Kuwait J. Sci. **41** (3) pp. 103-127, 2014

Multimodal biometric authentication using adaptive decision boundaries

FUAD M. ALKOOT

HITN, PAAET , P.O.Box 313, AlShuhada-S.Surra, 48554, Kuwait
f_alkoot@yahoo.com, fm.alkoot@paaet.edu.kw

ABSTRACT

Recent advances in multimodal biometric authentication techniques have improved their reliability level. The use of combination methods was one of the main techniques to improve the performance. We aim to find the fusion strategy that yields the best classification rate in experiments on the MOBIO-Banca biometric data set. Scores from classifier outputs are ranked and fused using several fusion strategies. Our proposed method of combined use of adaptive boundaries and quality measure based ranking of scores yields significant improvement over existing fixed boundary methods. Sum fusion yields very large improvement and succeeds, where MProduct does not. We explain the degradation in the serial MOBIO combiner analytically and through synthetic experiments. We show that the performance of weak classifier combiners follow a hyperbolic curve as weaker classifiers are added.

Keywords: Banca data; biometric; classifier fusion; MOBIO; quality measure.

INTRODUCTION

The advances in portable networked electronics devices and the increasing need to authenticate users of these devices (Savvides *et al.,* 2005) have motivated research in many aspects of the biometric authentication field. Currently these devices require a PIN code or a password to verify the authenticity of the user. However, these methods are susceptible to fraud as they can be stolen, forgotten or compromised. The consumers are interested in robust and easy services that can easily identify them. With the advances in audio and video capabilities of these devices, we can clearly see that the most reliable identity verification methods would be based on face and voice authentication.

EU has initiated a project to investigate mobile biometrics under the name MOBIO (Bailly-Bailliere *et al.,* 2003; MOBIO). MOBIO concept is to develop new mobile services secured by biometric authentication means. The goal of this project is to study, develop and evaluate biometric authentication technologies in the context of portable and networked devices. The objective is to develop face authentication systems robust-to-illumination and robust-to-noise.

Biometric authentication has not gained a wide acceptance level due to its degraded reliability. However, recent advances in multimodal authentication techniques have improved the reliability level. The use of combination methods has improved the performance of biometric authentication techniques (Hong & Jain, 1998). However, there are many open questions and further research is required to reach an acceptable performance level (Ross & Poh, 2009)

Since face recognition systems based on a single shot of the client (Ross & Poh, 2009; Rouvier *et al.,* 2009) can be compromised, as it is possible to use a picture of the genuine identity, we aim to investigate authentication using a video sequence. We use quality measures associated with a sample to improve the performance of score based ranked fusion systems, where we experiment with three types of rank based combiners. The contributions can be summarized as follows:

1 - An experimental comparison, on MOBIO data of soft fusion strategies and three types of combiner systems that use quality measures.

2 - Proposal of an adaptive boundary method to improve the classification rate of combiner systems used on the MOBIO biometric authentication data.

3 - An analytical and experimental investigation of the performance of the weak classifier combiner and displaying its hyperbolic performance curve.

In the next section we present an overview of previous research on biometric authentication and verification using score fusion and quality measures. Experimental methodology is presented in the third section followed by a section on results of the various fusion strategies and systems. The paper is ended by a section on the weak classifier fusion concept followed by a conclusion.

RELATED WORK

Quality measures have been used by many to improve the performance of biometric authentication systems using different fusion architectures. Researchers have developed and used different types of quality measures for single or multi-modalities. Poh & Kittler (2012) survey recent work on quality based fusion followed by a presentation of several theoretical and experimental findings that show the best approach for the problem. They categorize quality based fusion algorithms in two types: feature based and cluster based. Their work yields a proposal of a methodogical approach to the problem of quality based fusion. A recent work that uses different quality measures than ours is by Nandakumar *et al.* (2006). They show the advantage of weighting the decisions of classifiers based on quality measures by proposing a likelihood ratio based fusion that takes into account the quality of the biometric samples. Experiments were performed on iris and fingerprint modalities. They show that quality based product fusion yields best results in comparison to non quality or quality based sum fusion. Fierrez-Aguilar *et al.* (2005) used quality measures in a rank based fusion with a SVM combiner, where they employ a SVM classifier to classify score vectors. Samples with good quality are assigned higher cost of misclassification during training. Score level fusion function is adapted every time biometric data is sensed, depending on estimated quality. They experiment with signature and fingerprint modalities. Abaza & Ross (2009) use finger print quality measures in a rank level based fusion of biometrics. Several simple but powerful modifications were suggested to enhance the performance of rank-level fusion schemes in the presence of weak classifiers or low quality input images. Instead of quality measures authors in Poh *et al.* (2012) use higher order moments of the video scores to improve the standard fixed fusion strategies by as much as 50 percent. Recently Tresadern *et al.* (2013) reported a multimodal system that experiments on a new biometric data that is most similar to the mobile environment and includes jitter.

Quality measures adopted in this paper were also used by Poh & Kittler (2008); Fatukasi *et al.* (2007); Kittler *et al.* (2007). Poh & Kittler (2008) successfully combine device information and quality measures in a face and fingerprint multi-biometric fusion scenario, using Biosecure DS2 dataset. In a multimodal biometric score level fusion, using XM2VTS (Matas *et al.,* 2000) data base, Kittler *et al.* (2007) use quality measures 9 and 16 to control the influence of each classifier on the final fused score. Results prove quality based fusion outperforms quality free fusion. Also Fatukasi *et al.* (2007) show that quality based fusion outperforms quality free fusion. They propose a two stage system. In the first stage they group scores according to the quality of the samples, and combine scores of each group using sum fusion. Resulting group scores are combined in a second stage using product fusion. This is similar to the fixed boundary two level fusion architecture experimented in this paper. However, we experiment with several soft fusion strategies besides product and report a comparison of their performance.

Our work aims at improving on previous work involving one and two level ranked fusion systems that fuse the classifier output scores obtained from comparing frames of a video sequence to a claimed identity, in a MOBIO biometric application. Scores were produced by a Local Binary Pattern, LBP (Chan, 2008; Fierrez-Aguilar *et al.,* 2005; Fukunaga, 1990; Heusch *et al.,* 2006 and Pietikainen *et al.,* 2011), and a Linear Discriminant Analysis, LDA (Chan, 2008; Fukunaga, 1990), classifier. Quality measures produced by

Omniperception Ltd. are used to improve the performance of one level and two level rank based fusion systems. At the decision level we propose using training set based adaptive boundaries to separate classes. We show that MOBIO combiner systems improve using our proposed adaptive boundaries in contrast to the commonly used fixed boundary MOBIO combiners. We also experiment with various fusion strategies to find the best one. Results show that adaptive boundary using sum fusion outperforms fixed boundary at all rank levels and with higher percentages than modified product fusion. We find that MProduct outperforms Sum only for the controlled genuine LBP classifier, where it reaches 100%.

Fused scores of each of the first level classifiers of our two level combiner systems are introduced to the combiner in sequence to show the combiner performance due to each classifier. Results show combiner improves before it degrades, as additional classifiers are combined. We propose a formula that, based on critical parameters of the classifier, can be used to find, if weak classifiers will improve the combiner performance. This formula explains the hyperbolic performance seen in the MOBIO combiner when mixed classifiers are combined.

EXPERIMENTAL METHODOLOGY

We experiment with the BANCA-MOBIO database, which is a multi-modal database intended for training and testing multi-modal verification systems. A full description of the database is available in Bailly-Bailliere *et al.* (2003). The BANCA database was captured in four European languages in two modalities (face and voice). For recording, both high and low quality microphones and cameras were used. The captures from the four languages yield a total of 208 people. The 52 subjects per language, half men and half women, were recorded in three different scenarios; controlled, degraded and adverse, over 12 different sessions spanning three months. Each gender of each language was divided into 2 groups of 13 subjects, g1 and g2. Each subject recorded 12 sessions. Each of these sessions containing 2 recordings: 1 true client access and 1 imposter. The 12 sessions were separated into 3 scenarios resembling different conditions: controlled for sessions 1 to 4, degraded for sessions 5 to 8, and adverse for sessions 9 to 12. All the subjects in group g recorded one imposter attempt against each other. Each identity was attacked 4 times in each of the 3 conditions or scenarios totaling 12 attacks.

