Pattern Recognition Letters
journal homepage: www.elsevier.com

A Confidence-Based Late Fusion Framework For Audio-Visual Biometric Identification

Mohammad Rafiqul Alam[a,**], Mohammed Bennamoun[a], Roberto Togneri[b], Ferdous Sohel[a]

[a] *School of Computer Science and Software Engineering, The University of Western Australia, Crawley, WA 6009, Australia*
[b] *School of Electrical, Electronic and Computer Engineering, The University of Western Australia, Crawley, WA 6009, Australia*

ARTICLE INFO

Article history:
Received XX XXX XXXX
Received in final form XX XXX XXXX
Accepted XX XXX XXXX
Available online XX XXX XXXX

Communicated by XXXX

Keywords:
Score-Level Fusion
Rank-Level Fusion
Audio-Visual Biometrics

ABSTRACT

This paper presents a confidence-based late fusion framework and its application to audio-visual biometric identification. We assign each biometric matcher a confidence value calculated from the matching scores it produces. Then a transformation of the matching scores is performed using a novel confidence-ratio (C-ratio) i.e., the ratio of a matcher confidence obtained at the test phase to the corresponding matcher confidence obtained at the training phase. We also propose modifications to the highest rank and Borda count rank fusion rules to incorporate the matcher confidence. We demonstrate by experiments that our proposed confidence-based fusion framework is more robust compared to the state-of-the-art late (score- and rank-level) fusion approaches.

[1]

1. Introduction

Identification systems have long been used for criminal investigations and are now increasingly being used for various real life applications, e.g., computer login, physical access control, time attendance management (Murakami and Takahashi, 2009). The identification task can be more challenging compared to the verification when the number of enrolled users is large. One way of developing an accurate identification system is to use instances from multiple modalities (Nandakumar et al., 2009), such as the face image, speech and fingerprint. Multiple modalities are usually combined either at an early or at a late stage of recognition.

Existing score fusion techniques can be categorized into four groups. The **first** group are the *transformation-based* fusion methods: the match scores are transformed into (not necessarily) a common range and then simple rules (e.g., product, sum, mean, max, etc.) (Kittler et al., 1998) are applied to them. The **second** group are the *density-based* fusion methods: underlying match score densities are first estimated and then the joint likelihood ratio is calculated (Nandakumar et al., 2008) (Abaza and Ross, 2009). The **third** group are the *classifier-based* fusion methods: the match scores are considered as features of a fusion classifier (Sanderson and Paliwal, 2002) (Ross et al., 2006) (Tao and Veldhuis, 2008). **Recently**, another framework reported in (Poh and Kittler, 2012) is known as the *quality-based* fusion approach: the modalities are weighted based on the quality measure of the corresponding biometric samples. Although score-level fusion is commonly adopted in multimodal biometrics, rank-level fusion is considered a more viable option (Abaza and Ross, 2009) for systems operating in the identification mode. The ranked lists from different matchers are combined to reach a final decision (Ho et al., 1994). Unlike score-level fusion, the accurate estimation of the underlying genuine/impostor score distributions and normalization are not required in rank-level fusion.

[**]Corresponding author: Tel.: +61-8-6488-4775; fax: +61-8-6488-1089;
e-mail: 21100717@student.uwa.edu.au (Mohammad Rafiqul Alam)

[1]**NOTICE: this is the author's version of a work that was accepted for publication in Pattern Recognition Letters. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in PUBLICATION, [VOL**, ISSUE**, (DATE)] DOI: 10.1016/j.patrec.2014.10.006**

1.1. Motivation and Contributions

In real life scenario, a biometric system may encounter noisy outdoor environments. For example, a missing/wanted person detection system being operated at an airport, train/bus station, or some other public place. The biometric traits to be used in these types of applications must be unobtrusive (e.g., audio-visual) and the user's claim of an identity for verification may not be available. The challenge is, in an outdoor environment, the captured biometric samples may contain noise or corruption due to various environmental conditions (e.g., windy/gloomy atmosphere and low configuration capture devices). Quality-based fusion (Poh and Kittler, 2012) offers a solution to this problem by measuring the quality of the input samples and passing this bit of additional information to the fusion module. However, measuring the quality at the signal level is particularly difficult from face image samples (Chetty and Wagner, 2008) because the source of statistical deviation is varied and difficult to model. Alternatively, the matching scores from a biometric matcher provide a good indication of the quality of the input samples, given the matcher's decision making ability is strong under normal circumstances. Incorporating a system's confidence in the participating modalities (matchers) has not been well studied and this lack of development has also been highlighted in (Marasco and Sansone, 2011).

Our motivation is to develop a fusion framework that works well when either or all input samples presented are contaminated by noise (e.g., detector noise, bit-error, transmission error and additive noise). The core contributions of this paper are listed below:

- We propose a novel *C-ratio* which is the ratio of the matcher confidence obtained from the matching scores during the test phase to the maximum value of the matcher confidence obtained at the training phase (Section 3.1).
- We also propose a *confidence factor* to be used in rank-level fusion (Section 3.2). Our proposed confidence-based rank-level fusion approach considers that only the ranked lists and the maximum matcher confidence obtained at the training phase are available to the fusion module.
- We evaluate the robustness of our proposed framework and compare its performance with state-of-the-art score fusion approaches (Section 5). We also present a comparative analysis of our proposed confidence-based rank-level fusion approach with state-of-the-art rank-level fusion approaches.

In Fig. 1, a typical audio-visual biometric recognition system is shown with all possible fusion approaches including our proposed confidence-based fusion. Our contribution as a whole lies in the shaded box where the matcher confidence values are calculated from the match scores.

