A COMPARATIVE STUDY OF DATA FUSION STRATEGIES IN FACE VERIFICATION

Mohammad T Sadeghi and Josef Kittler

Centre for Vision, Speech and Signal Processing
School of Electronics and Physical Sciences, University of Surrey, Guildford GU2 7XH, UK
phone: +44 1483 686030, fax: +44 1483 686031, email: {M.Sadeghi,J.Kittler}@eim.surrey.ac.uk

ABSTRACT

In this paper, the merits of fusing colour information in a face verification system is studied. Three different levels of fusion, namely, signal, feature and decision levels are considered. The study is performed on a fisherface-based (LDA) verification system considering the Gradient Direction metric as the scoring function. We show that almost all the fusion methods enhance the performance of the system. However, despite the common use of the fusion at the signal level realised by creating intensity images, the other fusion methods specially the decision level fusion using score averaging are more effective.

1. INTRODUCTION

Colour has been shown to play an important role in face verification. In the work of Nastar and Mitschke [6], the colour information, accounted for via a colour histogram in the R,G,B space, forms a basis of decision making in one of a number of experts, the outputs of which are combined by a voting process. Marcel and Bengio [4] use colour as an information source to generate complementary features for their face verification system. They combine face appearance, as captured by an intensity image, with face colour represented by a colour histogram. For computational reasons, a simplified histogram representation is adopted. Accordingly, rather than using the standard 3D colour histogram, 3 one dimensional histograms, one for each of the colour channels, are used instead. These 1D histograms, derived from the R, G,and B channels are concatenated to create a feature vector which is then appended to the pixel intensity measurements and input to a neural network. In experiments on the XM2VTS database [5], performed according to the Lausanne protocol, the authors have shown that the combination of texture (appearance) and colour features improves the performance of the face verification system.

In another study, reported in [1], the problem of colour conversion for face verification systems was considered. The conversion methods investigated include the Karhunen-Loeve transform, a line-fitting in the RGB space as well as a genetic algorithm to optimise the sum of eigenfaces picked from the respective colour channels. The authors argued that in comparison with the systems that operate using solely the intensity information, the recognition accuracy can be improved using one of the above RGB image conversion techniques. However, the performance of their conversion algorithms was not properly evaluated in a real face recognition/verification system.

A somewhat different approach has been adopted in [10], where each channel of the multispectral image is treated as a separate grey level image, which is projected into the corresponding Principal Component Analysis space. A face recognition system is then constructed in each of the three PCA spaces and their outputs combined at the decision level by weighted averaging. The authors explored different colour spaces for their method and concluded that the use of YUB and SV colour representations led to improved performance.

A similar approach has been adopted in [9] who performed a much more systematic evaluation of signal, feature and decision level fusion of data derived from a multispectral face image. In contrast to [10], the authors focused on face verification using normalised correlation in Linear Discriminant Analysis (LDA) spaces associated with the respective R,G, B colour channels. The results demonstrated that the most beneficial were the decision level and feature level fusion but the decision level fusion was computationally the simplest.

The aim of this paper is to repeat the above study using a novel similarity measure known as the Gradient Direction metric, which has been shown to have considerable promise [3]. As in [9] the assessment of signal, feature and decision level fusion is carried out on the BANCA database using the standard BANCA experimental protocols.The experimental results confirm the superiority of this metric to the Normalised Correlation. We also show that, in conjunction with this metric, one of the feature level fusion schemes and the score averaging method offer the best performances.

The paper is organised as follows. In Section 2 we overview the face verification process. In Section 3 the various sensory data fusion strategies investigated are described. In Section 4 the adopted scoring function, Gradient Direction metric, is discussed. The experimental set up adopted for the comparative study is detailed in Section 5. The results of experiments are presented in Section 6. A discussion of the results as well as the main conclusions can be found in Section 7.

