International Journal of Control and Automation
Vol. 8, No. 4 (2015), pp. 313-332
http://dx.doi.org/10.14257/ijca.2015.8.4.30

Audio-Visual Person Recognition Based on Rank Level Fusion and Gaussian Mixture Model

Di Wu

[1]*College of Electrical and Information Engineering ,Hunan Institute of Engineering , Xiangtan, China*
wudi6152007@163.com

Abstract

In this paper, A new multimodal system for reliable person identification based on rank level fusion and Gaussian Mixture Model is proposed. Our system fuses the information of face and speech under variant conditions. In logistic regression rank level fusion method, its often difficult to compute the fusion weight properly, so a estimate method based on estimating probability density functions under Gaussian Mixture Model is proposed in this paper. This method can achieve a robust estimator of the optimal weighting parameter which can enhance the rank level fusion performance. The experiments on the AMI database indicate the proposed method can enhance the accuracy of the recognition system effectively.

Keywords*: MultiBiometrics, Rank Level Fusion, Gaussian Mixture Model, Face Recognition, Speech Recognition*

1. Introduction

There is an increasing interest in biometric person recognition for commercial, surveillance, security and other applications, but recognition based on only one modality is unlikely to achieve acceptable performance for practical deployment [1]. A potential method to overcome this drawback is to combine information from more than one modality which called multibiometric. Several important studies have confirmed this potential. In this paper, we consider the particular situation of audio-visual person recognition where there are just two source of information: an audio signal (speech) and a video signal (face).

Multibiometrics is a relatively new approach to biometric knowledge representation that strives to overcome the problems by consolidating the evidence presented by multiple biometric sources [2]. Multibiometric systems can significantly improve the recognition performance in addition to deterring spoof attacks, increasing the degree of freedom and also reducing the failure to enroll rate. Although the storage requirements, processing time, and computational demands of a multibiometric system can be higher than a unimodal biometric system, but the aforementioned advantages present a compelling case for deploying multibiometric systems in real world large scale recognition systems.

The key to successful multibiometric system is an effective fusion scheme, which is necessary to combine the information presented by multiple domain experts [3]. The aim of fusion is to determine the best set of experts in a given problem domain and devise an appropriate function that can optimally combine the decisions rendered by the individual expert. Pieces of evidence in multibiometrics systems can be integrated in several levels, and we can subdivide them in five categories [4]: sensor level fusion, feature level fusion, matching level fusion, rank level fusion and decision level fusion.

For fusion to achieve the claimed performance enhancement, fusion rules must be chosen based on the type of application, biometric traits and level of fusion. Among all of the afore-mentioned fusion approaches, fusion at the sensor, match score, feature, and

ISSN: 2005-4297 IJCA

decision levels have been extensively studied in the literature [5]. Biometric systems that integrate information at an early stage of processing are believed to be more effective than those systems which perform integration at a later stage. Sensor-level fusion addresses the problem of noisy sensor data [6], but all other potential problems associated with unimodal biometric systems remain. Since integration at the feature level should provide better recognition results than other levels of integration [7]. However, integration at the feature level is difficult to achieve in practice due to the unknown relationship between the feature spaces of different biometric systems, because of dimensionality problem—the concatenated feature vector with a very large dimensionality, and the inaccessibility of the feature vectors of most commercial biometric systems. Fusion at the decision level is too rigid since only a limited amount of information is available at this level [8]. Therefore, integration at the matching score level is generally preferred due to the ease in accessing and combining matching scores [9]. However, computing a single matching score from the scores of different modalities is required for fusion at this level. Since the matching scores generated by different modalities are heterogeneous, a process called normalization is required to transform these scores into a common domain before combining them. Normalization is computationally expensive, and choosing inappropriate normalization technique can result in a very low recognition performance rate. Also, extra time is needed for this purpose.

Fusion at the rank level [10], however, is a new and significantly understudied problem, which has a high potential for efficient consolidation of multiple unimodal biometric matcher outputs. In a general way, the ranks of individual matchers are combined using the Borda count method, and the logistic regression method. For the logistic regression method of rank level fusion scheme, it important to compute an optimal weight coefficient, although various methods are proposed to choose weighting parameters, there is not unanimous agreement on how to do this.

In the work, we propose a new means to choose the weighting parameter for audio–visual person identification which is based on estimating the probability density functions (pdfs) for the classifier scores. We estimating the probability density functions based on Gaussian Mixture Model. To the best of our knowledge, this is the first time that the Gaussian Mixture Model is used to estimating the optimal weight for rank level fusion scheme, which can produce higher and more reliable recognition results.

The remainder of this paper is organized as follows: The rank level fusion method is discussed in Section 2. The proposed weight estimated method which based on Gaussian Mixture Model is described in Section 3. In Section 4, the construction of the speaker and face classifiers is briefly discussed. Hence, the experiments result is presented in Section5. Section 6 conclude our work and outlines our future work to generalize the method to audio-visual person recognition.

2. Rank Level Fusion

Rank Level fusion is a relatively new fusion approach and is not a well studied research problems, it is used in recognition system and is applicable when the individual matcher's output a ranking of the "candidate" in the template database. The goal of rank level fusion method is to consolidate the rank output by individual biometric matchers in order to derive a consensus rank for each identity. Three method described by Ross et al [11] for making the final decision in a general multiple classifier system, can be used for rank level fusion in multimodal biometric systems. These three methods are highest rank, Borda count and logistic regression methods. All of these methods are briefly discussed in the next subsections.

2.1 Highest Rank Method

The highest rank method is good for combining a small number of specialized matchers and hence can be effectively used for a multimodal biometric system where the individual matchers are the best [12]. In this method, the consensus ranking is obtained by sorting the identities according to their highest rank.

Suppose that we have m classifiers which assign ranks to all classes, and then the consensus rank of a particular class R is obtained by:

$$R = \min_{i=1}^{m} R_i \qquad (1)$$

The final identity recognition ranking is then obtained by sorting the consensus ranking of each class in the ascending order. The advantage of this method is the ability to utilize the strength of each matcher. Even if there is only one matcher assigns the highest ranks to the correct users. The disadvantage of this method is that the final ranking may have many ties, the number if classes sharing the same ranks depending on the number of classifiers used. So the highest rank method can not be a good choice for a security critical multimodal biometric recognition system.