2. Fusion In Multiobiometric Identification

Let N denote the number of enrolled users and M denote the number of modalities. If $s_{m,j}$ is the score and $r_{m,j}$ the rank provided for the j-th template by the m-th matcher, $j = 1 \ldots N$; $m = 1 \ldots M$ then for a given query we get $M \times N$ score and rank matrices as follows:

$$S = \begin{pmatrix} s_{1,1} & \cdots & s_{1,N} \\ s_{2,1} & \cdots & s_{2,N} \\ \vdots & \ddots & \vdots \\ s_{M,1} & \cdots & s_{M,N} \end{pmatrix}, \tag{1}$$

and

$$R = \begin{pmatrix} r_{1,1} & \cdots & r_{1,N} \\ r_{2,1} & \cdots & r_{2,N} \\ \vdots & \ddots & \vdots \\ r_{M,1} & \cdots & r_{M,N} \end{pmatrix}. \tag{2}$$

Our objective is to determine the true identity of the given query from S and (or) R. In this section, we briefly discuss the state-of-the-art in late fusion for multibiometric identification.

2.1. Existing Score-level Fusion Approaches

Existing score fusion approaches can be categorized into four groups (Section 1). Here, we briefly discuss all four approaches of score-level fusion for multimodal biometric identification.

2.1.1. Transformation-based Score Fusion

An example of a simple transformation-based score fusion approach is the use of the min-max score normalization to transform the raw scores into [0,1] range and then use the equally weighted sum rule (EWS) of fusion. The min-max normalization is performed as

$$s'_{m,j} = \frac{s_{m,j} - min(S_m)}{max(S_m) - min(S_m)}, \tag{3}$$

where $s_{m,j}$ is the match score provided by the m-th matcher to the j-th identity, S_m is the m-th row in S that corresponds to the matching scores from the m-th matcher. Then, the matching scores from all the matchers are added without any bias to a particular matcher:

$$f_j = \sum_{m=1}^{M} w_m s'_{m,j}, \tag{4}$$

where, f_j represents the fused score for the j-th identity and w_m is the weight assigned to the m-th matcher such that $w_1 = w_2 = \ldots = w_m$.

2.1.2. Density-based Score Fusion

In (Nandakumar et al., 2009) the authors used a likelihood ratio score fusion (Nandakumar et al., 2008) which was originally designed for verification under certain assumptions: **(i)** prior probabilities are equal for all the users, **(ii)** the match scores for different users are independent of one another, and **(iii)** genuine (impostor) match scores of all users are identically distributed. The aim of an identification system is to assign the query an identity I_{j_0} that maximizes the posterior probability. The decision rule for closed set identification is governed by

$$P(I_{j_0}|S) \geq P(I_j|S), \forall j = 1, \ldots, N. \tag{5}$$

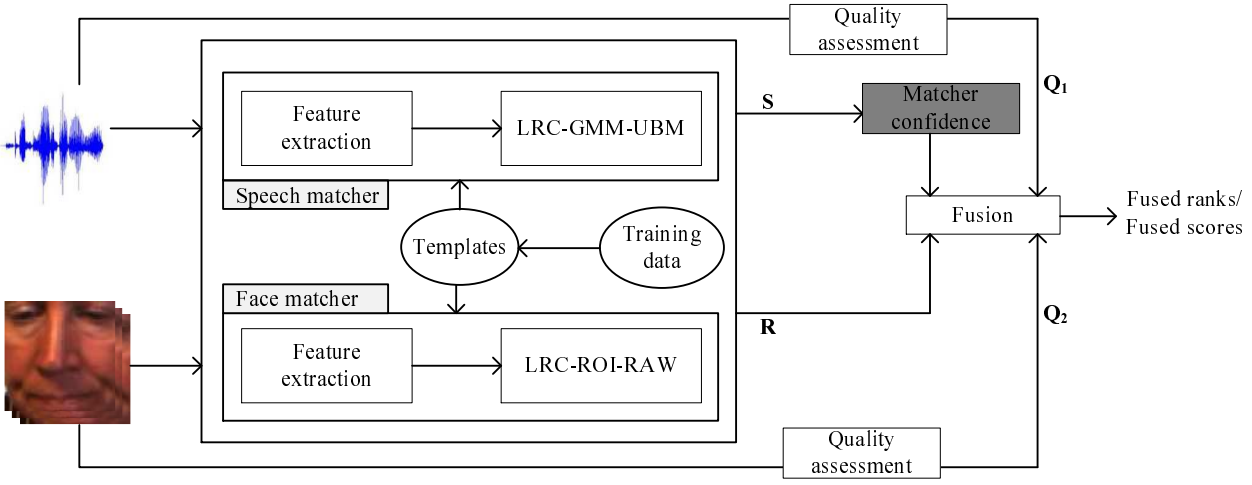


Fig. 1: Block diagram of an audio-visual biometric system that incorporates sample quality and (or) matcher confidence measures in the fusion. Although quality-based fusion has been studied extensively (Poh and Kittler, 2012), incorporating matcher confidence in the fusion has not been well studied (Marasco and Sansone, 2011). Moreover, achieving sample quality in audio-visual biometrics is challenging (Chetty and Wagner, 2008). The shaded box highlights our contribution.

For open set identification, the query is assigned identity I_{j_0} only when $P(I_{j_0}|S) > \tau$ in the above equation. According to Bayes theory (Duda et al., 2012) we can calculate $P(I_j|S)$ as follows:

$$P(I_j|S) = \frac{p(S|I_j)P(I_j)}{p(S)}. \tag{6}$$

Now, under the assumption of equal prior $P(I_j)$ for all users (Nandakumar et al., 2009), the posterior probability $P(I_j|S)$ is proportional to the likelihood $p(S|I_j)$. The likelihood $p(S|I_j)$ can be written as

$$p(S|I_j) = \frac{f_{gen}(s_j)}{f_{imp}(s_j)} \prod_{i=1}^{N} f_{imp}(s_i) \tag{7}$$

where $s_j = [s_{1,j}, \ldots, s_{M,j}]$ is the score vector corresponding to user j from M modalities, and $f_{gen}(s_j)$ and $f_{imp}(s_j)$ are the densities of genuine and impostor match scores, respectively, assuming that they are identically distributed for all users. Thus, the likelihood of observing the score matrix S given the true identity is I_j is proportional to the likelihood ratio for verification used by the authors in (Nandakumar et al., 2008). The authors in (Nandakumar et al., 2009) assumed that the scores from different matchers are conditionally independent. Hence, the joint density of the genuine (impostor) match scores can be estimated as the product of marginal densities, which we refer to as LRT-GMM in this paper:

$$\prod_{m=1}^{M} \frac{f^m_{gen}(s_{m,j_0})}{f^m_{imp}(s_{m,j_0})} \geq \prod_{m=1}^{M} \frac{f^m_{gen}(s_{m,j})}{f^m_{imp}(s_{m,j})}, \forall j = 1, \ldots, N. \tag{8}$$

