Research on Design Innovation Method Based on Multimodal Perception and Recognition Technology

Wu kan*, Men Longlong
School of art, Xi’an University of Science and Technology, Xi’an, Shaanxi, 710054, China
*Corresponding author’s e-mail: 2862369281@qq.com

Abstract. In order to solve the problems of low efficiency and poor robustness of single-mode speech emotion recognition, this paper uses a multi-mode fusion mechanism to fuse speech and visual information across modes to build an audio-visual speech recognition (AVSR) system. The results show that the modal attention mechanism can automatically adjust to a more stable and reliable state according to the quality of a single signal, so that the audio-visual multimodal perception is accurate and the recognition efficiency is high.

1. Introduction

Studies have shown that human beings understand language not only by hearing, but also by considering gestures and visual cues of the face. Although we do not rely much on visual information, speech interaction is multimodal in nature. This is because the auditory modality contains most useful information, which is enough for us to understand others. When sound signals are distorted by noise, visual modes become very important. The goal of AVSR is to use the visual information extracted from the visual motion to supplement the corrupted speech signal.

In the early days, people paid more attention to the study of simple signals such as sound and light, and the mechanism was relatively primary. In recent years, more and more attention has been paid to the research of complex signal perception, such as lip language, speech, gesture recognition and other interactive modes, but most of the current research in this field focuses on the in-depth study of single-mode signals. According to the results of current life science research, the learning and memory of multiple modes can be amplified nonlinearly through a cooperative mechanism, instead of simply linearly superposing the learning and memory of the two modes [1]. Therefore, the research on multimodal machine attention is more in line with the cognitive habits and forms of communication between human and nature, and the research in this field will become a way to break the bottleneck of current human-computer interaction.

In the research of multi-modal biometric identification, some foreign scholars and research institutions have carried out earlier research and achieved more advanced results: Brunelli [2] proposed to use multi-features of biology for identity authentication, and realized the fusion of face and voice at the decision-making level, which verified the feasibility and effectiveness of the multi-modal identification system. The concept of "Multi—modal" was first proposed by Bigun and Duc in 1997 [3]. In 2001, Ross and Jain proposed a hierarchical theory of multi-biological feature fusion [4]. In 2006, "Handbook of Multi.biometrics" edited by Ross and Jain et al. was published, marking the further maturity of multimodal identification technology [5]. Zhu [6] adopt multimodal methods based on knuckles, finger shapes, palm prints and the like, and adopt a 30-frame/second camera to ensure real-
time acquisition. Kumar\[7\] proposed a multimodal fusion method based on finger vein and finger abdomen texture, Prasanlakshmi\[8\] implemented a multimodal fusion method based on face, fingerprint and palm vein, etc. They developed a multi-modal biometric recognition system based on the fusion of face and iris features, and achieved good recognition results in the biometric recognition competition. However, the current research is still insufficient in feature extraction and feature fusion algorithms, and there is no unified standardized algorithm framework and evaluation standard in feature fusion.

Based on the research of emotion recognition of single-mode speech, this paper proposes a multi-mode attention method to fuse information from multi-mode input. The method integrates voice and visual information for bimodal information recognition, and establishes a bimodal emotion recognition system framework integrating voice acoustic features and visual features. Through cross-modal emotion fusion, feature extraction and emotion recognition, multi-modal emotion perception and recognition are realized, thus improving the recognition effect of a single mode.

Multimodal Information Fusion Technology.

1.1. Perception and Recognition of Monomodal Information

According to the existing research results of brain science and neuroscience, human perception and attention have multi-modal properties, such as hearing and vision are often complementary to each other. Psychologists define "Sensing" and "Perceiving" differently\[3\]. Current computers can digitally collect visual and auditory signals through various sensors, and filter noise and enhance signals through various digital filters to obtain a sense of target signals. However, the computer's ability to sense the collected signals is still very weak, and its ability to distinguish the meaning of video and audio signals (such as human state and emotion) is still very low.

In recent years, the development of monomodal biometric identification technology is in full swing, with a wide range of applications, involving advanced and popular technologies such as image processing and pattern recognition technology, machine vision technology, voice processing technology, multi-sensor technology, etc. Many mature products are widely used in the market, as shown in the follow:

![Figure 1. Biometric products](image)

However, no matter which kind of single-mode biometric technology, there is a short board, and "omnipotent" human biological characteristics do not exist. For example, face recognition accuracy decreases due to light occlusion, expression changes, age differences, etc. The universality, uniqueness, stability and accuracy of DNA are very good, but its acceptability and collectibility are poor, so its application scope is very small. Signatures and voices are easy to forge and their accuracy is not high. Fingerprint is not universal because it cannot clearly collect fingerprint images due to scar or cocoon.

Therefore, the single-mode biometric identification system can no longer meet people's needs in terms of recognition accuracy and robustness, and a biometric identification method with higher accuracy and robustness is urgently needed. In order to overcome this problem, a new biometric technology, Multi-modal Biometric, has emerged. In theory, it is an identification technology that
integrates multiple feature information. It combines multiple single biological features with different advantages to obtain an identification technology that does not have the comprehensive characteristics of any single-mode feature identification, so as to "learn from each other and complement each other's strengths".

1.2. Multimodal information fusion
Multimodal perception is an extension of active vision and behavioral perception. Multimodal perception uses multisensory cues to enhance its perception ability and actively establishes multisensory synchronous input. For example, repeated tapping on objects on the ground will produce visual and auditory changes. The existing practice shows that the multi-modal biometric identification technology can effectively improve the overall performance of the identity identification system and achieve the ideal effect that the single-modal biometric identification technology cannot achieve. Its advantages are as follows:

- Multimodal biological characteristics are obviously better than single-modal biological characteristics in characterizing human characteristics.
- It is usually possible to imitate biological modes to achieve the purpose of making false and true.
- Different biological characteristics can make up for their own limitations.