2. FACE VERIFICATION PROCESS

The face verification process consists of three main stages: face image acquisition, feature extraction, and finally decision making. The first stage involves sensing and image preprocessing the result of which is a geometrically registered and photometrically normalised face image. Briefly, the output of a physical sensor (camera) is analysed by a face detector and once a face instance is detected, the position of the eyes is determined. This information allows the face part of the image to be extracted at a given aspect ratio and resampled to a pre-specified resolution. Finally, the extracted face image is photometrically normalised to compensate for illumination changes. The output of these preprocessing algorithms is still in the form of an image, i.e. we are dealing with signals at this level. If the sensor is a colour camera, then the resulting representation will be three geometrically registered and photometrically normalised images corresponding to the R,G and B channels respectively. Any combination of these, at this stage, will be referred to as signal level fusion.

In the second stage of the face verification process the face image data is projected into a feature space. Different features can be extracted from each image. Regardless of the feature definition, at this processing stage the image nature of the data is lost. Each channel is transformed into a set of measurements and any combination of features derived from different channels will be referred to as feature level fusion.

The final stage of the face verification process involves matching and decision making. Basically the features extracted for a face image to be verified are compared with a stored template, that was

The financial support from the EU project VAMPIRE is gratefully acknowledged.

acquired on enrolment. The score output by the matching process, which measures the degree of similarity between the test image and the template, is then compared to a threshold in order to decide whether the claimed identity should be accepted or rejected. If this final stage of processing is applied to each data channel separately, we end up with a number of scores which then have to be fused to obtain the final decision. Any combination of the results of processing at this level will be referred to as decision level fusion.

3. FUSION METHODS

As we mentioned, in a face verification system, the fusion process can be performed in three different levels. In this section the adopted fusion strategies are detailed.

3.1 Signal level fusion

Two simple signal level fusion methods were considered. In the first method the raw sensor outputs, i.e. the R,G,B channels are combined by averaging to produce the intensity image, $I(x,y) = \frac{1}{3}[R(x,y) + G(x,y) + B(x,y)]$. This combined image is then photometrically corrected before feature extraction and decision making.

The second signal level fusion method simply involved concatenating the geometrically registered and photometrically normalised images $I_i(x,y), i = 1,2,3$ obtained for the three colour channels into a single multidimensional image, $\hat{I}(x,y)$.

3.2 Feature level fusion

Again, two methods have been considered. Let $\mathbf{f}_i$ be the set of features extracted from image $I_i(x,y)$. At the feature level, we fused the individual channel representations by creating a concatenated feature vector $\mathbf{f}$ as

$$\mathbf{f} = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \mathbf{f}_3 \end{bmatrix} \tag{1}$$

The feature vector $\mathbf{f}$ is then used to make the verification decision.

In the second method, the feature vector $\mathbf{f}$ was passed through a second feature extraction system, i.e. the feature vectors $\mathbf{f}_i, i = 1,2,3$ were combined to form the final feature vector.

3.3 Decision level fusion

In the score-based verification systems, a score s reflecting the quality of match between the claimed identity template and the observation is compared to a pre-set threshold η to determine whether the claim is genuine (class ω_a) or impostor(class ω_b).

$$s(\mathbf{f}) \underset{\omega_a}{\overset{\omega_b}{\lessgtr}} \eta \tag{2}$$

When multiple channels are available, a score $s_i = s(\mathbf{f}_i)$ can be obtained from each channel. The problem now is to find a function f so that the decision rule

$$f\left(s_1, s_2, s_3\right) \underset{\omega_a}{\overset{\omega_b}{\lessgtr}} \eta \tag{3}$$

leads to a higher verification performance. This level of fusion is also known as the confidence level or soft fusion [7]. Since the adopted experts in our case deliver a similar level of accuracy, their combination should either attach the same weight to all three scores or have a mechanism for selecting the best score. Thus in this study, two simple combining strategies have been considered. In the first method the samples average is considered as the final score.

$$f\left(s_1, s_2, s_3\right) = \frac{1}{3}[s_1 + s_2 + s_3] \tag{4}$$

The second method is to select the best of the three scores as the final score. In the case of a dissimilarity score like the Euclidean distance, this results in taking the minimum score for making the decision, i.e.

