matching scores. Uludag *et al*. [19] used the mean (or minimum) of similarity scores of query with templates of the claimed identity in fingerprint authentication system. Zhang and Martnez [20] explored score fusion with simple averaging in face recognition.

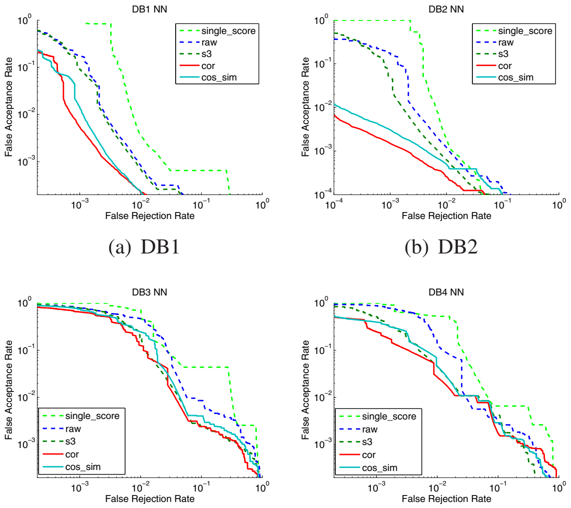
In this paper,we will consider verification system as Fig. 1 shows. Test template 1 and 2 from the same finger are to be authenticated to be person k. In addition to matching scores $s_{1,k}$ and $s_{2,k}$ between test templates and enrolled template for person k, other scores can be generated by matching the test templates to n enrolled templates, such as $s_{i,1}$, …, $s_{i,k}$, …, $s_{i,n}$ for test template i where $i = 1, 2$. These two sets of matching scores can be used to measure diversity or similarity of the two test templates because if test templates are more similar to each other, the two score sets would be similar as well. Since the two test templates are from the same person, we expect that, if the



(a) DB1 (b) DB2

(c) DB3 (d) DB4

Figure 2. ROC curves for utilizing likelihood ratio method in FVC2002 DB1 DB2 DB3 and DB4



(a) DB1 (b) DB2

(c) DB3 (d) DB4

Figure 3. ROC curves for utilizing multilayer perceptron in FVC2002 DB1 DB2 DB3 and DB4

second one is more diverse to the first one, the second one will give us more information and the performance will benefit from the fusion.

Our previous work [7] exploited the relationships between matching scores obtained for the same test template. For example, by considering score set $\{s_{1,k}\}$ obtained during matching first test template we can determine the confidence of matching results for this trial. For second test template, e.g. due to varying quality of the input, the confidence of matching results could be different. The algorithms of [7] try to model such variation of the confidence in matching results. But in the current paper, we consider the differences of sets $\{s_{1,k}\}$ and $\{s_{2,k}\}$ and utilize the measures of the diversity of test templates. In [6] we justified the use of a single matching score between two test templates as a quality measure. But this score can also be looked at as a simple diversity measure between test templates. In current paper, we present more complex measures of diversity which show better performance than algorithms of [6].

3. Similarity Measures for Vectors

Matching each test template to all enrolled templates generates one matching score set. Two test templates as in Fig. 1 give us two score sets $s_{i,1}, \ldots, s_{i,k}, \ldots, s_{i,n}$ where $i = 1, 2$. If two test templates are the same, then $s_{1,j}$ equals $s_{2,j}$, $1 \leq j \leq n$. So diversity between test templates could be reflected by using the two matching score sets. Each set of matching scores is actually a vector, so the problem is converted to find the similarity measure between vectors.

There are actually many similarity measures for vectors, such as Pearson's correlation, cosine similarity, Jaccard index. In our experiments we only use Pearson's correlation and cosine similarity.

The Pearson's correlation is:

$$corr(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (1)$$

where X and Y are two sets of matching scores, $\bar{x}$ and $\bar{y}$ are means of the sets. Pearson correlation is close to 1 in the case of a perfect positive linear relationship of two score sets, and it is near 0 in the case of the unrelated sets.

Cosine similarity which is often used in text mining is a measure of similarity between two vectors by measuring the cosine of the angle between them:

$$cos\_sim(X, Y) = \frac{\sum_{i=1}^{n}(x_i)(y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2 \sum_{i=1}^{n}(y_i)^2}} \qquad (2)$$

where parameters have the same meaning as in Eq. 1.