Associated with the database is the BANCA protocol. The protocol defines which sets of data to use for training, evaluation and testing. Performing experiments according to the protocol allows institutions to easily compare their results to others. Here, we use the configuration of the BANCA protocol involving only one language. When a separate development and evaluation sets are needed, g1 and g2 are used alternatively as development set and as evaluation set. In multi-modality experiments a third set is needed for tuning the fusion parameters. In this case a group from another language is used for the development set while g1 and g2 are used for evaluation and tuning, alternatively. Seven distinct experimental configurations have been specified, which identify which material can be used for training and which for testing. In all seven, the true client record from sessions 1, 5, and 9 are used for training. Our experiments are based on the seventh configuration, called the Grand test (G). It uses the three records mentioned above, (i.e. 1, 5 and 9), of each genuine client, for training, and the rest of the records for testing. The imposter records of the three training records were also used for testing. Based on this, the training set includes 26 identities with three genuine records for each, totaling 78 records. The remaining nine sessions for each of the 26 identities include genuine and imposter claims totaling 468 records for the test set. The three imposter records from sessions 1, 5 and 9 for each of the 26 identities were also added to the test set. The total test set records reaches 546 records for 26 identities. Therefore, based on the set protocols, no cross validation or randomized separation of data into training and test set is performed. The classification rate is therefore found by summing the number of correct classification of test samples divided by the number of test samples.

Each test sample is a video clip and a claimed identity that must be verified. Frames of the video sequence are compared to a claimed identity using LBP and LDA classifiers, resulting in frame output scores, obtained from Chan (2008). These scores are placed in files including the true identity number. We will fuse the classifier output scores using various combination systems and fusion strategies. Based on the scores output by the combiner, a decision is made to accept or reject the claim. The systems can make two types of errors; False Acceptance (FA) is to wrongly accept the imposter, while False Rejection (FR) would be to reject a true identity as an imposter. Each test sample is for a different identity with a genuine or an imposter claim. Therefore, we get a False Acceptance Rate (FAR) as the ratio between FA and number of imposter accesses. While (FRR) is the ratio between the number of FR and number of client accesses. To visualize the performance of the system DET curves are used, as in figures 6 to 9.

DET curves plot the FRR vs FAR. The point on the curve corresponding to FRR = FAR is called equal error rate, EER. The closeness of the DET curve to the origin is measured using the ERR. Another method of displaying the performance rate is the classification rate curve of the various fusion methods. Based on the number of test samples, we have a repeated number of system tests which yield an average classification rate, as shown in figures 4 and 5.

Along with each score file we have a file containing 16 quality measures (Poh & Kittler, 2008; Fatukasi *et al.,* 2007). These are as follows:

1 - Left eye coordinate (x,y) of the original image (extracted from video)

2 - Right eye coordinate (x,y) of the original image (extracted from video)

3 - Reliability of the face detector - this is the output of a classifier that has been trained to give an˜overall measure of quality given the quality measures 4-16 below

4 - Brightness

5 - Contrast

6 - Focus - this quantifies the sharpness of an image

7 - Bit per pixel - it measures the colour resolution in terms of bits

8 - Spatial resolution - the number of pixels between eyes

9 - Illumination

10- Uniform Background -- measuring the variance of the background intensity

11- Background Brightness -- the average intensity of the background

12- Reflection - or, specular reflection

13- Glasses -- face wearing glasses

14- Rotation in Plane

15- Rotation in Depth

16- Frontalness - it measures how much a face image deviates from a typical mug-shot face image

Our preliminary experiments showed that quality measures 3 and 16 yield best results. Therefore, only these quality measures are used in our experiments.

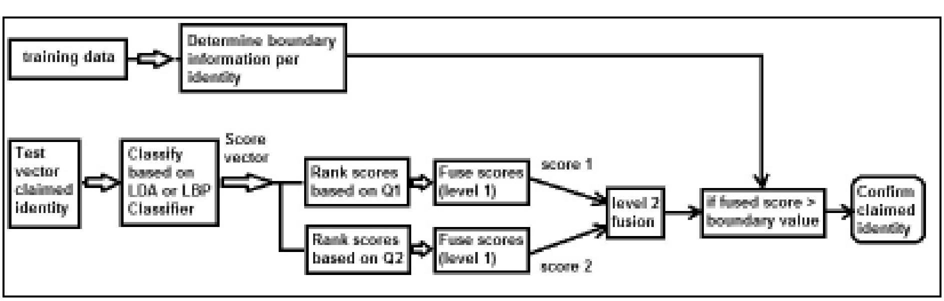


Fig. 1. Architecture of proposed system number 6; two level fusion using boundary information

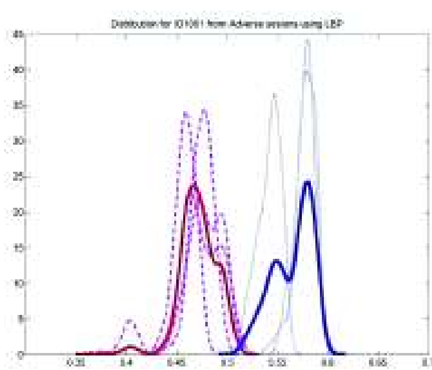
We experiment with six different fusion systems which use different ranking and fusion levels given a fixed or adaptive decision boundary, as described below. In all fusion tasks we experiment with five soft fusion strategies; Sum, Product, Modified Product (Alkoot & Kittler, 2002), Maximum, and Minimum (Alkoot & Kittler, 1999).

1 - Ranked fusion: The scores of a video are ranked and only the top k scores are fused using soft fusion strategies. We experiment with different rank levels such as 3, 9, and 20 in addition to fusion of all scores.

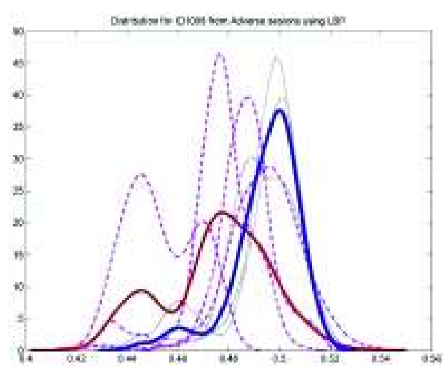
2 - One level fusion: the scores are ranked based on each of the two quality measures 3 and 16, in contrast to the ranked fusion which ranks based on the classifier score outputs. The two arrays of ranked scores are then fused in one array.

3 - Two level fusion, this is the fusion of fused scores. As in the one level fusion, we rank the scores based on each of the two quality measures. Next ranked scores of each array are fused separately to obtain two scores. At a second stage these two scores are fused again using one of the five fusion strategy mentioned above.