2.2 Borda Count Method

The Borda Count method is the most widely used rank aggregation method and uses the sum of the ranks assigned by individual matchers to calculate the final rank [13]. Suppose we have m classifiers, then the consensus rank of a particular class is obtained by:

$$R = \sum_{i=1}^{m} R_i \qquad (2)$$

The final identity recognition ranking is then obtained by sorting the consensus ranking of each class in the ascending order. This method assumes that the ranks assigned to the users by the individual matchers are statistically and the performances of all matchers are equal. The advantages of this method is that it is easy to implement and require no training stage, so the Borda Count method is feasible to incorporate in multimodal biometric systems. Because of it does not take into account the differences in the individual matcher's capabilities and assumes that all the matchers perform equally, which in usually no the case in most real biometric systems.

2.3 Logistic Regression Method

The logistic regression method [14], which is a variation of the Borda Count method, calculate the weighted sum of the individual ranks. In this method, the final consensus rank is obtained by sorting the identities according to the summation of their rankings obtained from individual matchers multiplied by the assigned weight.

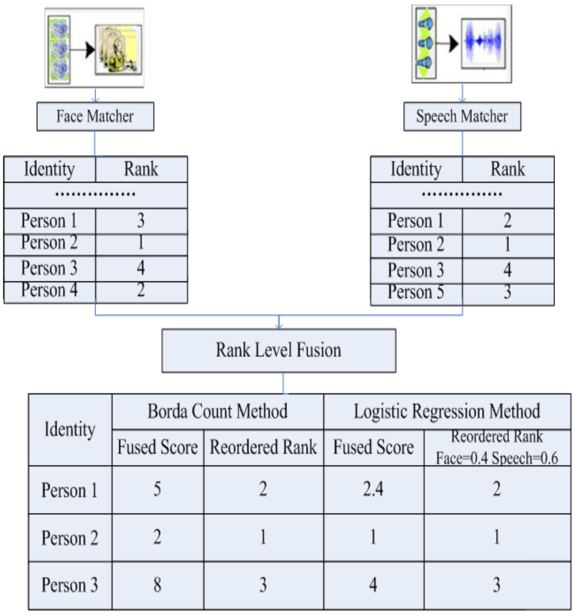
Suppose we have m classifier which assign ranks to all classes, then the consensus ranks of a particular class is obtained by:

$$R = \sum_{i=1}^{m} \omega_i R_i \qquad (3)$$

Where ω_i is the weight assigned to the i-th classifier. The final identity recognition ranking is then obtained by sorting the consensus ranking of each class in the ascending order.

Figure 1 show an example of the Borda Count method and the logistic regression method of rank level fusion. The less value of the rank, the more accurate the result.

Here, the ranks for "Person 1" are 3, 2 from the face and speech matchers respectively. Thus, for the Borda Count method, these ranks are added so we get 5 here, which is the second rank, while "Person 2 "get 2 here. For the logistic regression method, we have assigned 0.4 and 0.6 as the weights for face and speech respectively. The more the weight, the less the performance that means the speech matcher gives us less accurate than face matcher. These weights are chosen by reviewing the previous results obtained by different researchers and also by consequently executing the system. The identities which appear in the result of only one matcher have been discard or not be considered for the final rank in this system , the "Person 4" and "Person 5" appears only once in the matcher result, so it is not be considered in the final result.



Identity	Rank
....	
Person 1	3
Person 2	1
Person 3	4
Person 4	2

Identity	Rank
....	
Person 1	2
Person 2	1
Person 3	4
Person 5	3

Identity	Borda Count Method		Logistic Regression Method	
	Fused Score	Reordered Rank	Fused Score	Reordered Rank Face=0.4 Speech=0.6
Person 1	5	2	2.4	2
Person 2	2	1	1	1
Person 3	8	3	4	3

Figure 1. Example of Rank Level Fusion

The weight to be assigned to the different matchers is determined by a "logit" function using logistic regression, this method is very useful when the different matchers have significant difference in their accuracies but requires a training stage to determine the weights which can be computationally expensive. Inappropriate weight function can eventually reduce the recognition performance of this multimodal biometric system using logistic regression method compared to unimodal system. So, a new and novel weight estimate method based on Gaussian Mixture Model is given in next section.

3. The Proposed Weight Estimate Method

3.1 Theoretical Preparation

Some well known simple fixed rules for combining the matcher scores such as product rule, sum rule, maximum rule etc are described by Kittler[15]. These fixed rules can be sub-optimal and there exist rules which need a training set to adjust the parameters in order to get a better performance.

In this work, we present a method for estimating the optimal weight for combing two matchers under Gaussian assumptions.

Suppose $f_1(X)$ represent the scores obtained from the video classifier (face recognition), and $f_2(X)$ represent the scores obtained from the audio classifier (speech recognition). X represent the input of both video and audio signals. Suppose the training data have k classes, that $f_1^k(X)$ represent the score belong to the kth class. For a given weight ∂, the fusion score is obtained as follows [16]:

$$f_{fusion}^{k}(X,\alpha) = \alpha f_1^k(X) + (1-\alpha) f_2^k(X) \qquad (4)$$

The notation $f_{fuison}^{k}(X,\alpha)$ indicate that the fusion score depend not only on the input date but also on the weighting coefficient, in what follows, we simplify the notation for functions by dropping argument X and α, except when it is necessary to distinguish among different values of these arguments.

The first step of the estimate method proposed in this paper is normalize the genuine scores, here, we use the reduction of high scores effect normalization (RHE) method proposed by[17], the formula is taken as follows:

$$x^{'} = \frac{x - \min(X)}{mean(X) + std(X) - \min(X)} \qquad (5)$$

Where X denote the all raw scores, $mean(X)$ denote the arithmetic mean of X and $std(X)$ denote the standard deviation of X.

After normalization step, the next step is to estimate the probability distribution of the fusion scores, we assume that the values of the score functions are independent, so:

$$P(f_{fusion}^{1}, f_{fusion}^{2}, \dots f_{fusion}^{K} | X \in w_i) = \prod_{k=1,k\neq i}^{K} P(f_{fusion}^{K} | X \in w_i) \qquad (6)$$

The last step is calculated $C_i(\alpha)$ as equation 7:

$$C_i(\alpha) = \prod_{k=1,k\neq i}^{K} P(f_{fusion}^{K} < 0 \mid X \in w_i) \qquad (7)$$

3.2 Probability Density Estimate

To calculate the probability $P(f_{fusion}^{K} < 0 | X \in w_i)$, we first have to estimate the probability distribution $P(f_{fusion}^{K} < 0 | X \in w_i)$ from the training data in the form of Gaussian Mixture Model. Suppose there have M training data available for deciding the optimal weight, and there M_1 training date belong to class W_1, M_k training date belong to class W_k, which $M_1 + M_2 + M_3 + ... + M_k = M$. We denote the M_i training date belong to class W_i as $\{X_1, X_2, ... X_{M_i}\}$, the Gaussian mixture is obtained by:

$$P(f_{fusion}^{K} < 0 \mid X \in w_i) = \frac{1}{M_i} \sum_{j=1}^{M_i} \frac{1}{\sqrt{2\pi}\delta} \exp(-\frac{(f_{fusion}^{k} - \mu_{kj})^2}{2\delta^2}) \qquad (8)$$

The component means $\mu_{kj} = f_{fusion}^{k}(X_j, \alpha)$. From this, it shown that the means of the mixture component are the score of the training date. When α is large, the variance of each mixture component is large. In the extreme case when α is zero, the probability density shrinks to a series of impulse functions.