2.1.3. Quality-based Score Fusion

In (Nandakumar et al., 2008), the authors presented the quality-based likelihood ratio (QLR) fusion technique provided the sample quality information is available. Being inspired by their LRT-GMM method, we can define the QLR framework for identification problem as follows

$$\prod_{m=1}^{M} \frac{f^m_{gen}(s_{m,j_0}, Q_m)}{f^m_{imp}(s_{m,j_0}, Q_m)} \geq \prod_{m=1}^{M} \frac{f^m_{gen}(s_{m,j}, Q_m)}{f^m_{imp}(s_{m,j}, Q_m)}, \forall j = 1, \ldots, N. \tag{9}$$

We use the universal image quality index presented in (Wang and Bovik, 2002) to represent face image quality and the NIST-SNR as in (Kim and Stern, 2008) to represent speech signal quality.

2.2. Existing Rank-Level Fusion Methods

In rank-level fusion, the ranked lists from different matchers are combined using a number of methods, such as the highest rank, Borda count, logistic regression, etc.

2.2.1. Highest Rank Fusion

In the highest rank method (Ho et al., 1994), the combined rank r_j of a user j is calculated by taking the lowest rank (r) assigned to that user by different matchers. One of the shortcomings of the highest rank fusion is that it may produce the same final rank for multiple users. The authors in (Ho et al., 1994) proposed to randomly break ties between different users. On the other hand, in (Abaza and Ross, 2009) a perturbation factor ϵ was introduced to break ties:

$$r_j = \min_{m=1}^{M} r_{m,j} + \epsilon_j, \tag{10}$$

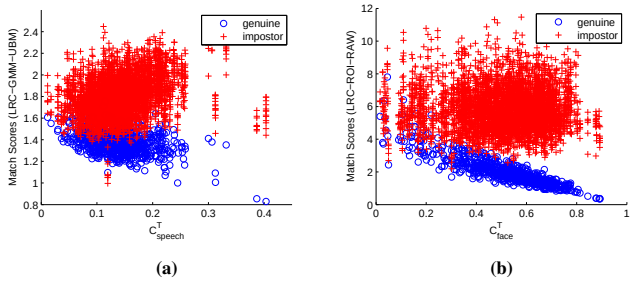
where,

$$\epsilon_j = \frac{\sum_{m=1}^{M} r_{m,j}}{K}. \tag{11}$$

The perturbation factor biases the fused rank by considering all the ranks associated with user j, by assuming a large value for K.

2.2.2. Borda Count Rank Fusion

In Borda count method (Ho et al., 1994), the fused rank is calculated by taking the sum of the ranks produced by individual matchers for a user j. The Borda count method accounts for the variability in ranks due to the use of a large number of matchers. The major disadvantage of this method is that it



(a) (b)

Fig. 2: Variation of match scores with the (a) speech and (b) face matcher confidence measures in the training dataset (T). Our proposed matcher confidence measure is able to separate the genuine scores from the impostor scores. For example, the difference between the genuine and impostor scores is high when our proposed matcher confidence measure is high and vice versa. The maximum value of C^T_{face} is 0.899 whereas the maximum value of C^T_{speech} is 0.403. We calculate the *C-ratio* of a modality by normalizing the matcher confidence obtained at the evaluation phase (c^E_m) by the maximum value of corresponding C^T_m.

assumes all the matchers are statistically independent and perform equally well. In practice, a particular matcher may perform poorly due to various reasons, such as the quality of the probe data, quality of the templates in gallery etc. In (Abaza and Ross, 2009), a method which is also known as the Nanson function (Fishburn, 1990), was used to eliminate the worst rank for a user:

$$\max_{m=1}^{M} r_{m,j} = 0. \tag{12}$$

This can be extended by eliminating the lowest rank k times before applying the Borda count on remaining ranks.

Another quality-based approach was proposed in the same paper (Abaza and Ross, 2009) with the inclusion of an input image quality in Borda count method as follows:

$$r_j = \sum_{m=1}^{M} Q_{m,j}.r_{m,j}, \tag{13}$$

where, $Q_{m,j} = min(Q_m, Q_j)$, and Q_m and Q_j are the quality factors of the probe and gallery fingerprint impressions, respectively. A predictor-based approach was proposed in (Marasco et al., 2010) which calculates the final rank for each user as the weighted sum of individual ranks assigned by M matcher. A higher weight was assigned to the ranks provided by the more accurate matcher:

$$r_j = \sum_{m=1}^{M} w_m.r_{m,j}, \tag{14}$$

where, w_m is the assigned weight for matcher m. An additional training phase was used for determining the weights. In (Kumar and Shekhar, 2011) a non-linear approach of rank-level fusion was proposed for palm-print recognition. On the other hand, in (Monwar and Gavrilova, 2009) the ranks of only those identities were fused which appear in at least two classifiers (face, ear and signature).

3. Proposed Fusion Framework

3.1. C-ratio Score Fusion

It is a well known fact that the difference between the genuine and impostor match scores is usually high under normal circumstances (e.g., clean conditions). In practice, noisy samples may be presented to the system and therefore the decision making task may become difficult for the matchers. We propose to set the confidence of a matcher as the normalized difference between the best match score and the mean of the k subsequent match scores.