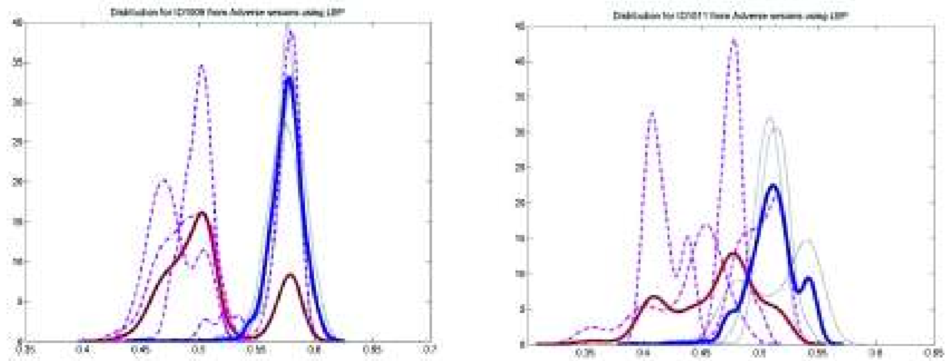
4 - Boundary fusion: The one level and two level fusion methods are repeated. However the decision boundary is found using the training set. This is found by initially finding the average of all genuine curves and the average of all imposter curves, where each curve is for a different video clip of the claimed identity. Next the genuine average curve is found and the closest imposter score is found. If the two curves don't intersect, we have perfect separation and boundary is exactly between the two curves. Otherwise, the boundary exists at the point of intersection of the two average curves.



(a)



(b)



(c) (d)

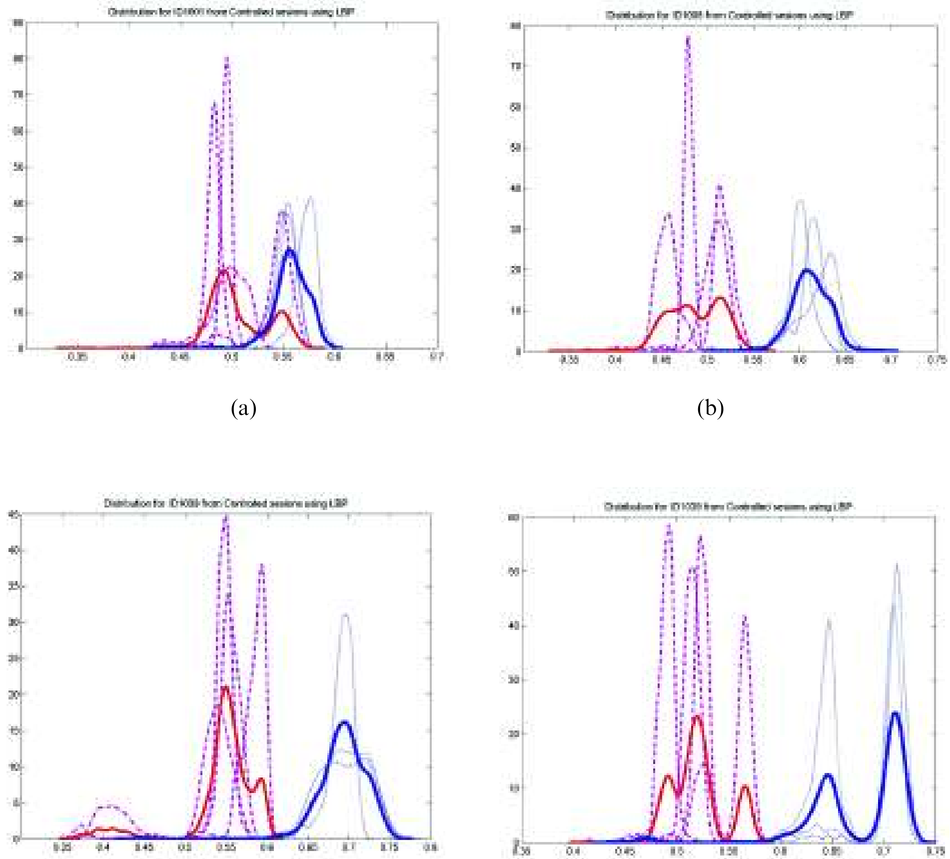
Fig. 2. Distribution of the two classes for the Adverse data using LBP classifier for identities (a)1001, (b)1008, (c)1009 and (d)1011

RESULTS

Data set distributions

Figures 2 and 3 show the distribution of the two classes based on the scores from the LBP classifier, at the adverse and controlled session types. Each figure is for a separate identity, where it shows the distribution of the genuine and imposter scores from the video frames. Therefore, each curve is for a different video clip, given a certain claim. The higher scores in the x axes are for the genuine claim while the lower are for the imposter. The average of all the curves for the genuine and the average of all the curves for the imposter are shown in solid blue and red, respectively. In gray scale printing the blue is the thicker darker line.

These figures show the difficulty in making the correct identity decision. For example identities 1008, 1009, 1011, 1030, 1031, 1032 and 1034 of the adverse session and identities 1001 and 1029 of the controlled session are cases that are difficult due to imposter scores being in the genuine region. A comparison of figures for cases 1008 and 1009 of the adverse session, Figures 2(b) and 1(c), to the same of the controlled session, Figures 3(b) and 3(c), shows that a controlled session can improve the separation between the two classes.



(a) (b) (c) (d)

Fig. 3. Distribution of the two classes for the controlled session using LBP classifier for identities (a)1001, (b)1008, (c)1009 and (d)1039.

Ranked Fusion using Quality Measures

Figure 4 presents results using the LBP and LDA classifiers, for the three different combiners at different rank levels and two boundary types, when sum fusion is used. It shows classification rates for the genuine, imposter and the average of the two, for adverse and controlled sessions. Looking at the controlled session results of the LBP classifier, we find that ranked fusion outperforms one and two level fusion on average, while if adaptive boundary is used, one and two level fusions outperform all. Additionally, when adaptive boundary is used, genuine and imposter rates are both high and very close, while for fixed boundary the imposter results are very low. This is true for all rank sizes. Adaptive boundary leads to imposter rates improving above genuine rates for ranked fusion. Increasing the rank level leads to degradation of the ranked fusion, while one and two level fusion are robust and yield relatively constant performance at various rank levels. This holds also when LDA classifier is used.

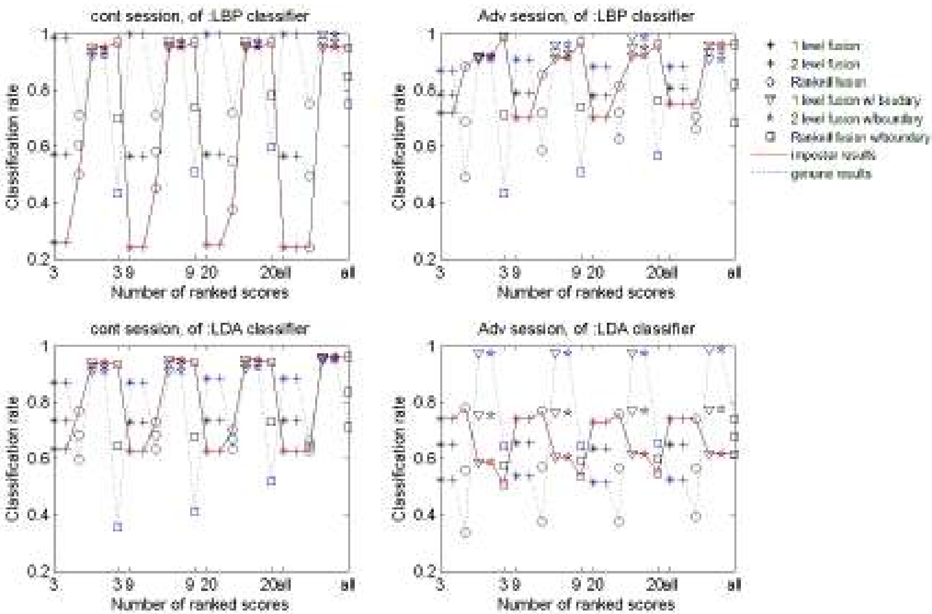


Fig. 4. Combiner method results using Sum fusion at different rank levels.