Using the estimated probability density function, the probability that $P(f_{fusion}^{K} < 0 | X \in w_i)$ is computed as follows:

$$\begin{aligned}&P(f_{fusion}^{K}<0\,|\,X\in w_i)\\&=\frac{1}{M_i}\sum_{j=1}^{M_i}\frac{1}{\sqrt{2\pi}\delta}\int_{-\infty}^{0}\exp(-\frac{(f_{fusion}^{k}(X,\alpha)-\mu_{kj})^2}{2\delta^2})d[f_{fusion}^{k}(X,\alpha)]\\&=\frac{1}{M_i}\sum_{j=1}^{M_i}\Phi(-\frac{\mu_{kj}}{\delta})\end{aligned} \quad (9)$$

Where $\Phi(x)$ is the integral of the Gaussian distribution:

$$\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx \quad (10)$$

Form equation (7), the $C_i(\alpha)$ finally obtained as follows:

$$C_i(\alpha)=\frac{1}{M_i^{\,k-1}}\prod_{k=1,k\neq i}^{K}(\sum_{j=1}^{M_i}\Phi(-\frac{\mu_{kj}}{\delta})) \quad (11)$$

The overall correct recognition rate, denoted as $C(\alpha)$, is given as

$$C(\alpha)=\sum_{i=1}^{K}C_i(\alpha)\frac{M_i}{M} \quad (12)$$

Thus, the problem of choosing optimal weight for combining two classifier were transformed to maximizing the $C(\alpha)$. Once the weighting coefficient is confirmed, we assume it does not change when it is applied to the test data. In this paper, we still assume that the probability distribution of the training data and test data is the same, because our purpose is on accurate estimation of the parameter.

Figure 2 shows the correct recognition rate $C_e(\alpha)$ as α increases from 0 to 1 with step of 0.01. The combined classifier achieves the highest recognition rate when α is 0.24, so we using this weight in this paper.

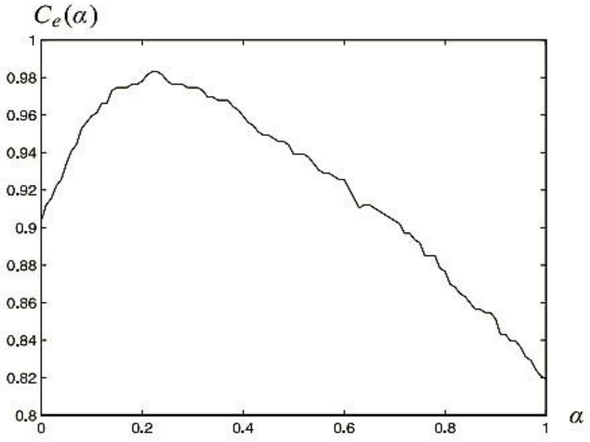


Figure 2. The Correct Recognition Rate Estimation with α Varying from 0 to 1

4. Multimodal Biometric Recognition System

This section deals with the procedures of the proposed mulitmodal biometric system through the rank level fusion. Face recognition based on kernel discriminant locality preserving projections (KDLPP) algorithm is given in Section 4.1, and speech recognition based on adaptive Gaussian Mixture Model and static and dynamic feature fusion is proposed in section 4.2.

4.1 Face Recognition

In this sections, we using KDLPP algorithm for face recognition which the author studied before [18].The KDLPP algorithm involve two major steps. The first step in to obtain the Gram matrix K and then to reduce the dimensionality of the original data features by applies the DLPP/QR algorithm.

The key idea of kernel Discriminate Local Preserve Projection Algorithm is to solve the problem of DLPP in an implicit feature space F, which is constructed by the kernel trick. Consider there is a feature mapping ϕ which maps the input data into a higher dimensional inner product space F [19]. So DLPP can be performed in F and it is equivalent to maximizing the following criterion:

$$J(G)=\frac{\left|G^T S_b^{L^\phi} G\right|}{\left|G^T S_w^{L^\phi} G\right|} \tag{13}$$

$$S_w^{L^\phi} = X^\phi L^\phi X^{\phi^T} \tag{14}$$

$$S_b^{L^\phi} = F^\phi H^\phi F^{\phi^T} \tag{15}$$

Referring to (13), any column of the solution G must lie in the span of all the samples in F, so there exit coefficients α_{ij} such that [20]:

$$g=\sum_{i=1}^{c}\sum_{j=1}^{n_i}\alpha_{ij}\phi(x_{ij}) \tag{16}$$

Where g represents any one column of the projection matrix G. In other words, we can project each vector onto an axis of F as follows:

$$g^t\phi(x)=\sum_{i=1}^{c}\sum_{j=1}^{n_i}\alpha_{ij}k(x_{ij},x)=\alpha^t\varepsilon_x \tag{17}$$