We also performed some preliminary experiments with Euclidean distance between these two vectors serving as a similarity measure, but it did not provide good performance. One explanation for this result is that the score sets could depend in a large degree on the quality of fingerprints, and two fingerprints of different quality could give distant (in an Euclidean distance sense) but well correlated matching score vectors (e.g. $X = kY$). Seemingly, such situations are frequent for the considered application, and thus the correlation and cosine measures have better performance.

4. Combination Rules

The combination methods we use in this paper are likelihood ratio and multilayer perceptron. The likelihood ratio

EER(%)	raw	S_3	cos_sim	cor
db1	0.33 ± 0.06	0.33 ± 0.08	0.26 ± 0.04	0.23 ± 0.04
db2	0.30 ± 0.06	0.30 ± 0.07	0.27 ± 0.07	0.26 ± 0.06
db3	3.30 ± 0.36	2.78 ± 0.37	2.46 ± 0.31	2.15 ± 0.28
db4	2.09 ± 0.28	2.09 ± 0.35	1.91 ± 0.27	1.58 ± 0.23

Table 1. Equal error rate (mean± standard deviation) for FVC2002 datasets using likelihood ratio method

EER(%)	raw	S_3	cos_sim	cor
db1	0.25 ± 0.08	0.23 ± 0.08	0.16 ± 0.05	0.14 ± 0.05
db2	0.22 ± 0.10	0.20 ± 0.09	0.11 ± 0.06	0.08 ± 0.03
db3	3.02 ± 0.57	2.23 ± 0.37	2.20 ± 0.42	1.86 ± 0.36
db4	1.97 ± 0.30	1.65 ± 0.25	1.52 ± 0.27	1.25 ± 0.21

Table 2. Equal error rate (mean± standard deviation) for FVC2002 datasets using multilayer perceptron

is theoretically optimal combination method for verification systems [16]. The formula for raw matching scores s_1 and s_2 is:

$$S = \frac{p_{gen}(s_1, s_2)}{p_{imp}(s_1, s_2)} \tag{3}$$

where $p_{gen}(s_1, s_2)$ is the probability density of genuine scores, and $p_{imp}(s_1, s_2)$ is the probability density of impostor scores. It is our baseline method. Likelihood ratio assigns the combined score a value of ratio between genuine and impostor score densities. In order to use similarity measure, three dimensional densities for genuine and impostor matchings are constructed:

$$S = \frac{p_{gen}(s_1, s_2, sim\_mes)}{p_{imp}(s_1, s_2, sim\_mes)} \tag{4}$$

where sim_mes is either Pearson's correlation or cosine similarity of two score sets generated from matching each test template to all enrolled templates.

In high dimensional space, direct approximation of score densities might not be accurate, so multilayer perceptron which is a feedforward artificial neural network model is also used. In our setting, the perceptron has four layers including two hidden layers, one with eight nodes and the other with nine nodes. The input layer for system using raw scores contains two nodes, in contrast, the input layer for similarity measures has three nodes with the additional node for the similarity measure. The output layer has one node with expected 1 for genuine matching and 0 for impostor matching.

5. Experiment

Each database captured by different sensors in FVC2002 has 110 different persons with 8 samples for the same finger of the person [2]. In genuine matching, two samples of one person are selected to be test templates and another one from the same person to be enrolled template. On the other hand, in impostor matching, two samples from one person are selected to be test templates and one from another person to be enrolled template. For each person, there are 168 genuine matching variations(Each variation contains one enrolled template and two tests exhaustively selected from person's ten templates. One variation has at least one template different from templates of another variation.) and totally it's $168 * 110 = 18480$. We also keep number of impostor matchings to be 18480. In both genuine and impostor matchings, score sets for similarity measures are generated by matching each test template to the first sample of other 109 persons.

The used fingerprint matching system is based on minutiae matching [10]. It is similar to the NIST fingerprint system [3] with a few modification to remove false minutiae on the edge of the fingerprint region. They got the optimal matching of two fingerprints by converting it to be a minimum cost flow problem.

Bootstrap sample testing technique is used [4] in both combination methods - likelihood ratio and multilayer peceptron. Twenty-five persons are selected as training and another twenty-five ones for validation in each bootstrap step. The left sixty persons are used for testing. Totally one hundred bootstrap steps are executed in each combination method. We use Parzen window with Gaussian kernel whose width is estimated by the maximum likelihood method. For multilayer perceptron, we use the Fast Artificial Neural Network Library [1] with default settings.