$$f\left(s_1, s_2, s_3\right) = \min\left(s_1, s_2, s_3\right) \tag{5}$$

while in the case of a similarity function such as normalised correlation scoring function, the maximum rule is considered.

$$f\left(s_1, s_2, s_3\right) = \max\left(s_1, s_2, s_3\right) \tag{6}$$

4. GRADIENT DIRECTION METRIC

In a face verification system, a matching scheme measures the similarity or distance of the test sample, x to the template of the claimed identity, μ_i. Note that x and μ_i are the projections of the test sample and class mean into the feature space respectively. In [3], it has been demonstrated that a matching score based on Normalised Correlation scoring function is more efficient than the simple Euclidean Distance function.

In [3] an innovate metric called the *Gradient Direction metric* (GD) has been proposed. In this method the distance between a probe image and a model is measured in the gradient direction of the aposteriori probability of the hypothesised client identity. A mixture of Gaussian distributions with Identity covariance matrix has been assumed as the density function of the possible classes of identity. In [8], we revisited the theory of the Gradient Direction metric and extended it to a Generalised Gradient Direction metric. We demonstrated that applying GD metric using either a general covariance matrix derived from the training data or an isotropic covariance matrix with a variance of the order of the variation of the image data in the feature space is even more efficient than the NC function. The proposed optimal matching score has been defined as

$$s_O = \frac{||(\mathrm{x} - \mu_i)^T \nabla_O P(i|\mathrm{x})||}{||\nabla_O P(i|\mathrm{x})||} \tag{7}$$

where $\nabla_O P(i|\mathrm{x})$ refers to the gradient direction. In the generalised form of the GD metric, the optimal direction would be

$$\nabla_G P(i|\mathrm{x}) = \Sigma^{-1} \sum_{\substack{j=1 \\ j \neq i}}^{m} p(\mathrm{x}|j)(\mu_j - \mu_i) \tag{8}$$

where $p(\mathrm{x}|j)$ is the j-th client measurement distribution. Considering an isotropic structure for the covariance matrix, i.e. $\Sigma = \sigma \mathrm{I}$, equation 8 could be simplified as:

$$\nabla_I P(i|\mathrm{x}) = \sum_{\substack{j=1 \\ j \neq i}}^{m} p(\mathrm{x}|j)(\mu_j - \mu_i) \tag{9}$$

Note that the magnitude of the σ will affect the direction through the values of $p(\mathrm{x}|j)$.

5. EXPERIMENTAL DESIGN

In this section the BANCA database and the main specifications of the experimental setup are briefly described.

5.1 BANCA database

The BANCA database contains 52 subjects (26 males and 26 females). Each subject participated to 12 recording sessions in different conditions and with different cameras. Sessions 1-4 contain data under *Controlled* conditions while sessions 5-8 and 9-12 contain *Degraded* and *Adverse* scenarios respectively. Each session contains two recordings per subject, a true client access and an informed imposter attack. For the face image database, 5 frontal face images have been extracted from each video recording, which are supposed to be used as client images and 5 impostor ones. In order to create more independent experiments, images in each session have been divided into two groups of 26 subjects.

In the BANCA protocol, 7 different distinct experimental configurations have been specified, namely, Matched Controlled (MC), Matched Degraded (MD), Matched Adverse (MA), Unmatched Degraded (UD), Unmatched Adverse (UA), Pooled test (P) and Grand test (G). Table 1 describes the usage of the different sessions in each configuration. "TT" refers to the client training and impostor test session, and "T" depicts clients and impostor test sessions. The decision function can be trained using only 5 client images per person from the same group and all client images from the other group.

	MC	MD	MA	UD	UA	P	G
1	TT			TT	TT	TT	TT
2	T					T	T
3	T					T	T
4	T					T	T
5		TT					TT
6		T		T		T	T
7		T		T		T	T
8		T		T		T	T
9			TT				TT
10			T		T	T	T
11			T		T	T	T
12			T		T	T	T

Table 1: The usage of the different sessions in the BANCA experimental configurations.