We obtain a two column matrix S^- by first sorting the score matrix S and then keeping the best matching score (i.e., column 1) and the mean of k subsequent matching scores:

$$S^- = \begin{pmatrix} s^1_1 & \mu_1 \\ s^1_2 & \mu_2 \\ \vdots & \vdots \\ s^1_M & \mu_M \end{pmatrix}, \tag{15}$$

where,

$$\mu_m = \frac{1}{k-1} \sum_{n=2}^{k} s^n_m. \tag{16}$$

Then, the matcher confidence for modality m is calculated as

$$c_m = \frac{|s^1_m - \mu_m|}{\mu_m}. \tag{17}$$

Here, a higher value of c_m refers to a strong classification (i.e., clean probe data), and a smaller value of c_m refers to a weak classification (see Fig. 2).

We propose a novel confidence-ratio (*C-ratio*) for a matcher m as follows:

$$\gamma_m = c^E_m / max(C^T_m), \tag{18}$$

where c^E_m is the matcher confidence for modality m obtained at the evaluation (E) phase and $max(C^T_m)$ is the maximum matcher confidence for modality m from the training (T) phase. This approach requires that the most likely identity is assigned the lowest matching score and other identities get higher scores (e.g., Euclidean distance). If the match scores follow opposite trend (e.g., likelihoods or probability) they must be inverted before our proposed matcher confidence and *C-ratio* can be applied. We then transform the matching score matrix (S) and perform fusion as follows:

$$f = S^T \mathrm{x} \tag{19}$$

where, $\mathrm{x} = [\gamma_1 \dots \gamma_M]^T$ is the transformation vector containing the *C-ratio* of all the matchers and $f = [f_1 \dots f_N]^T$ is the fused score vector. The decision is ruled in favor of the template achieving highest fused score.

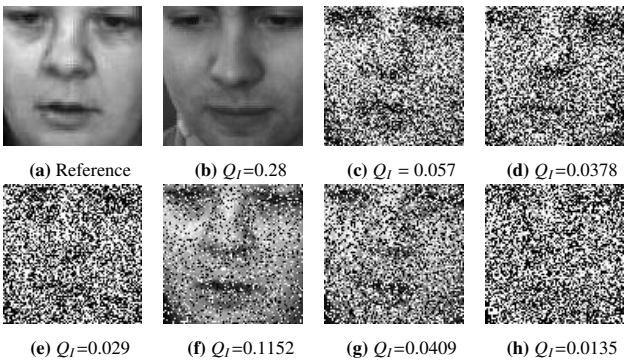
3.2. Confidence-Based Rank-Level Fusion

In this section, we discuss the confidence-based approach in rank-level fusion. We propose to use a novel *confidence factor* to be used with the highest rank fusion rule. We also propose a modification to the Borda count rank fusion.

3.2.1. Confidence-Based Highest Rank Fusion

The confidence measures obtained by Eq. (17) can be consolidated into a confidence-based highest rank fusion rule as follows:

$$r_j = \min_{m=1}^{M} r_{m,j} + \eta_j, \tag{20}$$



(a) Reference **(b)** Q_I=0.28 **(c)** Q_I = 0.057 **(d)** Q_I=0.0378

(e) Q_I=0.029 **(f)** Q_I=0.1152 **(g)** Q_I=0.0409 **(h)** Q_I=0.0135

Fig. 3: Universal image quality index (Q_I) for a reference image (a) matched against clean input image (b) from a different user, and against the same image (a) corrupted with AWGN of levels of (c) σ^2 = 0.3 (d) σ^2 = 0.3 and (e) σ^2 = 0.9 as well as (f) 25% (g) 50% and (h) 75% salt and pepper noise.

where the term η_j is the *confidence factor* which can be calculated as follows:

$$\eta_j = \frac{\sum_{m=1}^{M} max(C_m^T).r_{m,j}}{\sum_{m=1}^{M} r_{m,j}}, \tag{21}$$

We use the novel *confidence factor* (η_j) so that the ranks produced by a more confident classifier get more emphasis. The denominator in Eq. (21) transforms the *confidence factor* for a user (j) into the range [0, 1].

Here, we analytically show how the use of a *confidence factor* η_j can handle ties in the highest rank fusion better than the modified highest rank fusion rule in Eq. (10). Let the ranks for a user (j = 1) from two matchers of a mutibiometric system be $r_{1,1}$ = 1 and $r_{2,1}$ = 2, while for another user (j = 2), let $r_{1,2}$ = 2 and $r_{2,2}$ = 1. By the modified highest rank fusion in Eq. (10), we obtain r_1 = 1.03 and r_2 = 1.03, when K = 100 as in (Abaza and Ross, 2009). On the other hand, let the confidence measure $max(C_1^T)$ for a matcher be 0.3 and $max(C_2^T)$ for another matcher be 0.9. By using Eq. (21), we get $r_1 = 1 + \frac{(0.3\times1)+(0.9\times2)}{(1+2)} = 1.7$ and $r_2 = 1 + \frac{(0.3\times2)+(0.9\times1)}{(1+2)} = 1.5$. Thus, not only a tie between the final ranks of the users j = 1 and j = 2 is avoided but also the ranking of the more confident classifier is emphasized.

3.2.2. Confidence-Based Borda Count Fusion

We propose to modify the Borda count method as follows:

$$r_j = \sum_{m=1}^{M} max(C_m^T).r_{m,j}. \tag{22}$$

The proposed confidence-based Borda count fusion rule is indeed the numerator of Eq. (21) and similar to the quality based Borda count fusion in (Abaza and Ross, 2009). Here, instead of quality measures for the probe data, we propose to use confidence measures for the classifiers.

