A Survey of Multimodal Fusion for Identity Verification

Yongmin Li¹, Guanyu Li²*, Pengqi Li¹, Sixuan Li¹, Xinyu Yuan¹, Ding Liu¹ and Xuebin Yang¹

¹ Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730000, China
² Key Laboratory of China's Ethnic Languages and Intelligent Processing of Gansu Province, Northwest Minzu University, Lanzhou, Gansu 730000, China

*Corresponding author’s e-mail: 296286778@qq.com

Abstract With the rapid development of computer and network technology, the digitization and invisibility of people's identities have become an important feature of today's highly informationized society. This article introduces the knowledge of multimodal fusion, and lists some commonly used multimodal fusion techniques in the field of identity verification, and puts forward some prospects for the future multimodal fusion identity verification.

1. Introduction
With the development of network technology and the popularization of intelligent devices, there are more and more application scenarios of biometrics in life today, such as mobile payment, station gates, check-in systems, access control systems, and mobile phone unlocking. Biometric technology has become an indispensable part of people's daily life. It brings convenience to people and also brings some hidden dangers to information security. Although the single-modal biometrics already has high accuracy, it still has problems such as low security, and being easily restricted by the environment. In order to effectively solve this series of problems, multi-modal identity verification technology has begun to attract more and more attention[1].

2. Introduction to Multimodal Fusion
For the same description object, the data obtained through different fields or perspectives, and each field or perspective describing these data is called a modal, and the data composed of two or more modals is multimodal data. Compared with multi-modal data, single-modal data is difficult to completely represent the described object, and there is a certain degree of incompleteness. In order to solve this problem, it becomes particularly important to enhance the interaction between different modal information and achieve information complementation[2].

2.1. Data Layer Fusion
Data layer fusion is also called sensor layer fusion or pixel layer fusion. It is directly fused on the original data collected by the sensor, and the data is synthesized and analyzed before each sensor is issued without processing. Most of the data layer fusion adopts many-to-one fusion, which is a low-level fusion process.
2.2. Feature Layer Fusion

Feature layer fusion can also be called model layer fusion. It belongs to the middle level of fusion. First, the original data collected by the sensor is extracted, and then the extracted features are comprehensively analyzed. Feature layer fusion can effectively compress information, which is beneficial to real-time analysis, and because the extracted feature information is closely related to decision analysis, the fusion result can give the information required for decision to the greatest extent.

2.3. Decision Layer Fusion

Decision layer fusion can also be called fractional layer fusion. Each single-modal classifier performs preprocessing, feature extraction, recognition or judgment, etc. to obtain several preliminary conclusions, and then makes decision layer fusion decisions by correlating each result[7].

Finally, the results of joint inference are obtained.

3. Multimodal Fusion for Identity Verification

3.1. Multimodal fusion based on Bayesian theory

In pattern recognition, Bayesian theory is often used to reduce the risk of classification errors. We can also use it in multi-modal fusion to obtain reliable decisions. For a binary classification problem about $x$, $H_1$ and $H_2$ represent the two categories of the problem, $p(H_1)$ and $p(H_2)$ represents the prior probability of the category state, and $p(x|H_i)$ represents the probability of observing $x$ given the category $H_i$, $i=1,2$. That is, when $H_i$ appears, the conditional probability density of the observed $x$.

According to Bayesian theory, there are

$$p(H_i|x) = \frac{p(x|H_i)p(H_i)}{\sum_{j=1}^{2} p(x|H_j)p(H_j)}$$ (1)
\( p(H_i|x) \) represents the posterior probability of category status. Through the above operations, the prior probability \( p(H_i) \) can be converted into the posterior probability \( p(H_i|x) \). Finally, it is determined that \( x \) belongs to \( H_i \) class, where

\[
i = \text{argmax} \ p(H_i|x)
\]  \hspace{1cm} (2)