For the adverse session, imposter rate improves in contrast to the controlled session. The different combiner methods yield similar performance to the controlled session. When fixed boundary is used, imposter rates using ranked fusion yield better rate than genuine. As in the controlled session, ranked fusion outperforms one and two level fusion on imposter rates, but this drops to equal as the rank size is increased. One and two level fusion outperform ranked on genuine rate and on average. Using an adaptive boundary, again one and two level fusion yield equally good results on imposter and genuine. On average they outperform ranked fusion. They improve as the rank size is increased to 20 ranks and drop slightly when all scores are fused. The same holds for the LDA classifier except that the one and two level fusion using an adaptive boundary yield higher rate on genuine than on imposter. On average they yield best results at all rank levels. Figure 5 is using MProduct (Alkoot & Kittler, 2002) fusion. Here we find a difference in performance between the two level and one level fusion, where two level fusion leads.

For the genuine and average rates, Table 2 shows that one level quality measure ranked fusion, using adaptive boundary with MProduct, improves over fixed boundary at small rank sizes, but fails to outperform the fixed boundary for the adverse imposter LDA classifier. For the imposter rate two level fusion using adaptive boundary with quality measure based ranked scores yields best performance. However, Table 1 shows that using Sum, adaptive boundary outperforms fixed boundary at all rank levels and with higher percentages than MProduct. MProduct outperforms Sum only for the controlled genuine LBP classifier, where it reaches 100%.

One case where fixed boundary outperforms adaptive boundary is when the two level quality based fusion is used on adverse imposter using LDA classifier.

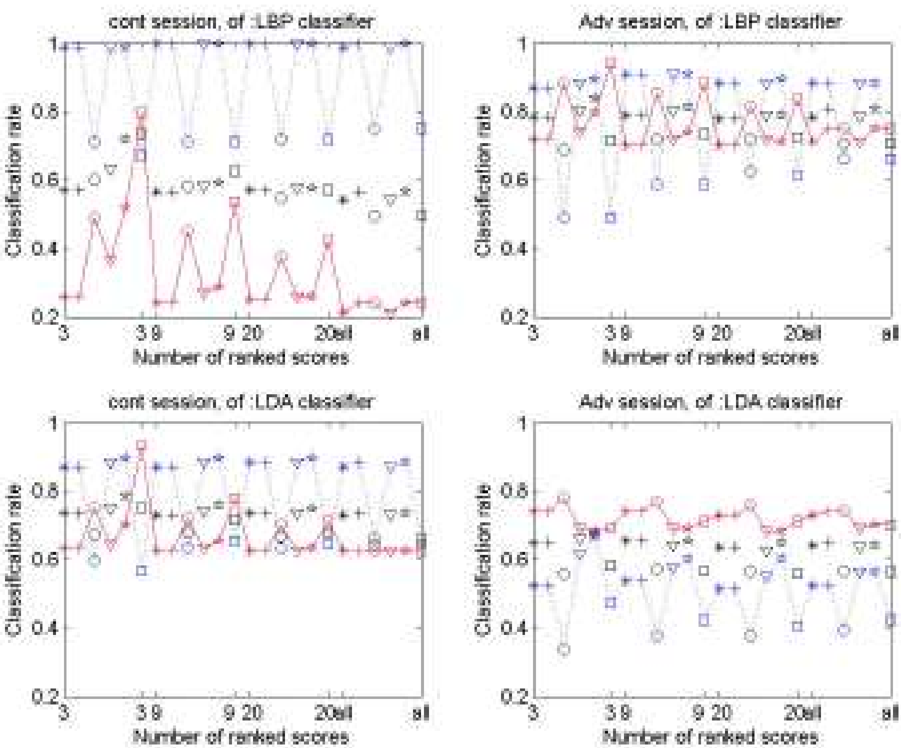


Fig. 5. Combiner method results using MProduct fusion at different rank levels

Table 1. Sum fusion strategy classification rates for the six combiner methods at both sessions and classifier types

	Classifier	Session	Rank size	Combiner Method 1	2	3	4	5	6
Average rate (imposter & genuine)	LBP	Controlled	3	57.14	57.14	60.57	**93.95**	**93.95**	70.19
			9	56.59	56.59	58.17	**96.15**	**96.15**	74.03
			20	57.14	57.14	54.80	**96.15**	**96.15**	78.36
			all	56.59	56.59	49.51	**97.25**	**97.25**	85.09
		Adverse	3	78.57	78.57	68.75	**91.75**	**91.75**	71.15
			9	79.12	79.12	72.11	**93.40**	**93.40**	74.03
			20	78.02	78.02	72.11	**95.05**	**95.05**	76.44
			all	80.76	80.76	70.67	**93.95**	**93.95**	82.21
	LDA	Controlled	3	73.62	73.62	68.26	**92.85**	**92.85**	64.42
			9	73.076	73.07	68.26	**93.40**	**93.40**	67.78
			20	73.62	73.62	66.82	**93.95**	**93.95**	73.07
			all	73.62	73.62	64.42	**95.60**	**95.60**	83.65
		Adverse	3	64.83	64.83	55.76	**75.27**	**75.27**	57.69
			9	65.38	65.38	57.21	**76.37**	**76.37**	59.13
			20	63.73	63.73	56.73	**76.92**	**76.92**	60.09
			all	64.83	64.83	56.73	**77.47**	**77.47**	67.78
Genuine rate	LBP	Controlled	3	98.71	98.71	71.15	92.30	92.30	43.26
			9	**100.00**	**100.00**	71.15	97.43	97.43	50.96
			20	**100.00**	**100.00**	72.11	97.43	97.43	59.61
			all	**100.00**	**100.00**	75.00	**100.00**	**100.00**	75.00
		Adverse	3	87.17	87.17	49.03	**91.02**	**91.02**	43.26
			9	91.02	91.02	58.65	**96.15**	**96.15**	50.96
			20	88.46	88.46	62.50	**98.71**	**98.71**	56.73
			all	88.46	88.46	66.34	**91.02**	**91.02**	68.26
	LDA	Controlled	3	87.17	87.17	59.61	**91.02**	**91.02**	35.57
			9	87.17	87.17	63.46	**91.02**	**91.02**	41.34
			20	88.46	88.46	63.46	**92.30**	**92.30**	51.92
			all	88.46	88.46	66.34	**94.87**	**94.87**	71.15
		Adverse	3	52.56	52.56	33.65	**97.43**	**97.43**	64.42
			9	53.84	53.84	37.50	**97.43**	**97.43**	64.42
			20	51.28	51.28	37.50	**97.43**	**97.43**	65.38
			all	52.56	52.56	39.42	**98.71**	**98.71**	74.03

Cont. Table 1. Sum fusion strategy classification rates for the six combiner methods at both sessions and classifier types

	Classifier	Session	Rank size	Combiner Method 1	2	3	4	5	6
Imposter rate	LBP	Controlled	3	**25.96**	25.96	50.00	95.19	95.19	**97.11**
			9	24.03	24.03	45.19	95.19	95.19	**97.11**
			20	25.00	25.00	37.50	95.19	95.19	**97.11**
			all	24.03	24.03	24.03	**95.19**	**95.19**	**95.19**
		Adverse	3	72.11	72.11	88.46	92.30	92.30	**99.03**
			9	70.19	70.19	85.57	91.34	91.34	**97.11**
			20	70.19	70.19	81.73	92.30	92.30	**96.15**
			all	75.00	75.00	75.00	**96.15**	**96.15**	**96.15**
	LDA	Controlled	3	63.46	63.46	76.92	**94.23**	**94.23**	93.26
			9	62.50	62.50	73.07	**95.19**	**95.19**	94.23
			20	62.50	62.50	70.19	**95.19**	**95.19**	94.23
			all	62.50	62.50	62.50	**96.15**	**96.15**	**96.15**
		Adverse	3	74.03	74.03	**77.88**	58.65	58.65	50.96
			9	74.03	74.03	**76.92**	60.57	60.57	53.84
			20	73.07	73.07	**75.96**	61.53	61.53	54.80
			all	74.03	74.03	74.03	61.53	61.53	61.53