4. Databases And Systems

4.1. AusTalk

In our experiments, we used a new audio-visual database, namely the AusTalk (Burnham et al., 2011). The AusTalk is a large collection of audio-visual data captured at several university campuses across Australia. We used audio-visual data from 248 individuals that consists of 4-digits utterances recorded in twelve sessions. We used the data in the first six sessions for enrollment of speaker models/templates and the data in the remaining six sessions as probes.We randomly selected half the users to be in the training set (T) and the remaining half in the evaluation set (E). This process was repeated five times and therefore the recognition performances reported in this paper on the AusTalk are the averages of five test runs. We needed the training set (T) to estimate the genuine/impostor score densities to implement LRT-GMM and QLR. We used the GMM fitting algorithm presented in (Figueiredo and Jain, 2002) for density estimation. Since the AusTalk database consists of clean speech and videos recorded in room environment, we degraded the data using additive white Gaussian noise (AWGN) and salt and pepper noise. In Fig. 3(b), a value of the universal image quality index is shown for the face image from a person when it is matched with a reference image in Fig. 3(a). In Fig. 3(c-h), the values of the universal image quality are shown when the reference image is compared with itself, but corrupted at different levels of AWGN and salt and pepper noise.

4.2. VidTIMIT

We also used the VidTIMIT database (Sanderson and Lovell, 2009) to evaluate the performance of our prosed fusion framework. It comprises audio-visual data from 43 persons (19 females and 24 males) reciting short sentences in 3 sessions. There are 10 sentences per person, with the first six sentences captured in Session 1, the next two sentences in Session 2 and the remaining two in Session 3. In our experiments, we used Session 1 and Session 2 data for the enrollment of speaker models/templates and Session 3 data as probes. We randomly selected 21 speakers (9 females and 12 males) to be in the training set (T) and the remaining 22 speakers in the evaluation set (E). This process was repeated five times; therefore, the recognition performances reported in this paper on the VidTIMIT database are the averages of five test runs. We used the same mechanisms for density estimation and data degradation as described in Section 4.1.

4.3. System

We used the LRC-GMM-UBM and LRC-ROI-RAW frameworks that we previously used in our works in (Alam et al., 2013) and (Alam et al., 2014) as the matchers of the audio and visual modalities, respectively. The main concept is that the samples from a specific user lie on a linear subspace, and therefore the task of person identification is considered to be a linear regression problem (Naseem et al., 2010). In the LRC-GMM-UBM, a Universal Background Model (UBM) is trained using the MFCCs extracted from the enrollment data of all the speakers in the training set (T). Then, the enrollment data from an

Table 1: Rank-1 identification at various levels of additive white Gaussian noise on speech and face probes in AusTalk

Noise Level (SNR,σ^2)	Speech (%)	Face (%)	LRT-GMM (%)	QLR (%)	EWS (%)	C-ratio (%)
(clean, 0.3)	99.02	86.81	95.99	98.25	99.6	99.19
(clean, 0.6)	99.02	48.27	94.32	98.17	97.23	96.43
(clean, 0.9)	99.02	26.87	93.38	97.92	95.83	94.35
(30dB, clean)	75.56	97.95	98.95	98.95	98.11	98.84
(20dB, clean)	30.99	97.95	95.69	96.68	89.22	98.41
(10dB, clean)	3.93	97.95	89.40	92.01	65.88	98.09
(30dB, 0.3)	75.56	86.81	75.15	81.18	91.56	96.39
(30dB, 0.6)	75.56	48.27	68.12	74.84	82.24	83.19
(20dB, 0.3)	30.99	86.81	34.11	42.71	64.15	89.50
(20dB, 0.6)	30.99	48.27	27.82	33.54	51.96	58.51
(20dB, 0.9)	30.99	26.87	27.17	32.76	46.15	51.55
Average	59.23	68.43	72.73	77	80.17	**87.67**

Table 2: Rank-1 identification at various levels of additive white Gaussian noise on speech and face probes in VidTIMIT

Noise Level (SNR,σ^2)	Speech (%)	Face (%)	LRT-GMM (%)	QLR (%)	EWS (%)	C-ratio (%)
(clean, 0.3)	80.90	73.36	88.18	88.63	91.67	90.45
(clean, 0.6)	80.90	61.81	81.36	83.63	89.39	90.00
(clean, 0.9)	80.90	49.99	73.18	79.54	84.09	88.18
(30dB, clean)	77.95	77.81	86.36	86.80	92.42	87.27
(20dB, clean)	58.86	77.81	83.63	84.09	87.88	85.45
(10dB, clean)	26.36	77.81	69.99	71.81	69.70	81.36
(30dB, 0.3)	77.95	71.36	88.18	87.72	90.15	89.09
(30dB, 0.6)	77.95	61.81	80.90	81.36	85.61	87.73
(20dB, 0.3)	58.86	71.36	83.18	83.63	85.61	85.45
(20dB, 0.6)	58.86	61.81	70.00	76.36	81.06	84.09
(20dB, 0.9)	58.86	49.99	56.00	68.63	75.76	79.09
Average	67.12	66.81	78.26	81.11	84.84	**86.19**

Table 3: Rank-1 identification at various levels of additive white Gaussian noise on speech and salt and pepper noise on face probes in AusTalk