Our previous work [6] demonstrated the possibility of using matching score between two test templates (called s_3, this notation is also used in this paper) in order to improve performance. According to that work, the score between test templates can represent a quality measure for templates. We can note, that s_3 can also be served a simple measure of diversity. In this paper, we compare the results of utilizing

	cos_sim	cor
db1	0.3271	0.3273
db2	0.3589	0.3665
db3	0.5616	0.5551
db4	0.4444	0.4530

Table 3. Correlations between S_3(matching score between test templates) and cosine similarity or Pearson's correlation

more complex similarity measures to the simple use of score s_3.

Fig. 2 shows the ROC curves for FVC2002 four fingerprint databases using likelihood ratio method. The results of multilayer perceptron are shown in Fig. 3. The curve noted as 'Single_Score' is the ROC for utilization of one score. The notation 'raw' means we use score s_1 and s_2 as in Eq. 3. It is obvious that the performance of using two scores is much better than using one score. The notation 's_3' says besides s_1 and s_2, we also use the score s_3 between the two test samples as in [6]. Utilization of similarity measures such as Pearson's correlation and cosine similarity can further improve the performance than using matching scores s_1 and s_2. In all cases, similarity measures can compete the use of matching score between test templates s_3, in particular for DB1 and DB2, similarity measure is much better.

Table 1 and Table 2 show the mean and standard deviation of equal error rate (EER) from 100 bootstrap steps for FVC2002 using likelihood ratio and multilayer perceptron. Both Pearson's correlation and cosine similarity have smaller equal error rate than using s_3 and only raw matching scores. The relationship of EER is:

$$EER_{cor} < EER_{cos\_sim} < EER_{S_3} <= EER_{raw} \qquad (5)$$

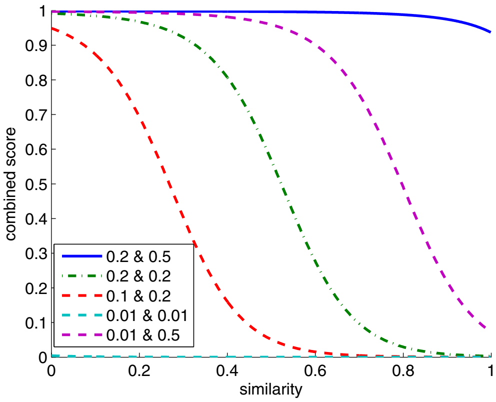


Figure 4. Relationship between combined score and similarity measure for few pairs of matching scores.

The correlations between s_3 and cosine similarity or Pearson's correlation are shown in Table 3. We can notice that all databases exhibit relatively high dependencies between s_3 and similarity measures. This implies that accounting for similarity measures can produce similar effects on combination results as utilizing s_3 in [6].

It would be interesting to look at Fig. 4 for the impact of similarity measures on the final combined score of trained combination algorithm. Specifically, we trained a multilayer perceptron to accept two matching scores and a similarity measure, and output the final combined score. Then, we fix the two input matching scores and calculate the combined score for the range of possible values of similarity measure. In Fig. 4, the horizontal coordinate is the similarity, namely Pearson's correlation, and the vertical coordinate is the combined score. The dependence graphs are given for five pairs of matching scores $\{s_1, s_2\}$, that is, {0.2,0.5},{0.2,0.2},{0.1,0.2},{0.01,0.01} and {0.01,0.5}. Since the actual mean of impostor scores is near 0.02 and the actual mean of genuine scores is near 0.62, the selected score pairs reflect typical genuine pairs({0.2,0.5},{0.2,0.2},{0.1,0.2}), impostor pairs{0.01,0.01}, as well as, an indeterminate case ({0.01,0.5}).

We can make some observations regarding the experiment:

1. When similarity increases, the final combined score for each pair decreases.

2. When both scores are small enough, for instance 0.01 and 0.01, the combined score is near 0 no matter what the similarity measure is.

3. When both scores are relatively large, in the case of 0.2 and 0.5, the combined score is still big even even for large similarity value.

4. The combined score for {0.1,0.2} can be bigger than the combined score for {0.2,0.2} at some values of similarities. For instance, the combined score for {0.1,0.2} with similarity at 0.2 is almost 0.6 But the combined score for {0.2,0.2} with similarity at 0.6. is only 0.28

As we can see, similarity measure indeed provides valuable information during combination. Smaller similarity or larger diversity indicates that the information contained in each template can compensate each other to achieve higher matching.

6. Conclusions

In this paper, we have used similarity measures in multisample biometric systems. Both Pearson's correlation and cosine similarity are used in our cases. Experiments show that it provides us a better performance than using raw matching scores.