Table 2. MProduct fusion strategy classification rates for the six combiner methods at both sessions and classifier types

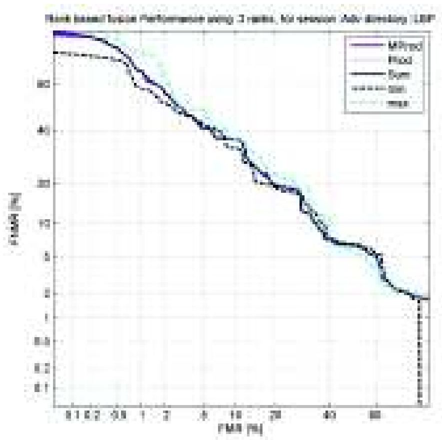
	Classifier	Session	Rank size	Combiner Method 1	2	3	4	5	6
Average rate (imposter & genuine)	LBP	Controlled	3	57.14	57.14	60.10	63.19	**71.98**	**73.56**
			9	56.59	56.59	58.17	58.24	59.34	62.50
			20	57.14	57.14	54.81	57.69	57.69	57.21
			all	54.39	56.59	49.52	54.40	56.59	49.52
		Adverse	3	78.57	78.57	68.75	80.22	**84.06**	71.63
			9	79.12	79.12	72.11	80.22	**81.31**	73.55
			20	78.02	78.02	72.11	79.12	79.12	72.59
			all	78.57	80.76	70.67	78.57	80.77	70.67
	LDA	Controlled	3	73.62	73.62	67.31	74.73	**78.57**	**75.00**
			9	73.07	73.07	67.79	74.18	75.82	71.63
			20	73.62	73.62	66.83	73.63	74.72	67.79
			all	73.07	73.62	64.42	73.08	73.62	64.42
		Adverse	3	64.83	64.83	55.76	65.93	**67.58**	58.17
			9	65.38	65.38	57.21	64.29	65.38	56.73
			20	63.73	63.73	56.73	62.64	64.83	55.77
			all	64.28	64.83	56.73	63.74	64.28	56.25

Cont. Table 2. MProduct fusion strategy classification rates for the six combiner methods at both sessions and classifier types

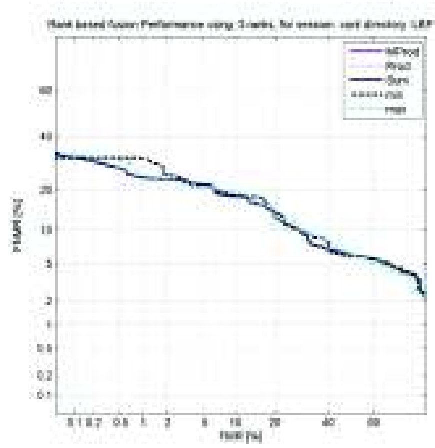
	Classifier	Session	Rank size	Combiner Method 1	2	3	4	5	6
Genuine rate	LBP	Controlled	3	98.72	98.72	71.15	98.72	98.72	67.31
			9	**100.00**	**100.00**	71.15	**100.00**	**100.00**	71.15
			20	**100.00**	**100.00**	72.12	**100.00**	**100.00**	72.12
			all	98.72	**100.00**	75.00	98.72	**100.00**	75.00
		Adverse	3	87.18	87.18	49.04	88.46	89.74	49.04
			9	**91.03}**	**91.03}**	58.65	**91.03}**	**91.03}**	58.65
			20	88.46	88.46	62.50	88.46	89.74	61.54
			all	88.46	88.46	66.35	88.46	88.46	66.35
	LDA	Controlled	3	87.18	87.18	59.62	88.46	**89.74**	56.73
			9	87.18	87.18	63.46	88.46	**89.74}**	65.38
			20	88.46	88.46	63.46	88.46	**89.74**	64.42
			all	87.18	88.46	66.35	87.18	88.46	66.35
		Adverse	3	52.56	52.56	33.65	61.54	**66.67**	47.12
			9	53.85	53.85	37.50	57.69	**60.26**	42.31
			20	51.28	51.28	37.50	55.13	**60.26**	40.38
			all	52.56	52.56	39.42	56.41	56.41	42.31
Imposter rate	LBP	Controlled	3	25.96	25.96	49.03	36.53	**51.92**	**79.81**
			9	24.03	24.03	45.19	26.92	28.84	53.85
			20	25.00	25.00	37.50	25.96	25.96	42.31
			all	21.15	24.03	24.03	21.15	24.03	24.04
		Adverse	3	72.12	72.12	88.46	74.03	79.80	**94.23**
			9	70.19	70.19	85.58	72.11	74.03	**88.46**
			20	70.19	70.19	81.73	72.12	71.15	**83.65**
			all	71.15	75.00	75.00	71.15	75.00	75.00
	LDA	Controlled	3	63.46	63.46	75.00	64.42	70.19	**93.26**
			9	62.50	62.50	72.12	63.46	65.38	**77.88**
			20	62.50	62.50	70.19	62.50	63.46	**71.15**
			all	62.50	62.50	62.50	62.50	62.50	62.50
		Adverse	3	74.04	74.04	**77.88**	69.23	68.26	69.23
			9	74.04	74.04	**76.92**	69.23	69.23	71.15
			20	73.07	73.07	**75.96**	68.26	68.26	71.15
			all	73.07	74.04	74.04	69.23	70.19	70.19

Comparing Fusion methods using DET curves

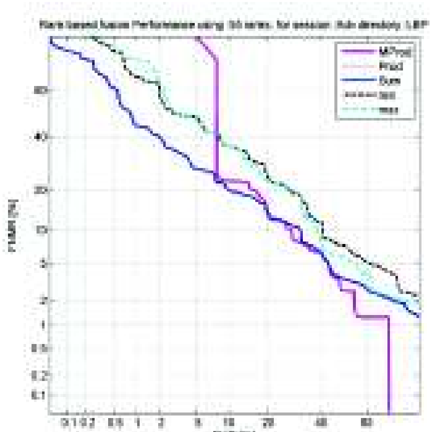
Figures 6 to 9 present DET curves of result for the LBP classifier ranked score fusion (Figure 6), two level fusion (Figure 7), LDA classifier two level fusion Figure .8 and one level fusion Figure 9. These figures are for results of rank size 3 and none ranked all score fusion. DET yield another understanding of the performance, as it considers both genuine and imposter performance relative to each other. Much information can be gathered from these DET figures, but we highlight the most notable as follows; LBP results for adverse session at 20 ranks, using rank based fusion of scores, indicates Minimum is best, followed by Sum and MProduct, while Maximum underperforms all. This is also true for the controlled session. Using two level fusion MProduct and Sum are close with Sum leading on the controlled session. For one level fusion it is more obvious, especially for the controlled session, that Sum and MProduct outperform other methods; they are followed by Minimum. For the adverse session the same occurs, however, MProduct degrades below all at small FMR below 2%.



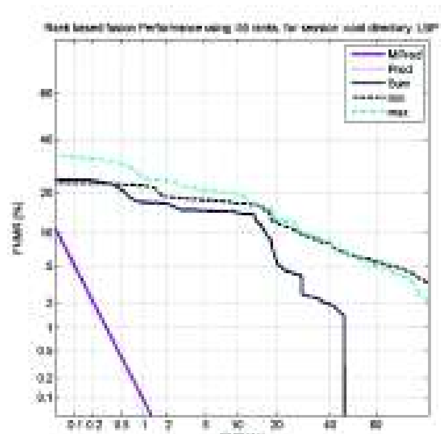
(b)



(d)



(c)



(d)

Fig. 6. Rank based results for the LBP classifier using different soft fusion methods. Modified product and product curves in Figs (a) and (b) are under the sum curve due to exact performance.