Noise Level (SNR,σ^2)	Speech (%)	Face (%)	LRT-GMM (%)	QLR (%)	EWS (%)	C-ratio (%)
(clean, 0.25)	99.02	96.98	99.35	99.65	99.81	99.33
(clean, 0.50)	99.02	57.06	95.61	98.65	98.36	97.39
(clean, 0.75)	99.02	6.16	93.46	97.68	91.07	89.49
(30dB, 0.25)	75.56	96.98	92.18	95.61	95.83	99.06
(30dB, 0.50)	75.56	57.06	71.85	77.39	84.54	86.18
(20dB, 0.25)	30.99	96.98	65.88	77.87	76.96	98.06
(20dB, 0.50)	30.99	57.06	30.43	35.32	55.24	69.78
(20dB, 0.75)	30.99	6.16	26.72	30.83	32.51	33.52
Average	67.64	59.30	71.93	76.62	79.29	**84.10**

individual speaker is used to adapt a Gaussian Mixture Model (GMM) from the UBM. An adapted GMM from the UBM is also commonly known as the GMM-UBM speaker model. Finally, the means from all the GMM-UBMs are concatenated to form a supervector. Speaker-specific templates are created stacking all the feature vectors (GMM-UBM means) from the enrollment data. Similarly, in the LRC-ROI-RAW framework, user-specific templates are created by stacking the feature vectors obtained from down-sampled raw face images. In the test phase, a feature vector is first extracted from the probe data and a response vector is then predicted as a linear combination of the templates of each speaker stored in the gallery. Finally, the Euclidean distance between the test and a predicted response vector is used as a matching score.

5. Experiments, Results and Analysis

In this section, we present experimental results on the AusTalk and the VidTIMIT databases. We evaluated the robustness of our proposed fusion framework considering additive white Gaussian noise (AWGN) and salt and pepper noise on the face images as well as AWGN in the speech samples. We compared the performance of our fusion framework with the LRT-GMM, QLR and min-max normalized equal weighted sum (EWS) methods for score-level fusion.

Then, we tested the robustness of our proposed framework for rank-level fusion on the AusTalk database with AWGN only. We compared the performance of our proposed confidence-based rank-level fusion (conBordaCount and conHighestRank) with the Borda count (bordaCount) and the highest rank (highestRank) fusion as well as the perturbation factor based highest rank (pFactorHighestRank) and the predictor based Borda count (predictorBasedBorda) methods of rank-level fusion. The weights w_m in Eq. (14) were computed using the probe data in the training set (T). The ratio between correct identification and the total number of probes (Marasco et al., 2010) as determined by the matchers were used as weights. In our predictor-based experiments, the audio sub-system weight $w_1 = 0.98$ and the visual sub-system weight $w_2 = 0.97$.

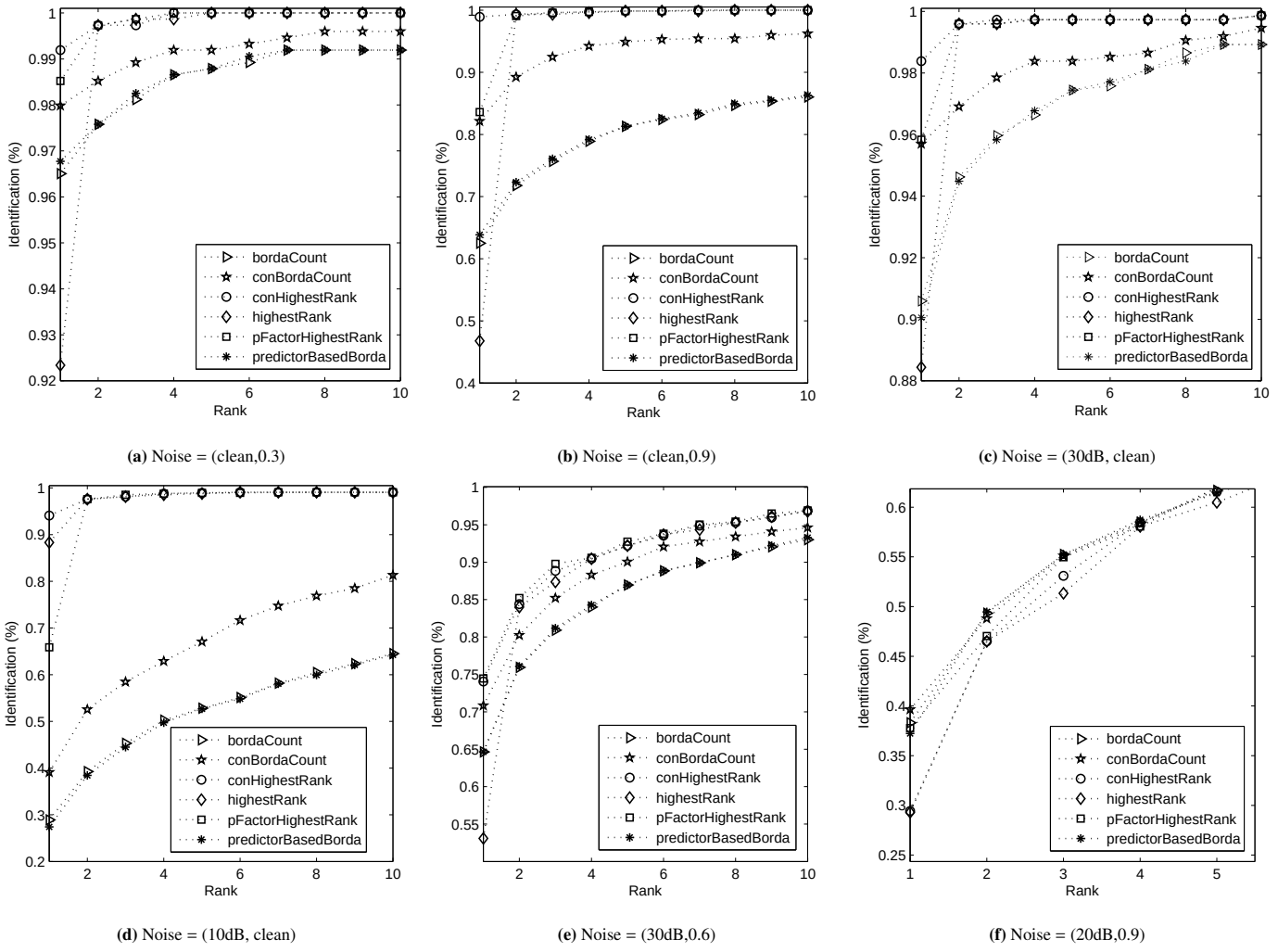
5.1. Robustness to AWGN

The additive white Gaussian noise is always an important case-study in the context of robustness because it models the detector noise of the imaging system (Nakamura, 2005) . The input face images were distorted by adding zero-mean Gaussian noise with three different error variances (see Fig. 3(c-e)). The speech samples were distorted by adding white noise at three different SNR levels. In Table (1-2), the rank-1 identification accuracies for the LRT-GMM in Eq. (8), QLR in Eq. (9), and our proposed confidence-based (C-ratio) score fusion in Eq. (19) as well as the EWS in Eq. (4) with min-max score normalization are listed for the AusTalk and the VidTIMIT databases, respectively.