Where

$$\varepsilon_x=(k(x_{11},x),...,k(x_{1n_1},x),...,k(x_{c1},x),...,k(x_{cn_c},x))^t \tag{18}$$

$$\alpha=(\alpha_{11},...,\alpha_{1n_1},...,\alpha_{ij},...,\alpha_{c1},...,\alpha_{cn_c})^t \tag{19}$$

$$K(x_1,x_2)=\langle\phi(x_1),\phi(x_2)\rangle \tag{20}$$

Thus, by using the definitions of $S_w^{L^\phi}$, $S_b^{L^\phi}$ and (17), we can obtain:

$$G^T S_b^{L^\phi} G = A^T K_b^{L^\phi} A \tag{21}$$

$$G^T S_w^{L^\phi} G = A^T K_w^{L^\phi} A \tag{22}$$

$$K_w^{L^\phi} = K(X)LK(X)^T \tag{23}$$

$$K_b^{L^\phi} = K(F)HK(F)^T \qquad (24)$$

$$K(X) = [K(x_1), K(x_2), ..., K(x_N)] (25)$$

$$K(F) = [K(f_1), K(f_2), ..., K(f_C)] \quad (26)$$

$$\begin{aligned} K(x_i) = \varepsilon_{x_i} = (k(x_{11}, x_i), ..., k(x_{1n_1}, x_i), ..., \\ k(x_{c1}, x_i), ..., k(x_{cn_c}, x_i))^t i = 1,2,...,N \end{aligned} \qquad (27)$$

$$\begin{aligned} & K(f_i) = (\frac{1}{n_i}\sum_{k=1}^{n_i} k(x_{11}, x_{ik}), ..., \frac{1}{n_i}\sum_{k=1}^{n_i} k(x_{1n_1}, x_{ik}), ..., \\ & \frac{1}{n_i}\sum_{k=1}^{n_i} k(x_{c1}, x_{ik}), ..., \frac{1}{n_i}\sum_{k=1}^{n_i} k(x_{cn_c}, x_{ik}))^t \qquad (28) \\ & i = 1,2,...,C \end{aligned}$$

So the objective of KDLPP can be written as follows:

$$J(A) = \frac{\left| A^T K_b^{L^\phi} A \right|}{\left| A^T K_w^{L^\phi} A \right|} = \frac{\left| A^T K(F)HK(F)^T A \right|}{\left| A^T K(X)LK(X)^T A \right|} \quad (29)$$

Therefore, similar to DLPP algorithm, the optimal solution of equation (29) can be computed by finding the leading r eigenvalues $\{\alpha_i\}_{i=1,2,...,r}$ of $(K_w^{L^\phi})^{-1} K_b^{L^\phi}$ corresponding to the nonzero eigenvalues. Once $A = [\alpha_1, \alpha_1, ..., \alpha_r]$ is obtained, for a given pattern x, we can map it to a r-dimensional space spanned by the column of A. This projection is given by $y = A^T x$.

The solution of A is complexly and always suffer from the small sample size problem, so we using QR decomposition matrix analysis to handle this issue[21,22]. The first step is to decompose $K_b^{L^\phi}$ as follows:

$$K_b^{L^\phi} = H_b^{L^\phi} (H_b^{L^\phi})^T \qquad (30)$$

Therefore we do QR decomposition on $H_b^{L^\phi}$ by $H_b^{L^\phi} = QR$. For any given matrix $G \in R^{r \times r}$, with $r = rank(H_b^{L^\phi})$, the solution of A is given by $A = QG$, that

$$J_\phi(A) = \frac{\left| (QG)^t K_b^{L^\phi} QG \right|}{\left| (QG)^t K_w^{L^\phi} QG \right|} = \frac{\left| G^t \tilde{k}_b G \right|}{\left| G^t \tilde{k}_w G \right|} \qquad (31)$$

The final step is to compute an optimal G by solving the largest r eigenvalues problem on $(\tilde{K}_w)^{-1}\tilde{K}_b$. Table **Error! Reference source not found.** resume the step of KDLPP algorithm.

Table 1. Procedure of KDKPP Algorithm

urpose: compute projection matrix A
Steps:
1. compute $K(X)$ and $K(F)$ from (25)and (26).
2.Compute $K_w^{L\phi}$ and $K_b^{L\phi}$ from (23) and (24).
3.Construct $H_b^{L\phi}$ from equation (30).
4.Perform QR decomposition on $H_b^{L\phi}$, $H_b^{L\phi} = QR$.
5.Compute $\tilde{K}_b = Q^t K_b^{L\phi} Q$, $\tilde{K}_w = Q^t K_w^{L\phi} Q$.
6.Compute the r eigenvalues g_i of $(\tilde{K}_w)^{-1}\tilde{K}_b$, corresponding the r largest eigenvalues.
7.The projection matrix is $A = QG$ with $G = [g_{1,} g_2, ..., g_r]$.

The main procedure of our method is depicted in Figure 3.

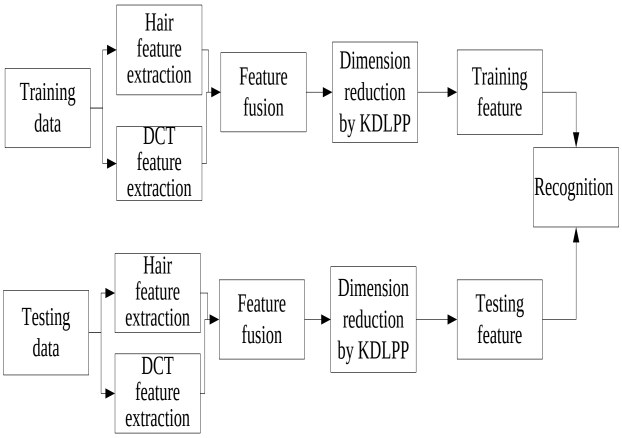


Figure 3.Block Diagram of Recognition Process

4.2 Speech Recognition

In this section, we recognize the speech signal using adaptive Gaussian Mixture Modal and statistic and dynamic feature fusion. The feature used in this paper is Gamma tone Filter Cepstral Coefficients (GFCC) and Gamma tone Filter Shifted Delta Cepstral Coefficients (GFSDCC) based on Gamma tone Filter.

Gaussian Mixture Model is the common and popular used recognition algorithm for speech recognition [23], a Gaussian mixture is modeled by M Gaussians (GMMS) as follows:

$$P(x_t|\lambda)=\sum_{i=1}^{M}\omega_i b_i(x_t) \tag{32}$$

Here , x_t represent the speech feature vectors, ω_i is the weight of Gaussian m, and $b_i(x_t)$ is the component mixture of the i th mixture and is parameterized by mean vector μ_i and the diagonal covariance matrix $\sum_i$ [24],

$$b_i(x_t)=\frac{1}{(2\pi)^{D/2}\,|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}(x_t-\mu_i)^T\Sigma_i^{-1}(x_t-\mu_i)\right\} \tag{33}$$

The mixtures satisfy the constraint criteria that $\sum_{i=1}^{M}\omega_i=1$, the complete Gaussian mixture density is parameterized by the mean vectors, diagonal covariance matrix and weight which can represent as the 3-tuple:

$$\lambda=\{(\omega_i,\mu_i,\textstyle\sum_i),i=1,2,\cdots,M\} \tag{34}$$

The performance of Gaussian Mixture Model (GMM) declines rapidly when the length of the training data is reduced under different unexpected noise environment, so the adaptive process for each GMM model with sufficient training data is transformed to the shift factor based on Factor Analysis[25], when the training data is insufficient, the coordinate of the shift factor is learned from the GMM mixtures of insensitive to the training data and then it is adapted to compensate other GMM mixtures. In this paper, we sorting the Gaussian mixture model component according to the value of weight ω_i, the components which ranking at first third are constitute the collection H, ranking at last third are constitute the collection L, ranking at the middle are constitute the collection M.