The observed improvement appears to be the result of diversity between multiple samples. Pre-determined rules such as averaging doesn't account this information and is clearly inadequate. In this paper we proposed to utilize similarity measures between input samples based on vectors of matching scores and experiments confirm the virtue of using them. The proposed similarity measures also have a better performance than simply using a matching score between input samples.

We don't have specific information whether fingerprint templates for each person are taken in the same session in this FVC2002 dataset. If they are taken at different time, two templates for the same person might bear a higher diversity and the results might need to be adjusted.

Our work can be extended to more general cases, such as more than two test template or enrolled templates. In particular, it can be applied to matching a single face image to video of one person. It would be even beneficial to combine scores for each frame of the video in the case some frames bear errors.

References

[1] Fast artificial neural network library. http://leenissen.dk/fann/wp/. 4

[2] Fingerprint verification competition. 2006. http://bias.csr.unibo.it/fvc2002/. 4

[3] Nist fingerprint software. Internet page, National Institute of Standards and Technology(NIST), March 2006. 2006. http://fingerprint.nist.gov/NFIS/. 4

[4] R. M. Bolle, N. K. Ratha, and S. Pankanti. Error analysis of pattern recognition systems–the subsets bootstrap. *Computer Vision and Image Understanding*, 93(1):1–33, 2004. doi: DOI: 10.1016/j.cviu.2003.08.002. 4

[5] R. Chellappa, V. Kruger, and Z. Shaohua. Probabilistic recognition of human faces from video. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, pages I–41–I–44 vol.1, 2002. 2

[6] X. Cheng, S. Tulyakov, and V. Govindaraju. Multiple-sample fusion of matching scores in biometric systems. *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 120–125. 3, 4, 5

[7] X. Cheng, S. Tulyakov, and V. Govindaraju. Combination of multiple samples utilizing identification model in biometric systems. In *International Joint Conference on Biometrics (IJCB2011)*, Washington, USA, 2011. 3

[8] A. Jain and A. Ross. Fingerprint mosaicking. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 4, pages IV–4064 –IV–4067, May 2002. 2

[9] A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Trans. on Circuits and Systems for Video Technology*, 14:4–20, 2004. 1

[10] T.-Y. Jea and V. Govindaraju. A minutia-based partial fingerprint recognition system. *Pattern Recognition*, 38(10):1672–1684, 2005. 4

[11] T. Kanade, A. Jain, N. Ratha, C. Ryu, Y. Han, and H. Kim. Super-template generation using successive bayesian estimation for fingerprint enrollment. volume 3546 of *Lecture Notes in Computer Science*, pages 261–277. Springer Berlin / Heidelberg, 2005. 2

[12] E. Kelkboom, J. Breebaart, R. Veldhuis, X. Zhou, and C. Busch. Multi-sample fusion with template protection. In *BIOSIG 2009: Proceedings of the Special Interest Group on Biometrics and Electronic Signatures*, Darmstadt, Germany, 2009. 2

[13] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(3):226–239, 1998. 1

[14] L. Kuang-Chih, J. Ho, Y. Ming-Hsuan, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, 1:I–313–I–320 vol.1, 2003. 2

[15] L. I. Kuncheva, C. J. Whitaker, C. A. Shipp, and R. P. W. Duin. Is independence good for combining classifiers? In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 2, pages 168–171 vol.2, 2000. 1

[16] S. Prabhakar and A. K. Jain. Decision-level fusion in fingerprint verification. *Pattern Recognition*, 35(4):861–874, 2002. doi: DOI: 10.1016/S0031-3203(01)00103-0. 4

[17] G. Rigoll and A. Kosmala. A systematic comparison between on-line and off-line methods for signature verification with hidden markov models. In *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, volume 2, pages 1755–1757 vol.2, 1998. 2

[18] A. A. Ross, K. Nandakumar, and A. K. Jain. *Handbook of Multibiometrics*. International Series on Biometrics. Springer-Verlag New York, Inc., 2006. 1

[19] U. Uludag, A. Ross, and A. A. Jain. Biometric template selection and update: a case study in fingerprints. *Pattern Recognition*, 37(7):1533–1542, 2004. doi: 10.1016/j.patcog.2003.11.012. 2

[20] Y. Zhang and A. M. Martnez. A weighted probabilistic approach to face recognition from multiple images and video sequences. *Image and Vision Computing*, 24(6):626–638, 2006. doi: DOI: 10.1016/j.imavis.2005.08.004. 2