On the AusTalk database, our proposed *C-ratio* score fusion outperforms the state-of-the-art *density-based*, *quality-based* and *transformation-based* fusion techniques, particularly when probes from both modalities are degraded due to AWGN. For example, in Table 1, when the speech signal is corrupted with 30dB and face images with $\sigma^2 = 0.3$ of AWGN, our proposed C-ratio score fusion achieves a rank-1 recognition accuracy of 96.39%, which is significantly higher than the state-of-the-art in score fusion. We achieved 75.15%, 81.18% and 91.56% rank-1 recognition rates with the LRT-GMM-UBM, QLR and EWS fusion methods, respectively and at the same noise level. The overall rank-1 identification accuracy using our proposed *C-ratio* score fusion on the AusTalk is 87.67% which is also at least 7.5% higher than any other method. Similarly, our proposed method outperforms state-of-the-art score fusion techniques on the VidTIMIT database. The overall rank-1 identification accuracy (Table 2) obtained using our proposed C-ratio score fusion is 86.19% which is slightly better than the rank-1 identification accuracy obtained using the EWS method and at least 5% higher than the accuracies achieved using the LRT-GMM and QLR methods.

5.2. Robustness to Salt and Pepper Noise

In the next set of experiments, we tested the robustness of our proposed confidence-based score fusion by considering in-



(a) Noise = (clean,0.3) **(b)** Noise = (clean,0.9) **(c)** Noise = (30dB, clean)

(d) Noise = (10dB, clean) **(e)** Noise = (30dB,0.6) **(f)** Noise = (20dB,0.9)

Fig. 4: CMC curves for our confidence-based rank fusion method (conBordaCount and conHighestRank) at different (audio,visual) noise levels and compared against the Borda count (bordaCount), the highest rank (higestRank), perturbation-factor highest rank (pFactorHighestRank) and the predictor based Borda count (predictorBasedBorda) methods.

Table 4: Rank-1 identification at various levels of additive white Gaussian noise on speech and salt and pepper noise on face probes in VidTIMIT

Noise Level (SNR,σ^2)	Speech (%)	Face (%)	LRT-GMM (%)	QLR (%)	EWS (%)	C-ratio (%)
(clean, 0.25)	80.90	76.13	88.18	89.54	90.91	89.55
(clean, 0.50)	80.90	65.90	82.73	84.54	87.88	90.00
(clean, 0.75)	80.90	29.31	64.54	69.54	75.00	82.73
(30dB, 0.25)	77.95	76.13	87.72	89.09	91.67	86.82
(30dB, 0.50)	77.95	65.90	81.36	83.63	85.61	88.64
(20dB, 0.25)	58.86	76.13	85.00	85.00	87.12	85.45
(20dB, 0.50)	58.86	65.90	75.91	79.54	82.58	83.64
(20dB, 0.75)	58.86	29.31	43.18	55.90	59.85	65.91
Average	71.89	60.58	76.07	79.59	82.57	**84.09**

put probes contaminated with data *drop-out* and *snow in the image* simultaneously, usually referred to as salt and pepper noise (Gonzalez et al., 2009). This type of noise can be caused by analog-to-digital converter errors and bit errors in transmission (Naseem et al., 2012).

In Table 3, a summary of rank-1 identification accuracies on the AusTalk databse has been presented. It shows that the proposed *C-ratio* score fusion also performs better than the state-of-the-art score fusion methods when salt and pepper noise was considered for face images and AWGN on speech signals. For example, when the speech signal is corrupted with 30dB SNR and 25% of the pixels on the face image are assumed to be contaminated, our proposed *C-ratio* fusion achieves 99.06% rank-1 accuracy. The overall rank-1 identification accuracy using the proposed *C-ratio* score fusion is at least 4.8% higher than any other method. Similarly, our proposed C-ratio score fusion outperforms state-of-the-art on the VidTIMIT database. The overall rank-1 identification accuracy (Table 4) using our proposed C-ratio score fusion on VidTIMIT is 84.09% which is at least 1.5% higher than the accuracies obtained using the EWS, LRT-GMM, and QLR fusion methods.

5.3. Rank-level Fusion With AWGN

We also performed experiments on rank-level fusion under various levels of audio and visual degradations on the AusTalk database. In Fig. 4(a-f), the Cumulative Match Characteristics (CMC) curves for different (audio,visual) noise levels are shown. Our proposed confidence-based rank fusion approach achieved better rank-1 identification rates than the state-of-the-art highest rank fusion approaches. For example, in Fig.

4(a) the CMC curve for clean speech and slightly corrupted face images is shown. Our proposed confidence-based highest rank fusion achieved 99% rank-1 identification accuracy that is 2.5% higher than the conventional highest rank fusion (highestRank) and slightly higher than the modified highest rank (pFactorHighestRank) approach in (Abaza and Ross, 2009). Fig. 4(b-d) shows the CMC curve at other (audio,visual) noise levels. In all settings, the rank-1 recognition rate obtained using the confidence-based highest rank fusion was higher than the highest rank (highestRank) and the modified highest rank (pFactorHighestRank) . In Fig. 4(e-f), we show the CMC curves of our proposed rank fusion approach considering that data from both modalities are degraded. Although the confidence-based rank fusion approach outperforms the conventional highest rank fusion (highestRank) on both the occasions, its performance is almost equal at (30dB,0.6) noise level and slightly worse at (20dB,0.9) noise level when compared with the modified highest rank fusion (pFactorHighestRank) approach.