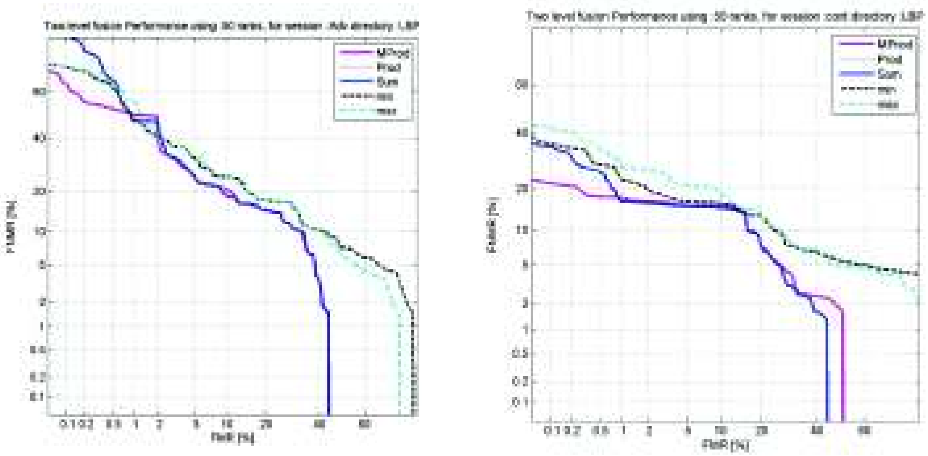


Fig. 7. Two level fusion DET curves for the LBP classifier using different soft fusion methods.

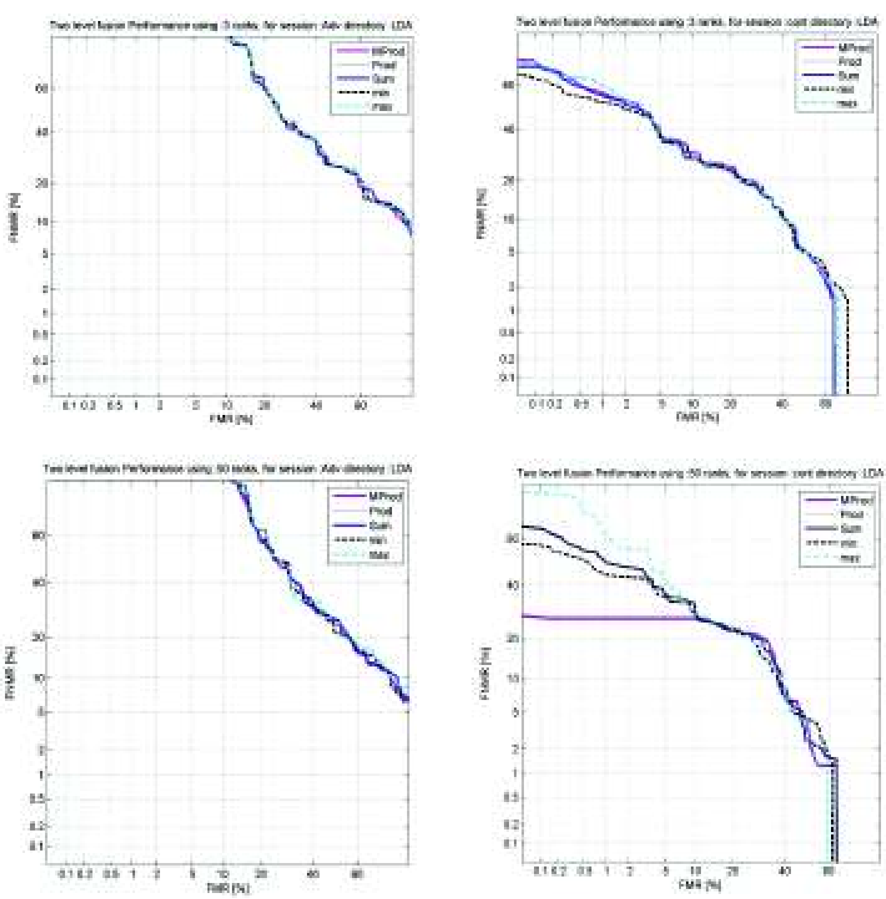


Fig. 8. Two level fusion DET curves for the LDA classifier at 3 ranks and full rank fusion.

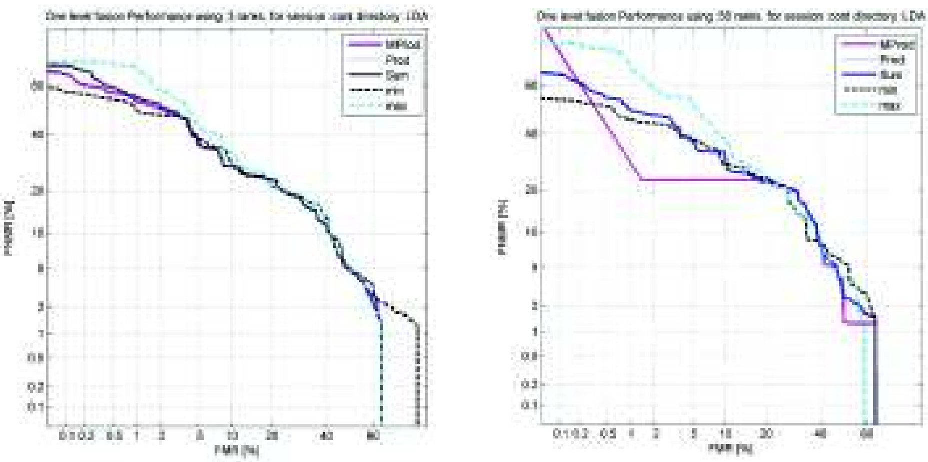


Fig. 9. One level fusion DET curve results for the controlled session LDA classifier, at 3 ranks and full score fusion.

WEAK CLASSIFIER FUSION

All classifiers suffer from a bayes error due to uncertainties in the problem space. Therefore, classifiers designed on different problem spaces have different bayes errors. Classifiers with large bayes error are considered weak. Assuming we have several classifiers with different underlying bayes error that are arranged in a sequence of increasing bayes error. The amount by which each classifier bayes error increases in comparison to the previous ones is called weakness factor, m. The classifiers are added to the combiner in sequence from strongest to weakest. This scenario is used to explain the behavior of the MOBIO combiner.

The MOBIO combiner consists of six classifiers arranged from strongest to weakest based on their performance on an evaluation set g2. For the three different fusion systems, described in section 3, we have scores using two classifier types, LDA and LBP. Therefore, we have six different decision scores that can be considered as classifier decisions. At a second stage some or all of these decisions can be combined together to yield one final combined score. We experiment with gradually adding the six classifiers to the combiner and find its error rate after the inclusion of each of the six classifiers. Figure 10(b) shows the relation between the number of added classifiers and the combiner error rate. We see that as more classifiers are added, the error rate decreases to a minimum, then it increases. We investigate this phenomenon by simulating synthetic experiments of a fusion system that combines increasingly weaker classifiers. Additionally, in the next subsection an equation relating fused weak classifiers in a multiple classifier system is proposed. It shows if the inclusion of a weak classifier yields an improved system performance, and when this improvement is achievable. Synthetic experiment results indicate that as weaker classifiers are added, the system improves. This improvement increases reaching a peak value, and then gradually decreases as additional weak classifiers are combined.

Mathematical preliminaries

Kittler & Alkoot (2003) presented a mathematical analysis showing the basic components of the total classification error. We represent the mathematical background based on which the hypothesis and experiments were conducted.