In a multi-modal identity verification system, it is assumed that there are \( n \) models, and their corresponding matching scores are \( s \). Each mode has two verification states \( H_0 \): system verification fails, \( H_1 \): system verification succeeds, and their prior probabilities meet \( p(H=0)+p(H=1)=1 \), and the modes are independent of each other. Therefore, the matching score class conditional probability density of the multimodal fusion system can be obtained as:

\[
p(s|H) = \prod_{i=1}^{n} p_i(s_i|H)
\]  \hspace{1cm} (3)

We use the 0-1 function to evaluate the risk of misjudgment, and assuming that \( p(H=0)=p(H=1)=0.5 \), the decision function of the multimodal fusion system can be

\[
D = \begin{cases} 
1, & \prod_{i=1}^{n} \frac{p_i(s_i|H = 1)}{p_i(s_i|H = 0)} \\
0, & \text{other}
\end{cases}
\]  \hspace{1cm} (4)

### 3.2. DBN-Base Multimodal Fusion Model

The Bayesian network (BN) is composed of a directed acyclic graph with nodes representing variables and directed line segments connecting the nodes. Each node represents a random variable. The directed line segments between nodes represent the mutual relationship between random variables. In the network, the conditional probability is calculated to describe the strength of the relationship between random variables. If a node does not have a parent node, then use prior probability to express it. We regard it as a dynamic Bayesian network (DBN) is the expansion of the Bayesian network in the time dimension[6].

![Figure 2. Network Structure. (a) Bayesian network (b) dynamic Bayesian network](image)

Dynamic Bayesian networks have excellent performance in describing complex random processes with multiple channels. Using the characteristics of the dynamic Bayesian network, we treat each modality as a dynamic Bayesian chain. Multiple dynamic Bayesian chains are aligned by external knowledge, and they are regarded as a whole. It can achieve the fusion of multi-modal data using external knowledge as a benchmark. Several common structures using dynamic Bayesian network for multi-modal fusion are:

#### 3.2.1. Sentence-coupled multi-modal joint modeling

Sentence-coupled multi-modal joint modeling is a common one in the Bayesian network benchmark model. For example, text information is used to align audio and video in audiovisual database to achieve fusion of audio and video data[4].
3.2.2. Frame Coupled multi-modal joint modeling. The dynamic Bayesian network is the expansion of the Bayesian network in the time dimension. Therefore, we can use time-related frames to align the data. This method is only applicable to a limited number of modal data.

3.2.3. State-coupled multi-modal joint modeling. A parallel state sequence is introduced outside the multi-modal data dynamic Bayesian chain. The C node of the multi-modal data affects each single-mode O node. We align the C node with the state node in the parallel state chain, thus enabling multimodal data fusion.

3.2.4. Feature-coupled multi-modal joint modeling. Each single mode in the multi-modal data set has a natural connection with each other. The jump probability of each single mode in the time dimension affects other modes, and the jump of multi-modal data in the time dimension also affects each single mode. Based on this feature, we can fuse multimodal data.

3.3. Multimodal fusion based on multivariate polynomial

From the perspective of decision layer fusion, multimodal fusion is the fusion of scores and decision[5]. We assume that there are \( M \) modal features, and \( S_m \) is the set of matching scores of the \( m \)-th modal, \( m=1,2,3,...,M \), \( m \) represents the number of equation, \( n \) represents the number of unknown parameters and, \( n>m \). The following fusion function can be constructed:

\[
\text{Fusion}(N) = y = s \cdot p
\]  

Among

\[
y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad s = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nm} \end{bmatrix}, \quad p = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{bmatrix}
\]

\( y_i \) is the classification result of the \( i \)-th classifier \( y_i = [0,1] \).