It well known that the GMM-UBM modal is the most commonly used GMM model. Suppose the mean super vector of UBM model which sorting the weight based on descending order is μ_{ubm} , the mean super vector Sorting consistent for the speaker s together with the Gaussian components belong to UBM model is μ_s. So the difference between the two μ_{s_u} is represent the speaker's mean super vector moving in the high dimension space under its adaptive process, then we divide the μ_{s_u} into collection $\{\mu_{s\_u\_H},\mu_{s\_u\_M},,\mu_{s\_u\_L,}\}$ according to the weight value ranking of Gaussian component. From the above suppose, we can find the confidence of $\mu_{s\_u\_H}$ is good because of the length of training samples are influence the high weight collection H lesser. While the confidence of $\mu_{s\_u\_L}$ is bad because of the length of training samples are influence the low weight collection L larger, sometimes will occur distortion phenomenon. So we analyze μ_{s_u} and decompose it into equation (35) based on factor analyze technique [26]:

$$\mu_{s\_u}=\mu_s-\mu_u=Vy(s)+Dz(s) \tag{35}$$

That the k dimension of μ_{s_u} can represented as:

$$\mu_{s\_u\_k}=\sum_{i=1}^{h}v_{ki}y(s)_i+\eta_k \tag{36}$$

Combining with equation (35), we can obtain the next equation:

$$\mu_{s\_u} = \mu_s - \mu_u = \begin{Bmatrix} \mu_{s\_u\_H} \\ \mu_{s\_u\_M} \\ \mu_{s\_u\_L} \end{Bmatrix} = \begin{bmatrix} V_H \\ V_M \\ V_L \end{bmatrix} y(s) + Dz(s) \qquad (37)$$

From the equation (36), the matrix V is divide into three parts: $\{V_H, V_M, V_L\}$, while the value of $Dz(s)$ is little and its can negligible in this paper, and $\mu_{s\_u\_M}$ also can negligible because its null vector. So the equation (37) can represent as:

$$\mu_{s\_u\_H} = V_H y(s) \qquad (38)$$

$$\mu_{s\_u\_L} = V_L y(s) \qquad (39)$$

As discussed above, $\mu_{s\_u\_H}$ influenced by the length of the training samples lesser, so we can estimate moving factor $y(s)$ through $\mu_{s\_u\_H}$ and V_H by equation (38), then the $\mu_{s\_u\_L}$ is computed by V_L and the estimated $y(s)$ using equation (39).

4.3 Rank Level Fusion

After getting the recognition result with ranks by unimodal face recognition system and speech recognition system, the ranked output then we fusion it at rank level by using Borda Count method and logistic regression method, and the weight using in the logistic regression method it's computed by Gaussian Mixture Modal estimate which proposed in this paper.

As mentioned above, we choose 0.24, 0.76 as the weights for the face and speech matchers respectively. The more the weight, the less the recognition performance of the system. This means the speech matcher give us less accurate results than the face matchers.

The capacity of a system is an important issue for biometric design, as we consider only the top ten matched samples for the last rank level fusion, so the system face no problem to work for a large database. No matter what the number of the samples that we put into the training set, the system will output only the first ten ranked samples from those training set.

5. Experiment and Results

This section consists of a description of the database used in this paper and the description of the extensive experimental step.

5.1Experiment Overview

The simulation experiments in this paper are consisting of three parts:

(1) Unimodal face recognition using KLDPP algorithm under different conditions.

In this experiment, the subset of AMI database named AMIES2016 was used. For this experiment, we captured 5 video segments from each people's video that at last a total of 20 small video segments were obtained. We denoted it's as S1 to S20. For the reason of most of the image frames in the video have poor quality and no nose in the images, so we should delete it and then regular the image to guarantee nose is in the center of the image. Then select 10 frames from the video and record as 1 to 10 to construct the AMIES2016 face database. For each image, we normalized it to form the uniform size of 64*64. Figure

4 show 10 frames selected from one video.

Figure 4. Face Images of AMIES2016 Database Videos

(2) Unimodal speech recognition using Adaptive Gaussian Mixture Model and GFCC with GGSDCC feature fusion under different noise conditions.

The experiment data also come from AMI database, which consist of 200 voice segments, each voice segment is corresponding a face figure proposed in the above, and the length of each segment is 1 minute. Those 100 segments are used for training GMM parameters, and the rest used for testing.

(3) Visual and audio recognition system result fusion at rank level fusion.

In this experiment, we divide the data using in this paper into 2 parts, one is training part and the other is test part. Each part composed of 10 sets, each set consist of 5 face images and 5 voice segments. We compare the performance of the Borda Count method and Logistic Regression method in this experiment. The weight using for the logistic regression method in this experiment is 0.24.

5.2 Face Recognition Experiment

We randomly take k images from each class as the training data, with $k \in \{2,3,...,9\}$, and leave the rest $10-k$ images as the test data. The Nearest Neighbor algorithm was employed using Euclidean distance for classification. There are two small experiments taken in our experiment as follows:

Experiment A. Compare recognition accuracy based on KDLPP algorithm under different kernel functions.

The input data of the LDLPP is the kernel matrix and it is necessary to choose an adequate kernel function to construct this matrix. In this paper, we used Polynomial kernel function, Gaussian RBF kernel function and Fractional polynomial kernel function. Table 2 present the kernel functions used in our studies.

Table 2. Kernel Functions

1.Polynomial kernel function
$K(x,y)=(1+xy^d),d \in N$
2.Gaussian RBF kernel function
$K(x,y)=\exp(-\|x-y\|^2 \big/ 2\delta^2)$
3.Fractional polynomial kernel function
$K(x,y)=(1+xy^d),0 \prec d \prec 1$

In order to illustrate the effect of kernel function choice, Figure 5 to Figure 7 shows the results of KDLPP algorithm with different kernel functions. In Figure 5 we can show that for Polynomial kernel the performance decrease with the parameter d increasing. And globally gives less result than Gaussian RBF kernel function and Fractional polynomial kernel function. For Gaussian RBF kernel function, the value $\delta^2 = 10^9$ gives maximum recognition rate compares to others values of δ . The performance of Fractional polynomial kernel function with value $d = 0.4$ is good but it is lower than Gaussian RBF kernel function.