5.2 Experimental Setup

The effects of different fusion methods are experimentally evaluated on the BANCA database using the configurations discussed in the previous section. The original resolution of the image data is 720 × 576. The experiments were performed with a relatively low resolution face images, namely 64 × 49. The results reported in this article have been obtained by applying a geometric face registration based on manually annotated eyes positions. Histogram equalisation was used to normalise the registered face photometrically. The feature selection process is performed using the linear Discriminant Analysis (LDA). The XM2VTS database [5] was used for calculating the LDA projection matrix. In this study, the Gradient Direction metric (GD) was used as the scoring function. An isotropic structure was considered for the covariance matrix of the clients distribution. So, the matching scores were calculated using Equations 7 and 9. The thresholds in the decision making system have been determined based on the Equal Error Rate criterion, i.e. where the false rejection rate (FRR) is equal to the false acceptance rate (FAR). The thresholds are set either globally (*GT*) or using the client specific thresholding (*CST*) technique [2]. As we mentioned earlier, in the training sessions of the BANCA database 5 client images per person are available. In the case of global thresholding method, all these images are used for training of the clients template. The other group data is then used to set the threshold. While using the client specific thresholding strategy, only two images are used for the template training and the other three along with the other group data are used to determine the thresholds. Moreover, in order to increase the number of data used for training and to take the errors of the geometric normalisation into account, 24 additional face images per each image are generated by perturbing the location of the eyes position around the annotated positions

6. EXPERIMENTAL RESULTS

Table 2 shows the performance of the face verification system considering the R, G and B channels individually using the GT and CST methods. We can see that the verification results based on the respective colour channels are highly correlated. In general as one would expect, for matched protocols the performance is better than for the unmatched protocols due to generalisation problem posed by the latter.

It can be seen that the CST technique is superior in the matched scenarios while the GT method gives a better performance on the unmatched protocols. The reason is that, as we mentioned earlier, images in each session have been divided into two groups of 26 subjects. In the GT method the global threshold of each test group is calculated using the other group of test data. It means that the evaluation and test data have always the same image quality. In the case of the CST technique, we need to have available the client and impostors scores of each client individually. Thus we divided the training images into two subsets, client template training and client scores evaluation images. The impostor scores are then calculated using the other group of the test data (the group which the client does not belong to). It means that in the threshold evaluation of the unmatched experiments, images with different quality are used for calculating the client and impostors scores, while in the test stage all probe images (clients and impostors) have the same quality as the ones which are used for the impostors scores evaluation. Note that P protocol which is a collection of the MC, UD and UA protocols can mainly be considered as an unmatched protocol while the G protocol which involves data from different scenarios for both training and test can be seen as a matched protocol. The results of various fusion experiments reported below were obtained with the CST method in the case of the matched protocols and GT method in the unmatched cases.

		GT			CST		
		FAR	FRR	TER	FAR	FRR	TER
Mc	R	1.635	4.615	6.25	0.865	5.513	6.378
	G	2.308	5.897	8.205	0.481	3.205	3.686
	B	2.596	6.026	8.622	0.769	5.513	6.282
Md	R	8.75	8.462	17.21	1.058	7.821	8.878
	G	9.423	8.718	18.14	1.058	9.103	10.16
	B	6.154	7.051	13.21	2.019	4.615	6.635
Ma	R	7.589	7.051	14.64	2.019	7.692	9.712
	G	6.923	8.205	15.13	1.442	8.333	9.776
	B	6.058	5.769	11.83	1.923	5.641	7.564
Ud	R	13.65	14.74	28.4	2.019	33.08	35.1
	G	14.33	15.51	29.84	0.577	43.08	43.65
	B	20.58	21.03	41.6	0.865	56.54	57.4
Ua	R	16.83	17.56	34.39	1.154	35.00	36.15
	G	14.52	14.36	28.88	0.577	49.87	50.45
	B	17.02	18.08	35.1	1.538	50.13	51.67
P	R	11.31	10.85	22.17	1.378	25.94	27.32
	G	10.35	10.13	20.48	0.417	33.76	34.18
	B	13.94	14.19	28.13	1.026	43.76	44.79
G	R	2.276	3.034	5.31	2.596	1.239	3.835
	G	2.212	2.393	4.605	1.891	1.752	3.643
	B	2.147	2.607	4.754	2.179	2.65	4.829

Table 2: ID verification results on BANCA configurations using NC method in the R, G and B colour spaces considering global and client specific thresholding methods.