Let us denote the aposteriori probability of class ω_i given observation (pattern) x by $P(\omega_i|x)$. Suppose class ω_s satisfies

$$P(\omega_s|x) = \max_i^N P(\omega_i|x) \tag{1}$$

Where N denotes the number of classes. Thus the bayes optimal decision would be to assign pattern x to class ω_s. Let class ω_j satisfy

$$P(\omega_j|x) = \max_{i=1,i\neq s}^N p(\omega_i|x) \tag{2}$$

Thus in the presence of estimation errors the most likely suboptimal decision will be to assign pattern x to class ω_j. The probability of the label switching error $e_s(x)$ will depend on the distribution $p(\varepsilon_i(x))$ of errors $\varepsilon_i(x)$ corrupting the estimate of the ith class aposteriori probability. It has been shown in [11] that the switching error $e_s(x)$ is given by

$$e_s(x) = \int_{\Delta P}^{\inf} P(\varepsilon_i(x))dt \tag{3}$$

Where $\Delta P(x)$ is the margin between the aposteriori probabilities of the two classes likely to be swapped, i.e.

$$P(x) = P(w_s|x) - P(w_j|x)$$

The additional error $e_A(x)$ [4] is then given as

$$e_A(x) = e_s(x)\Delta P(x) \tag{4}$$

Assuming that the probability of switching between class ω_s and any other class $\omega_i, i \neq j$, is negligible, the actual classifier error e(x) will then be

$$e(x) = e_B(x) + e_s(x)\Delta P(x) \tag{5}$$

Note that in a two class case $\Delta P(x)$ in (3) can be expressed as

$$\Delta P(x) = 1 - 2e_B(x) \tag{6}$$

Thus the error in (4) can be written as

$$e(x) = e_S(x)[1 - 2e_B(x)] + e_B(x) \tag{7}$$

In the multiclass case the margin $\Delta P(x)$ will, in general, be greater than $1 - 2e_B(x)$. However, the above assumption that the switching error between the Bayes optimal decision and any other class $\omega_i, i \neq j$ is negligible, implies, that $P(\omega_i|x) = 0$, for all, $i \neq s$, j in the equation in (5) will be valid and it will represent the worst case scenario.

Kittler & Alkoot (2003) explain that if the estimation error is gaussian, then the distribution of the difference of the two class errors is gaussian with twice the variance. The probability that the difference between the errors is larger than the difference between the two class posterior probability estimates is given by the area under the gaussian tail with a cutoff at the margin between the two classes.

Based on the above and an analysis of the condition where classifiers with different bayes error are combined we reach at the following equation for the total combiner error;

$$E_T = 0.5 + (mN - b) \times erf(\frac{b\sqrt{N} - mN\sqrt{N}}{\sigma})$$

Where $b = 1 - 2e_B + m$, N is the number of combined classifiers, m is the amount by which the bayes error of each additional classifier added to the combiner is higher than the previous classifier. The standard deviation of the classifier Gaussian error is σ. In the following we run synthetic experiments to find the combiner error under the conditions mentioned above. Then compare these results to real MOBIO data experiments to show that our analysis matches real data results.

Experimental methodology

We experiment with synthetic classifier outputs simulating different parameters using Matlab. The parameters under investigation are:

- Margin between class output probabilities.
- Distribution of error on the posterior probability estimates.
- Number of combined classifiers.
- The weakness factor, which is the amount by which the bayes error of each added classifier, is higher than previous one.

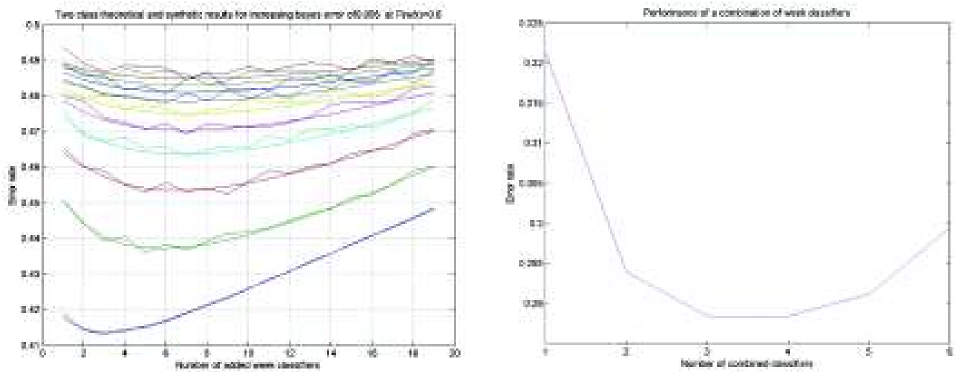
A classifiers class posterior probability estimate is an approximation of the true probability distribution. Consequently, the decision boundary is an approximation of the Bayesian decision boundary. This shift in decision boundary leads to estimation errors. In a synthetic data experiment we simulate the fused classifiers outputs by selecting a true posterior probability value and add to it an estimation error having a zero mean normal distribution. We experiment with different standard deviation values ranging between 0.1 and 1 at steps of 0.1. The experiments are repeated for different posterior probability levels ranging from 0.6 to 1 at steps of 0.1. All experiments are repeated 100000 times then averaged to achieve statistical reliability and smother curve of the synthetic results.

To compare the theory and synthetic experiments with experiments on a real world problem using MOBIO, data we use six different classifiers that are based on different problem spaces. The different problem spaces are a close representation of different underlying bayes distributions per classifier. To obtain the six classifier outputs we use scores of two classifiers LBP and LDA. These scores are ranked based on score value, ranked according to quality measure 3 and ranked according to quality measure 16. The three rankings yield three different classifier outputs for a single problem. Therefore, using two classifier types, LDA and LBP, we have a total of six classifier outputs or scores for fusion. We have more than 50 scores for each test sample out of which we may use the top R ranks, where R in Figure 10(b) is 9.

Experimental results

Figure 10(a) shows experimental results of the synthetic classifiers, under various parameter values, for the two class case. The results show when an optimum performance may be achieved.

In all of the figures, the x-axis represents the number of classifiers added to the combined system. Each added classifier is weaker by a constant weakness factor, m. That is, each added classifier has a larger bayes error by a constant amount, m. For experiments of this report we select the weakness factor to be 0.005. The y-axis represents classification error. Figure 10(a) shows results for Sum fusion method at 5 posterior probability values at different standard deviation values. The figure contains ten double curves representing the ten standard deviations. For each standard deviation one of the two curves is smooth and is for the error rate based on the theoretical error from equation 5, while the noisy curve is based on the synthetic experiments. Each posterior probability value also represents a bayes error value. Hence, we have initial classifier bayes error values ranging from 0 to 0.4, at steps of 0.1. Again the smother curve is using theoretical equation 5 while the noisy curve is based on synthetic experiments. It is an increasing curve because we gradually introduce weaker classifiers.



(a) (b)

Fig. 10 (a). Total synthetic and theoretical error rate at fixed posterior probability of 0.6 for ten Gaussian error standard deviation values.

Fig. 10(b) Error rate of a combiner using six weak classifiers tested on the MOBIO data set, using rank size 9.

We find that as σg ncreases switching error beyond bayes error appears. It is here that fusion becomes useful. As we added classifiers, the error rate decreased until it reached a minimum. Beyond this minimum the addition of more classifiers would not reduce the total error rate. We find that the slope of the error curve depends on four parameters. The weakness factor, number of classifiers, the degree of estimation error, represented by the standard deviation, σ and the initial classifier bayes error, also called the margin between classes. We find that, the larger the standard deviation, the more we gain from combining weak classifiers. We note that the switching error above the bayes error is reduced rapidly as more classifiers are added. However, the average bayes error increases due to the classifiers being weaker. For some initial bayes error values, at some σ, the total error curve and the average bayes curve meet at a point beyond the minimum. Ideally we would prefer if they meet at the minimum. Based on the initial bayes error, below a certain minimum standard deviation,no gain is achieved from combining any classifier that is weaker than the initial by a certain weakness factor.