On the other hand, the performance improvement by using the confidence-based Borda count method was higher for all rank levels (rank-1 to rank-10). Therefore, the confidence-based rank-level fusion clearly improves the recognition accuracy. Another interesting observation is that the predictor-based Borda count method (Marasco et al., 2010) does not improve recognition performance if there is noise on probe data because the predictor-based method uses fixed weights for the matchers.

6. Conclusions

We have presented a confidence-based late fusion framework and its application to audio-visual biometrics. We showed that matcher confidence can be calculated from the output match scores and a novel *C-ratio* can be calculated for transforming the match scores before they are fused. We have also proposed a novel *confidence-factor* that can successfully break the ties in the highest rank fusion. Finally, experimental results have been presented which give us a clear indication that confidence-based fusion can be considered as a robust and accurate fusion method for biometric systems operating in the identification mode.

Acknowledgment

This research is partially supported by Australian Research Council grants DP110103336 and DE120102960.

References

Abaza, A., Ross, A., 2009. Quality based rank-level fusion in multibiometric systems, in: Biometrics: Theory, Applications, and Systems, 2009. BTAS'09. IEEE 3rd International Conference on, IEEE. pp. 1–6.

Alam, M.R., Bennamoun, M., Togneri, R., Sohel, F., 2014. Confidence-based rank-level fusion for audio-visual person identification system. Pattern Recognition Applications and Methods, 2014. 3rd International Conference on , pp. 608–615.

Alam, M.R., Togneri, R., Sohel, F., Bennamoun, M., Naseem, I., 2013. Linear regression-based classifier for audio visual person identification, in: Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on, IEEE. pp. 1–5.

Burnham, D., Estival, D., Fazio, S., Viethen, J., Cox, F., Dale, R., Cassidy, S., Epps, J., Togneri, R., Wagner, M., et al., 2011. Building an audio-visual corpus of australian english: large corpus collection with an economical portable and replicable black box, in: Twelfth Annual Conference of the International Speech Communication Association.

Chetty, G., Wagner, M., 2008. Robust face-voice based speaker identity verification using multilevel fusion. Image and Vision Computing 26, 1249–1260.

Duda, R.O., Hart, P.E., Stork, D.G., 2012. Pattern classification. John Wiley & Sons.

Figueiredo, M.A.T., Jain, A.K., 2002. Unsupervised learning of finite mixture models. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24, 381–396.

Fishburn, P., 1990. A note on a note on nanson's rule. Public Choice 64, 101–102.

Gonzalez, R.C., Woods, R.E., Eddins, S.L., 2009. Digital image processing using MATLAB. volume 2. Gatesmark Publishing Knoxville.

Ho, T.K., Hull, J.J., Srihari, S.N., 1994. Decision combination in multiple classifier systems. Pattern Analysis and Machine Intelligence, IEEE Transactions on 16, 66–75.

Kim, C., Stern, R.M., 2008. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis., in: INTERSPEECH, pp. 2598–2601.

Kittler, J., Hatef, M., Duin, R., Matas, J., 1998. On combining classifiers. Pattern Analysis and Machine Intelligence, IEEE Transactions on 20, 226–239.

Kumar, A., Shekhar, S., 2011. Personal identification using multibiometrics rank-level fusion. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 41, 743–752.

Marasco, E., Ross, A., Sansone, C., 2010. Predicting identification errors in a multibiometric system based on ranks and scores, in: Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on, IEEE. pp. 1–6.

Marasco, E., Sansone, C., 2011. An experimental comparison of different methods for combining biometric identification systems, in: Image Analysis and Processing–ICIAP 2011. Springer, pp. 255–264.

Monwar, M.M., Gavrilova, M.L., 2009. Multimodal biometric system using rank-level fusion approach. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 39, 867–878.

Murakami, T., Takahashi, K., 2009. Accuracy improvement with high convenience in biometric identification using multihypothesis sequential probability ratio test, in: Information Forensics and Security, 2009. WIFS 2009. First IEEE International Workshop on, IEEE. pp. 66–70.

Nakamura, J., 2005. Image sensors and signal processing for digital still cameras. CRC Press.

Nandakumar, K., Chen, Y., Dass, S.C., Jain, A.K., 2008. Likelihood ratio-based biometric score fusion. Pattern Analysis and Machine Intelligence, IEEE Transactions on 30, 342–347.

Nandakumar, K., Jain, A.K., Ross, A., 2009. Fusion in multibiometric identification systems: What about the missing data?, in: Advances in Biometrics. Springer, pp. 743–752.

Naseem, I., Togneri, R., Bennamoun, M., 2010. Linear regression for face recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32, 2106–2112.

Naseem, I., Togneri, R., Bennamoun, M., 2012. Robust regression for face recognition. Pattern Recognition 45, 104–118.

Poh, N., Kittler, J., 2012. A unified framework for biometric expert fusion incorporating quality measures. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34, 3–18.

Ross, A., Nandakumar, K., Jain, A., 2006. Handbook of multibiometrics. volume 6. Springer.

Sanderson, C., Lovell, B.C., 2009. Multi-region probabilistic histograms for robust and scalable identity inference, in: Advances in Biometrics. Springer, pp. 199–208.

Sanderson, C., Paliwal, K., 2002. Information fusion and person verification using speech & face information .

Tao, Q., Veldhuis, R., 2008. Hybrid fusion for biometrics: Combining score-level and decision-level fusion, in: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on, IEEE. pp. 1–6.

Wang, Z., Bovik, A.C., 2002. A universal image quality index. Signal Processing Letters, IEEE 9, 81–84.