Signal level fusion: The simplest fusion method which is commonly used in the face verification systems is to produce an intensity image by averaging the R,G,B channels. The results using this method has been reported in the left part of table 3 (Signal 1). At the signal level, the other strategy is to generate a multidimensional image by concatenating the R,G,B images. The right part of the table 3 (Signal 2) indicates the results using this method.

	Signal 1			Signal 2		
	FAR	FRR	TER	FAR	FRR	TER
Mc	1.058	4.744	5.801	0.769	1.41	2.179
Md	1.25	7.051	8.301	1.25	10.77	12.02
Ma	1.346	6.538	7.885	2.981	7.179	10.16
Ud	13.94	15.26	29.2	17.69	18.46	36.15
Ua	16.06	16.15	32.21	16.73	17.05	33.78
P	11.57	10.64	22.21	11.76	12.18	23.94
G	2.019	1.581	3.6	1.859	0.940	2.799

Table 3: Verification results using the signal level fusion methods.

Feature level fusion: In the feature level, we fused the individual channel representations as shown by equation 1. The decision is then made either directly using the acquired feature vector, **f**, (Feature 1) or after combining the individual channel features using a second stage LDA (Feature 2). Table 4 contains the experimental results using these methods of fusion.

	Feature 1			Feature 2		
	FAR	FRR	TER	FAR	FRR	TER
Mc	0.4808	3.846	4.327	0.865	3.846	4.712
Md	0.9615	5.513	6.474	0.769	9.872	10.64
Ma	1.25	5.385	6.635	1.538	4.744	6.282
Ud	14.23	13.85	28.08	18.75	19.74	38.49
Ua	14.42	15.26	29.68	16.06	16.15	32.21
P	10	10.26	20.26	12.69	12.69	25.38
G	1.635	1.239	2.874	1.41	1.197	2.607

Table 4: Results when the features obtained from different channels were fused (Feature level fusion).

Decision level fusion: In this level, scores from individual channels were calculated first. The resulted scores were then combined using two simple methods, the average (Decision 1) or the best (Decision 2) rules. Note that since the Gradient Direction metric is a dissimilarity (distance) function, Equation 5 was used to find the best score. The associated results are shown in table 5.

	Decision 1			Decision 2		
	FAR	FRR	TER	FAR	FRR	TER
Mc	0.288	4.103	4.391	0.288	4.872	5.16
Md	0.769	6.154	6.923	0.865	6.282	7.147
Ma	1.154	5.897	7.051	1.635	5.897	7.532
Ud	14.33	14.1	28.43	15.96	16.79	32.76
Ua	14.52	13.97	28.49	15.1	15.26	30.35
P	10.19	10.09	20.28	12.02	10.98	23.00
G	1.795	1.282	3.077	2.051	1.197	3.248

Table 5: Decision level fusion results using the averaging and the best rules.

7. DISCUSSION AND CONCLUSIONS

From the fusion results, a number of conclusions can be derived. First of all, almost all the fusion methods, enhance the performance. However, despite the common use of the intensity space in the face verification systems, not too much benefit is gained from working with the intensity image. The other fusion methods appear to be more beneficial.