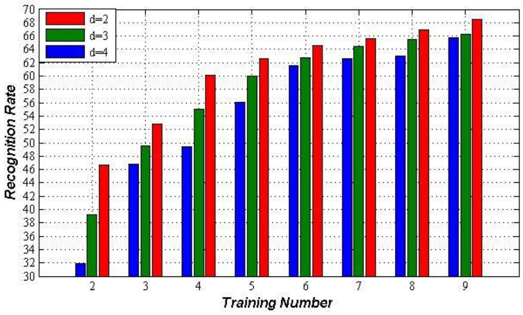


Figure 5. Recognition Accuracy of KDLPP Algorithm under Polynomial Kernel Function

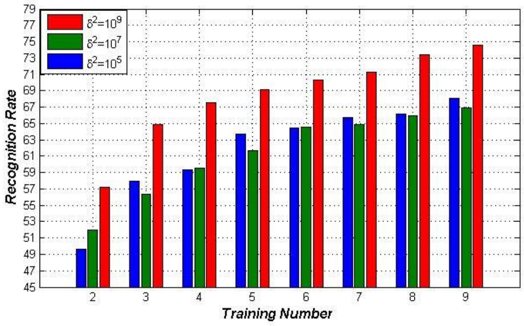


Figure 6. Recognition Accuracy of KDLPP Algorithm under Gaussian RBF Kernel Function

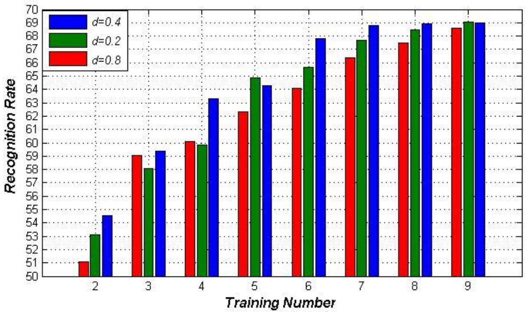


Figure 7. Recognition Accuracy of KDLPP Algorithm under Fractional Polynomial Kernel Function

Experiment B. Compare recognition accuracy based on different algorithms.

In this small experiment, we tested the FDA and DLPP methods compare to our proposed KDLPP algorithm, the kernel function used in this experiment is Gaussian RBF kernel function, the value $\delta^2 = 10^9$. Figure 8 give the recognition rate result. From the result we can show that KDLPP algorithm gives the best result under any training number situations, and FDA method give the worst result. From the figure we can also know that the face recognition rate under smart meeting environment is less than the standard face database environment because of the problem of poor quality image, lighting condition and facial expression change and so on.

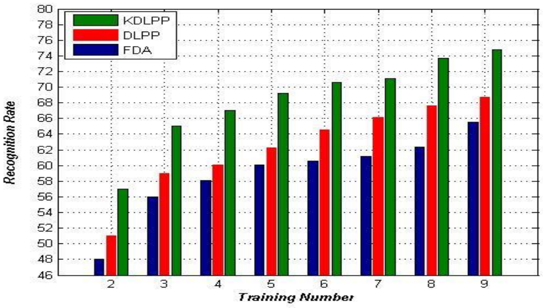


Figure 8. Recognition Accuracy of Different Algorithms based on Hair and DCT Feature Fusion

5.3 Speech Recognition Experiment

In this experiment, the dimension of MFCC is 13, and the dimension of GFSDCC is 39, then we fusion them at feature level. There are two small experiments taken in our experiment as follows:

Experiment A. Compare recognition accuracy based on different acoustic features under different conditions.

We compare the recognition performance of speech recognition based on baseline GMM which using MFCC feature, GFCC feature, GFSDCC feature and GFCC, GFSDCC feature fusion under different noise conditions, the result are taken as follows:

From the results shown in the Tab. 3, it clears that whether under clear voice conditions or under different noise conditions, the recognition performances of GFCC features are better than MFCC features. On average, it will be higher than 2.1% that because the discrimination performance of GFCC features which based on Gamma tone filters is better than the MFCC features which based on Mel filters. The Gamma tone auditory model reflects the anti-noise capacity of human auditory system preferably.

Table 3. Recognition Rate based on Different Voice Features under Different Voice Environments

Voice conditions feature		MFCC	GFCC	GFSDCC	GFCC+ GFSDCC
Clean background	SNR				
	> 40dB	86.72%	89.02%	91.38%	93.54%
White noise	0dB	34.63%	34.17%	42.01%	49.76%
	5dB	47.77%	48.17%	54.19%	58.72%

	10dB	50.25%	54.13%	57.48%	61.33%
	15dB	59.58%	62.66%	64.24%	69.24%
	20dB	61.56%	63.32%	66.22%	73.30%
Babble noise	0dB	32.72%	33.71%	34.64%	36.95%
	5dB	47.03%	48.4%	50.99%	56.84%
	10dB	54.13%	56.04%	59.00%	64.62%
	15dB	63.44%	65.27%	66.71%	71.17%
	20dB	67.79%	73.64%	75.24%	78.09%
average		55.05%	57.13%	60.19%	64.86%

The GFSDCC features based on Shifted Delta Cepstral not only using the anti-noise capacity of Gamma tone filters but also can fusing a long sequence information in a feature vector, thus depict the dynamic information of audio features strongly. The recognition performance of GFSDCC features is 91.38% under clean background which 3.06% higher than the performance based on GFCC features. The recognition performances of GFSDCC features are improve obviously when under White noise and babble noise environments.

Because the GFCC feature and the GFSDCC feature only reflect the one side of audio features, either static or dynamic, so we fusing them at feature level which considering the static and dynamic character comprehensively. The recognition performance based on fusion feature is 93.54% under clean background, compared to the performance based on MFCC、GFCC、GFSDCC features, is higher than 6.82%, 4.52% and 2.16% respectively. The recognition performance based on fusion features are higher than single features under different noise environments, then recognition rate is 78.09% when the SNR equal 20 under babble noise environments.

Experiment B. Compare recognition accuracy based on Adaptive GMM and Static with Dynamic feature fusion under different conditions.

The results shown in the Tab.4 are the recognition performance using fusion features and adaptive GMM proposed in this paper for speech recognition under different noise environments. Form the result shown in the Tab.4 it clear that the recognition rate using the method proposed in this paper is reached 95.46% under clean background. That 1.92% higher than using the baseline GMM and fusion feature method thus reflect the superiority of the method proposed in this paper.