CONCLUSION

Video sequences are used in a biometric authentication application, where we aim to improve the performance of existing score fusion systems in experiments on the MOBIO-Banca biometric data set. Scores from LBP and LDA classifier outputs are ranked and fused using various fusion strategies. Results show that the use of an adaptive boundary using information from the training set distribution curves, and the use of quality measure based ranking of scores yields higher classification rates than fixed boundary. It also leads to equally high imposter and genuine rates. Among other fusion strategies Sum and MProduct are used and their performances are compared. We found that Sum leads all fusion strategies and mostly outperforms MProduct except for one case where fixed boundary results reach an optimum rate of 100%, and only adaptive boundary using MProduct is able to reach this rate. Using adaptive boundary, Sum yields higher improvement percentages than MProduct and at most rank sizes, while MProduct yields smaller improvements and only at small rank sizes.

Additionally, we aim to explain the degradation of the MOBIO combiner as the number of combined classifiers increases. Combiner error rate based on synthetic experiments and on an equation relating fused weak classifiers follows a hyperbolic curve. We show that, based on certain conditions and on values of critical parameters, the performance of a combiner improves before degrading due to the addition of weaker classifiers. Experiments on real world classifiers using MOBIO data show the combiner follows the synthetic curves and degrades after the third classifier, as predicted.

ACKNOLEDGMENT

The author would like to sincerely thank PAAET for supporting this project under grant number TR-09-03.

REFERENCES

Abaza, A. & Ross, A. 2009. Quality based rank-level fusion in multibiometric systems. IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems. BTAS. 1-6.

Alkoot, F. M. & Kittler, J. 2002. Modified product fusion, Pattern Recognition Letters. **23**(8): 957-965.

Alkoot, F. M. & Kittler, J. 1999. Experimental evaluation of expert fusion strategies. Pattern recognition letters. **20**(11-13):1361-1369.

Bailly-Bailliere, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., Matas, J., Messer, K., Popovici, V., Porée, F., Ruiz, B. & Thiran, J.P. 2003. The BANCA database and evaluation protocol. Lecture Notes in Computer Science, **2688/1057**.

Chan, C. 2008. Multi-scale local binary pattern histogram for face recognition. PhD thesis, University of Surrey.

Fatukasi, O., Kittler, J. & Poh, N. 2007. Quality controlled multimodal fusion of Biometric experts, Progress in pattern recognition, Image Analysis and Applications Lecture Notes in Computer Science. **4756**: 881-890

Fierrez-Aguilar, J. Ortega-Garcia, J., Gonzalez-Rodriguez, J. & Bigun, J. 2005. Discriminative multimodal biometric authentication based on quality measures. Pattern recognition. **38** (4) 777-779.

Fukunaga, K. 1990. Introduction to statistical pattern recognition. Academic Press, 2nd edition.

Heusch, G., Rodriguez, Y. & Marcel, S. 2006. Local binary pattern as an image preprocessing face authentication, in Proc. 7th Int'l Conf. Automatic Face and Gesture Recognition (FGR06). Washington, DC. 9-14.

Hong, L. & Jain, A. 1998. A. Integrating faces and fingerprints for personal identification, IEEE transactions on pattern analysis and machine Intelligence, Vol. **20**, No. 12, Dec.

Kittler, J. & Alkoot, F.M. 2003. Sum versus vote fusion in multiple classifier systems. IEEE transactions on pattern analysis and machine intelligence **25**(1): 110-115.

Kittler, J., Poh, N., Fatukasi, O., Messer, K., Kryszczuk, K., Richiardi, J. & Drygajlo, A. 2007. Quality dependent fusion of intramodal and multimodal biometric experts, in SPIE Biometric technology for human identification IV. **6539**, 653903

Matas, J., Hamouz, M., Jonsson, K., Kittler, J., Li, Y., Kotropoulos, C., Tefas, A., Pitas, I., Tan, T., Yan, H., Smeraldi, F., Begun, J., Capdevielle, N., Gerstner, W., Ben-Yacoub, S., Abdeljaoued, Y. & Mayoraz, E. 2000. Comparison of face verification results on xm2vts database, in Proc. 15th int'l Conf. Pattern Recognition in Barcelona 858-863.

MOBIO, The project, http://www.mobioproject.org/

Nandakumar, K., Yi Chen, Jain, A. K. & Dass, S.C. 2006. Quality-based score level fusion in multibiometric systems. ICPR. 18th International Conference on ˜Pattern Recognition **4**: 473 - 476.

Pietikainen, M., Hadid, A., Zhao, G. & Ahonen, T. 2011. Computer vision using local binary patterns. Springer.

Poh, N. & Kittler, J. 2008. A family of methods for quality-based multimodal biometric fusion using generative classifiers.10th International conference on control, Automation, Robotics and vision. ICARCV 1162 - 1167

Poh, N. & Kittler, J. 2012. A Unified framework for biometric expert fusion incorporating quality measures. IEEE transactions on battern analysis and machine intelligence, **34**(1): 3 - 18

Poh, N., Kittler, J. & Alkoot, F. 2012. A discriminative approach to video-based Score-level fusion for biometric authentication, in proceedings of the 21st ICPR conference, November, Tsukuba Science City, JAPAN.

Ross, A. & Poh, N. 2009. Multibiometric systems: Overview, Case studies and open issues in the Handbook of Remote Biometrics for Surveillance and Security- Advances in Pattern Recognition, Springer: 273-292.

Rouvier, M., Matrouf, D. & Linares, G. 2009. Factor analysis for audio-based video genre classification, in: International conference on speech communication and technology (Interspeech).

Savvides, C., Ng, M. & Khosla, P. K. 2005. Real-timeface verification system on a cell-Phone using advanced correlation filters. Fourth IEEE workshop on automatic identification advanced technologies in New York, USA. 57 - 62.

Tresadern, P.A. McCool, C., Poh, N., Matejka, P., Hadid, A., Levy, C., Cootes, T. F. & Marcel, S. 2013. Mobile biometrics: Combined face and voice verification for a mobile platform, IEEE pervasive computing, **12**(1):79-87.

Submitted : *29/07/2013*
Revised : *31/12/2013*
Accepted : *11/02/2014*

المصادقة البيومترية متعددة الوسائط باستخدام حدود القرار التكيفية

فؤاد محمد الكوت

المعهد العالي للإتصالات والملاحة، الهيئة العامة للتعليم التطبيقي والتدريب
ص.ب: 313، الشهداء، 48554، الكويت
fm.alkoot@paaet.edu.kw, f_alkoot@yahoo.com

خلاصة

لقد حسنت التطورات الأخيرة في تقنيات التوثيق التعريفية متعددة الوسائط من مستوى موثوقيتها، حيث كان لاستخدام أساليب الدمج المختلفة الأثر الأكبر في رفع مستوى الأداء. نهدف من خلال هذا البحث إلى إيجاد استراتيجية اندماج تؤدي إلى أفضل نسبة تصنيف في التجارب على مجموعة بيانات بنكا البيومترية – MOBIO. يتم ترتيب ودمج نتائج مخرجات المصنف باستخدام استراتيجيات متعددة للدمج. تعطي طريقتنا المقترحة التي تجمع ما بين حدود التكيف وقياس الجودة في ترتيب النتائج أداءاً ملحوظا بالمقارنة مع الأساليب الحالية للحدود الثابتة. لقد أعطى الدمج الجمعي تحسيناً كبيراً جدا ونجاحاً على خلاف اسلوب الضرب المطور. هذا ونفسر التدهور في موحد MOBIO المسلسل من الناحية التحليلية ومن خلال التجارب التركيبية. كما نبين أن أداء مصنفات المجمع الضعيفة تتبع منحنى قطعي كلما تم إضافة مصنفات ضعيفة.

الكلمات الرئيسية: بيانات بنكا؛ البيومترية؛ المصنف المدمج؛ MOBIO؛ قياس الجودة.