The second signal level fusion is surprisingly good in the Mc and G protocols i.e. either where both training and test data have been recorded in the controlled conditions or where a relatively good sized training data is available. In fact in this method of fusion, since we concatenate the image data in different spaces together, the dimensionality of the input data is relatively large. Now, when the data is projected into the LDA space, the number of the dimensions, D, is reduced to $d = M - 1$, where M is the number of the classes (subjects) which are used for the training of the LDA axes. Note that d does not depend on D. In the second method of signal level fusion, more information about the input image is available in the representation vector which is three times bigger than the signal extracted from the individual channels. The most important issue in this stage is to extract the useful d-dimensional information efficiently. This matter is even more important, when we are building the representative models (template) of the classes (clients). Obviously, when the test and training data both have been collected in a controlled conditions or the size of the template training data is larger a better model can be obtained.

At the feature level, in all cases, especially Md and Ud protocols, the first method leads to a better performance. The main reason that this method does not work well for the degraded data is that, as we mentioned, the XM2VTS database is used for the LDA training purpose. The quality of the images of this database is more similar to the controlled and adverse data in the Banca database. Therefore, the obtained LDA axes probably do not represent the degraded data very well. In the case of the 2 stage LDA, this problem is pronounced.

In the decision level fusion, the score averaging method is superior. Overall, the first method of fusion at the feature level and the decision level fusion by score averaging give the best performances. In [9], a similar study was performed using the Normalised Correlation rather than the Gradient Direction metric. The best performance was achieved again using either the feature level fusion by concatenating the R,G,B channels data in the LDA space or the score averaging method. Table 6 shows the corresponding results.

	Feature 1			Decision 1		
	FAR	FRR	TER	FAR	FRR	TER
Mc	1.442	6.538	7.981	0.7692	6.667	7.436
Md	2.308	9.359	11.67	2.692	9.872	12.56
Ma	3.942	9.744	13.69	3.75	10.13	13.88
Ud	13.46	13.33	26.79	15.38	16.03	31.41
Ua	17.21	17.95	35.16	18.08	17.05	35.13
P	12.82	13.38	26.2	13.65	15.26	28.91
G	5.577	2.906	8.483	6.154	2.735	8.889

Table 6: Verification results using NC considering the first feature level fusion and the score averaging methods.

These results demonstrate that the GD metric leads to a better performance in all possible scenarios. As far as the fusion method is concerned, considering other aspects such as the computational complexity involved both in the design and testing (routine operation) and the system robustness in the case of channel failure, the overall recommendation would be to go for the decision level fusion by averaging, which delivers good performance in the simplest possible way.

REFERENCES

[1] C.F. JonesIII and A.L. Abbott. Optimization of color conversion for face recognition. *EURASIP special issue on biometric signal processing*, to appear 2004.

[2] K. Jonsson, J. Kittler, Y. Li, and J. Matas. Support vector machines for face authentication. In *T. Pridmore and D. Elliman, editors, Proceedings of BMVC'99*, pages 543– 553, 1999.

[3] J. Kittler, Y. P. Li, and J. Matas. On matching scores for lda-based face verification. In M Mirmehdi and B Thomas, editors, *Proceedings of British Machine Vision Conference 2000*, pages 42–51, 2000.

[4] S. Marcel and S. Bengio. Improving face verification using skin color information. In *6th International Conference on Pattern Recognition*, volume 2, pages 20378–20382, 2002.

[5] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Second International Conference on Audio and Video-based Biometric Person Authentication*, March 1999.

[6] C. Nastar and M. Mitschke. Real-time face recognition using feature combination. In *Int. Conf. on Automatic Face and Gesture Reconition,AFGR98*, pages 312–317, 1998.

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

[8] M. Sadeghi and J. Kittler. Decision making in the lda space: Generalised gradient direction metric. submitted to the 6th Int. Conf. on Automatic Face and Gesture Reconition, Seoul, Korea, May 2004.

[9] M. Sadeghi, J. Kittler, and K. Messer. On data fusion in face verification. submitted to the 17th Int. Conf. on Pattern Recognition, ICPR'04, Cambridge, UK, August 2004.

[10] L. Torres, J.Y. Reutter, and L. Lorente. The importance of the color information in face recognition. In *International Conference on Image Processing, ICIP99*, pages 627–631, 1999.