Obviously, this system of equations has no solution in general, so we need to introduce the residual sum of squares equation \( R \) and find the appropriate \( p \) so that the equation holds as much as possible[3].

\[
R(p) = \| s \cdot p - y \|^2
\]  

When \( p = \hat{p} \), \( R(p) \) takes the minimum value and is recorded as

\[
\hat{p} = \arg\min \ R(p)
\]

By differentially calculating the maximum value of \( R(p) \), we can get

\[
S^T S \hat{p} = S^T y
\]

If the \( S^T S \) is a non-singular matrix, then \( p \) has a unique solution:

\[
\hat{p} = \left( S^T S \right)^{-1} S^T y
\]

3.4. Multimodal fusion based on neural network

With the advent of the era of big data, the performance of neural network-based systems has been significantly improved. The data of different modal are extracted after preprocessing to extract their respective features, trained by neural network, and fused at an appropriate stage. In a multi-modal fusion system based on neural network, feature \( \alpha \) and feature \( \beta \) are fused into feature \( F \), \( \alpha = [\alpha_1, \alpha_2, ..., \alpha_n] \) \( \beta = [\beta_1, \beta_2, ..., \beta_m] \).
3.4.1. The principle of addition.

\[ F = [\alpha + \beta] = [\alpha_1 + \beta_1, \alpha_2 + \beta_2, \ldots, \alpha_n + \beta_m] \quad (11) \]

3.4.2. The principle of series.

\[ F = [\alpha_1, \alpha_2, \ldots, \alpha_n, \beta_1, \beta_2, \ldots, \beta_m]_{1 \times (n+m)} \quad (12) \]

3.4.3. The principle of multiplication.

\[ F = [\alpha^T \cdot \beta] = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_m \end{bmatrix} = \begin{bmatrix} \alpha_1 \cdot \beta_1 & \alpha_2 \cdot \beta_1 & \cdots & \alpha_n \cdot \beta_1 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1 \cdot \beta_2 & \alpha_2 \cdot \beta_2 & \cdots & \alpha_n \cdot \beta_2 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1 \cdot \beta_m & \alpha_2 \cdot \beta_m & \cdots & \alpha_n \cdot \beta_m \end{bmatrix}_{n \times m} \quad (13) \]

4. Summary and prospect

The authentication of multi-modal fusion is gaining more attention from academia. Multi-modal fusion data effectively makes up for the deficiencies of single-modal data, and can be applied to the development of other tasks, such as lip-reading[1]. The accuracy of multi-modal fusion identity verification based on neural network has exceeded the traditional fusion method when the amount of data is sufficient. Later, more work will be done on the end-to-end model and integrated into the attention mechanism. With the deepening of the research on multi-modal fusion identity verification, digital identity verification will become more convenient, and it will continue to approach feeless authentication.

Acknowledgment

This research was funded by The National Natural Science Fund (NO.61633013).

Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730000, China

References

[1] Ngiam J, Khosla A, Kim M, et al. Multimodal Deep Learning[C]. International Conference on Machine Learning, 2011: 689-696.
[2] Atrey P K, Hossain M A, Saddik A E, et al. Multimodal fusion for multimedia analysis: a survey[J]. Multimedia Systems, 2010, 16(6): 345-379.
[3] Marquardt D W. An Algorithm for Least-Squares Estimation of Nonlinear Parameters[J]. Journal of The Society for Industrial and Applied Mathematics, 1963, 11(2): 431-441.
[4] Zadeh A, Chen M, Poria S, et al. Tensor Fusion Network for Multimodal Sentiment Analysis[C]. Empirical Methods in Natural Language Processing, 2017: 1103-1114.
[5] Toh K, Yau W, Jiang X, et al. A reduced multivariate polynomial model for multimodal biometrics and classifiers fusion[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2004, 14(2): 224-233.
[6] Wu Z, Wu Z, Cai L, et al. Weight Estimation for Audio-Visual Multi-level Fusion in Bimodal Speaker Identification[C]. International Conference on Intelligent Computing, 2006: 1107-1112.
[7] Jain A K, Nandakumar K, Ross A, et al. Score normalization in multimodal biometric systems[J]. Pattern Recognition, 2005, 38(12): 2270-2285.