Table 4. Recognition Rate Based on Proposed Method under Different Noise Backgrounds

Voice Environment	SNR	Recognition Rate
Clean Background	>40dB	95.46%
White Noise	0dB	53.41%
	5dB	62.72%
	10dB	67.33%
	15dB	73.58%

	20dB	80.93%
Babble Noise	0dB	43.46%
	5dB	66.13%
	10dB	74.46%
	15dB	76.43%
	20dB	83.24%

The recognition rates are enhanced significantly using the method proposed in this paper compared to the method using in Tab.3 under white noise environment and babble environment. The recognition rate is reached 80.93% when the SNR is 20 under white noise environment while under babble environment is 83.24%, compared to the results shown in Tab.3 which using fusion feature and baseline GMM, improved 7.63% and 5.15% respectively.

5.4 Multibiometric Recognition at Rank Level Fusion

Tab.5 given the recognition rate of Borda Count rank level fusion method and Logistic Regression rank level fusion method proposed in this paper under clean environment.

Table 5. Comparison of Different Rank Level Fusion Systems under Clean Environment

Approaches	Recognition Rate
Borda Count Fusion	97.25%
Logistic Regression Fusion	98.70%

From the result, it clear that the GMM estimate rank level fusion method can give a good performance under clear environment, which can attain 98.70%.

Fig .9 and Fig. 10 shown the performance rate comparing between face recognition using training samples are 5 for each set, the kernel function used in this experiment is Gaussian RBF kernel function, the value $\delta^2 = 10^9$. speech recognition using fusion features and adaptive GMM proposed in this paper under different noise environments, Borda Count rank level fusion method and Logistic Regression rank level fusion method proposed in this paper.

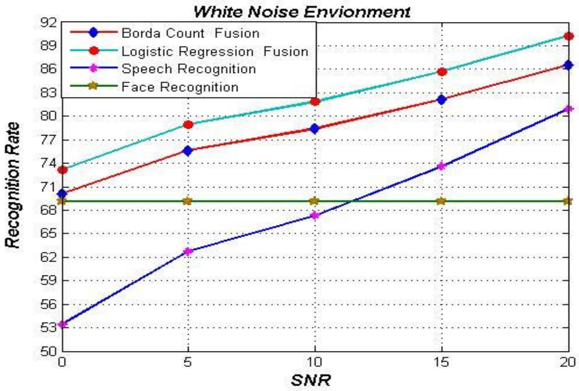


Figure 9. Different Recognition Methods Results Comparison under White Noise Environment

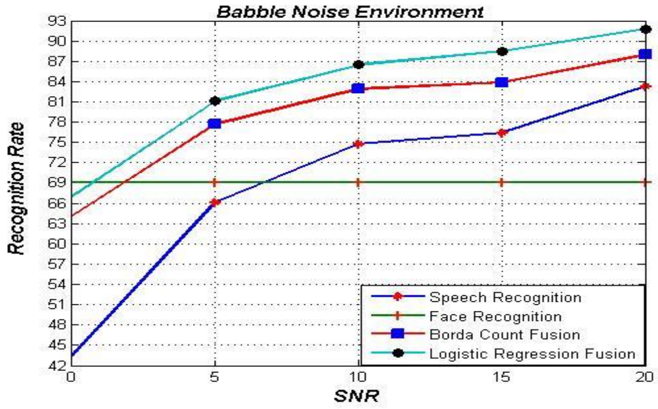


Figure 10. Different Recognition Methods Results Comparison under Babble Noise Environment

From the figure, it is clear that the recognition would be low without any fusion method, significant performance gain can be achieved with different rank level fusion methods. The best performance that we have received from these experiments is using the logistic regression approach based on GMM estimate proposed in this paper of the rank level fusion method. When the SNR is 20, the speech recognition rate is below 80% under white noise environment but using the logistic regression approach based on GMM estimate its can enhanced to 90.26%. Under babble noise environment, When the SNR is 20, the speech recognition rate is near 84% but using the logistic regression approach based on GMM estimate its can enhanced near to 93%. Because in this method, assigning different weights to individual matchers based on their accuracy which plays an important role in determining the final result.

The Borda Count rank level fusion method give less performance than the logistic regression approach because there is no weight assigning procedure in this method. But from the given results, the Borda Count rank level fusion method also can improve the recognition performance significantly.

Table 6. Comparison of Different Multi Biometric Systems

Systems	Biometric Recognition	Fusion Level and Approaches	Equal Error Rate(EER)
Current System (2013)	Face Speech	GMM estimate based Rank Level Fusion	1.57%
Carcia -Salicetti et al(2005)	Signature Speech	Match Score Arithmetic Mean Rule(AMR)with Min-Max normalization	1.88%
Kumar and Zhang(2004)	palm print	Match Score Product of Sum(POS) rules	3.20%
Nandakum et al(2007)	Fingerprint	Match ScoreQuality Weighted Sum(QWS) rules	3,39%

In the Tab.6, we compare our results with some other multibiometric systems which are proposed researchers. The comparison is made on the value of the EER systems, EER is the value of the position in the DET curve where FAR is equal to FRR. From the results, it is clear that the rank level fusion with the logistic regression approach based on GMM estimate can increase the recognition rate significantly.

Tab.7 shown the comparison of recognition time between different methods, due to the extra calculation for combining results from different systems, recognition time of multibiometric system is higher than the single biometric system, the difference in the recognition time between them depend on the number of computations involved in the adopted fusion techniques and the number of samples considered.

Table 7. Recognition time Comparison

Approaches	Recognition Time(min)
KDLPP Face Recognition	0.53±0.15
Adaptive GMM Speech Recognition	0.60±0.19
Borda Count Fusion	0.70±0.10
Logistic Regression Fusion	0.77±0.13

6. Conclusion

The domain of multibiometric is a new and exciting area of information science and pattern recognition research. Recent years have seen a significant increase in research activity directed at understanding all aspects of biometric information system representation and utilization for decision making support, this paper is specifically focused on understanding the complex mechanisms employed to find a good combination of multiple biometric traits and various fusion methods to get the optimal recognition results.

In this paper, a novel fusion scheme which using Gaussian Mixture Model(GMM) to estimate the optimal weight for logistic regression rank level fusion method is proposed, The estimate is based on estimation of probability destiny functions. Using the simulated date, the results indicate that this method can improve the recognition performance significantly compared to other methods.

The logistic regression approach provide the better performance, also the response time is a little than that of the Borda Count method. Future work should focus on following three fields:

(1) The face recognition is influenced by pose and illumination variation or facial vary, so our future work will focused on these problems.

(2) The logistic regression rank level fusion method given the better performance than other methods, our weight estimate is based on GMM, to our best knowledge; it is an optimal weight estimate method at present. Future work will focused on estimate the best optimal weight by using other methods, such as genetic algorithm and normalization technique.

(3) the rank level fusion method proposed in this paper are the most common approaches, recently, researchers are interesting on studying the Bayesian approach and Markov Chain approach, also is our future works.

Acknowledgement

This work was supported by in part the National Science-technology Support Plan Project of china under contract 1214ZGA008, in part by the Nature Science Foundation of china under contract 61263031, in part by the Science Foundation of Gansu Province of China under contract 1010RJZA046.

References

[1] Y. Wang, "The theoretical framework of cognitive informatics", Int. J.Cognit. Informat. Nat. Intell., vol. 1, no. 1, **(2007)**, pp. 10–22.

[2] Y. Xu, Q. Zhu and D. Zhang, "Combine crossing matching scores with conventional matching scores for bimodal biometrics and face and palmprint recognition experiments", vol. 74, **(2011)**, pp. 3946–3952.

[3] A. K. Jain, A. Ross and S. Pankanti, "Biometrics: A tool for information security", IEEE Trans. Inf. Forensics Security, vol. 1, no. 2, **(2006)**, pp. 125–143.

[4] A. K. Jain, A. Ross and S. Pankanti, "Biometrics: A tool for information security", IEEE Trans. Inf. Forensics Security, vol. 1, no. 2, **(2006)**, pp. 125–143.

[5] B. Khaleghi, A. Khamis and F. Karray, "Random finite set theoretic based soft/hard data fusion with application for target tracking", Proc. of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, **(2010).**

[6] M. P. Down and R. J. Sands, "Biometrics: An overview of the technology, challenges and control considerations", Inf. Syst. Control J., vol. 4, **(2004)**, pp. 53–56.

[7] A. Ross and R. Govindarajan, "Feature level fusion using hand and face biometrics", Proc. SPIE 2nd Conf. Biometric Technol, Human Identification, **(2005);** Orlando, FL.

[8] R. Giot and C. Rosenberger, "Genetic programming for multibiometrics", Expert Systems with Applications, vol. 39, **(2012)**, pp. 1837–1847.

[9] A. Omari and A. R. Figueiras, "Feature Combiners with Gate Generated Weights for Classification", IEEE Transaction on Neural Networks and Learning Systems, vol. 24, no. 1, **(2013)**, pp. 158-163.

[10] A. Ross and A. K. Jain, "Information fusion in biometrics", Pattern Recogn. Lett., vol. 24, no. 13, **(2003)**, pp. 2115–2125.

[11] A. Ross, K. Nandakumar and A. K. Jain, Handbook of Multibiometrics.

[12] J. Bhatnagar, A. Kumar and N. Saggar, "A novel approach to improve bio-metric recognition using rank level fusion", Proc. IEEE Conf. Comput.Vis. Pattern Recog. , Minneapolis, MN, **(2007).**

[13] M. M. Monwar and M. L. Gavrilova, IEEE Transcations on Systems, Man, and Cybernetics—Part B: Cybernetics, vol. 39, no. 4, **(2009)**, pp. 867-878.

[14] B. Khaleghi, A. Khamis, O. Fakhreddine and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art", Information Fusion, vol. 14, **(2013)**, pp. 28–44.

[15] J. Kittler and F. M. Alkoot, "Sum versus vote fusion in multiple classifier systems", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 1, **(2003)**, pp. 110–115.

[16] A. Jain, K. Nandakumar and A. Ross, "Score normalization in multimodal biometric systems", Pattern Recognition, vol. 38, no. 12, **(2005)**, pp. 2270-2285.

[17] A. Kumar, V. Kanhangad and D. Zhang, "A new framework for adaptive multimodal biometrics management", IEEE Transactions on Information Forensics and Security, vol. 5, no. 1, **(2010)**, pp. 92-102.

[18] D. Wu, J. Cao, J. H. Wang and W. Li, "Multi-feature fusion face recognition based on Kernel Discriminate Local Preserve Projection Algorithm under smart environment", Journal of Computers, vol. 7, no. 10, **(2012)**, pp. 2479-2487.

[19] W. Yu, X. Teng and C. Liu, "Face recognition using discriminant locality preserving projections", Image Vision Computing, vol. 24, **(2006)**, pp. 239–248.

[20] B. C. Zhang and Y. Qiao, "Face recognition based on gradient gabor feature and Efficient Kernel Fisher analysis", Neural Computing & Applications, vol. 19, no. 4, **(2010)**, pp. 617-623.

[21] M. A. X. Hong and L. L. Zhao, "A robust audio watermarking method based on QR decomposition and lifting wavelet transform", Journal of Dalian University of Technology, vol. 50, no. 2, **(2010)**, pp. 278-282.

[22] S. F. Xie, S. G. Shan, X. L. Chen and J. Chen, "Fusing Local Patterns of Gabor Magnitude and Phase for Face Recognition", IEEE Transactions on Image Processing, vol. 19, no. 5, **(2010)**, pp. 1349-1361.

[23] T. K. Perrachione, S. N. Del Tufo and J. D. E. Gabrieli, "Human Voice Recognition Depends on Language Ability", Science, vol. 333, **(2011)**, p. 595.

[24] J. Du and Q. Huo, "A Feature Compensation Approach Using High-Order Vector Taylor Series Approximation of an Explicit Distortion Model for Noisy Speech Recognition", IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 8, **(2011)**, pp. 2285-2293.

[25] Y. J. He and J. Q. Han, "Gaussian Specific Compensation for Channel Distortion in Speech Recognition", IEEE Signal Processing Letters, vol. 18, no. 10, **(2011)**, pp. 599-602.

[26] O. Dehzangi, B. Mab, E. S. Chng and H. Z. Li, "Discriminative feature extraction for speech recognition using continuous output codes", Pattern Recognition Letters, vol. 33, **(2012)**, pp. 1703-1709.

[27] http://www.amiproject.org/.

Author

Di Wu, he was born in Xiang Tan Hu Nan province of China in 1985. Received bachelor degree in communication system from Jiu Jiang university, China, in 2007, and received master degree and doctor degree in signal and information processing from Lan Zhou university of technology,China, in 2010 and 2014. His main research area is information fusion theory and application, multi-person speech recognition.