The future development trend of multi-modal fusion is inevitably to synthesize multi-modal data, capture complementary information between modes, and build a recognition model with strong robustness and high recognition degree. With the development of artificial intelligence, multi-modal interaction must be a more humanized and closer way to interact with people.

2. AVSR System Based on Speech and Visual Information

2.1. Audio-visual mode fusion
To understand language only from visual signals, scholars at home and abroad have done a lot of research, and many researchers are committed to extracting powerful visual features for accurate gesture models. It is always a challenge to integrate auditory and visual information into a single AVSR model. One reason is that gestures convey less information than languages, and the other reason is the need for an effective fusion strategy. Visual features can be extracted by convolution neural network and combined with audio features. Under the same space-time condition, the dual-mode common-mode recognition effect is better than the simple superposition of two isolated single-mode recognition effects.

2.2. Construction of AVSR system

![AVSR system identification framework](image)
As shown above, sound and video signals are collected from microphone array and controllable camera respectively. The voice signal is preprocessed, noise is filtered, voice objects are determined, voice positioning calculation is carried out, and azimuth parameters relative to the controllable camera are obtained; According to the result of sound positioning, the camera is adjusted to rotate to the direction of the target so that the target appears in the field of view of the camera.

Next, the processing receives the preprocessed visual signal context vector and the acquired sound signal context vector from the controllable camera and connects them with a decoder, associates each frame of audio features with the visual features obtained by cross-modal attention, and decodes the relevant features using an attention-based decoder; Then the output probability distribution is given, the speech and visual features are analyzed, and the AVSR system recognition is completed.

The attention weight is calculated in each decoding step, and the weight is dynamically adjusted along with the change of the time modal quality, so that the dryness resistance is strong, and the AVSR system has stronger robustness. Modal attention mechanism is integrated into end-to-end attention AVSR system, and attention weight can be automatically adjusted to a more stable and reliable state according to the quality of a single signal. When speech recognition encounters obstacles, poor recognition results will hinder users from completing tasks. For example, it is difficult to recognize a multi-meaning Chinese language. When this unavoidable error occurs, the system provides corresponding error mechanism, switches to visual mode to understand speech, realizes the complementary effect of video mode on speech recognition, and stores the corresponding relation of recognition results in the "speech feature <-> visual feature correspondence" database.

AVSR is a bimodal task. Auditory and visual information is stored in ASR system. The contribution of video depends on the signal-to-noise ratio of audio signal. The frame rates of audio and video are different to facilitate the use of modal attention mechanism. The recognition effect is much better than that of single mode recognition. In addition, in this system, we try to keep the video signal unchanged and change the speech signal under different signal-to-noise ratios, especially under noisy conditions, in order to explore the complementary effect of visual modes on speech recognition.

3. Design of Perception and Recognition Patterns Based on Multimodal Information

The research of multimodal signal perception theory mainly focuses on the comprehensive perception of multimodal signals and the selective attention to cross-modal transfer between multi-sensory systems. Relevant research has confirmed that the information processing mode of the human brain is the coding representation of the existing statistical sensory information, and the concept and perception of the external world are formed through the combination of multimodal sensory information. As reflected in Pavlov's conditioned reflex test, biological "perception" of external signals is manifold, and human brain's perception of external signals is multimodal.

Behavior signal perception is divided into two sub-models: behavior recognition and state monitoring. After the behavior recognition (multi-classification problem) model is used to perceive the user behavior, a simpler state monitoring (two-classification problem) model will be used to judge
whether the user behavior has changed. Assuming that the user behavior mode to be identified is \( m \), the task of the behavior identification model is to identify which behavior the user is in, which behavior can be characterized by \( m \) classifiers (MC); Different modal information is weighted and summed according to the modal attention mechanism, and the flow model is integrated into the representation of a single mode to complete identification. The task of the state monitoring model is to determine whether the user's behavior state has changed. M Binary Classifier (BC) can be used for detection. The principle is as follows. According to the result of the behavior recognition model, the state monitoring model to be used is determined, and the result of the conditional state monitoring model determines whether to call the behavior recognition model.

\[
\text{flag}^{\text{MC}} = \begin{cases} 
1, \cap_{i=1}^{m} f_{i}^{\text{MC}} = 0 \\
0, \cap_{i=1}^{m} f_{i}^{\text{MC}} \neq 0
\end{cases} 
\]

(1)

\[
\text{flag}^{\text{BC}}_{i} = \begin{cases} 
1, f^{\text{MC}} = i \\
0, f^{\text{MC}} \neq i
\end{cases} 
\]

(2)

Let \( f^{\text{MC}} \) represent the classification result of the \( m \)-class classifier MC, \( f^{\text{MC}} = \{1, 2, \ldots, m\} \). \( f^{\text{BC}} \) represents the classification result of the 2-classifier, \( f^{\text{BC}} = \{0, 1\} \) then the cooperative working mechanism of the behavior recognition model and the state monitoring model is shown in Formulas (1) and (2), Where \( \text{flag}^{\text{MC}} \) represents the cooperative flag of the state monitoring model 2-classifier, and \( \text{flag}^{\text{MC}} \) represents the cooperative flag of the behavior recognition model. The state monitoring determines whether the behavior signal expression of the user is recognized. 0 indicates abnormal identification and 1 indicates normal identification.

4. Conclusion
In this paper, aiming at the shortcomings of single-mode speech emotion recognition, multi-mode fusion mechanism is adopted to improve, voice and visual information are cross-mode fused, and AVSR system is constructed. AVSR is a bimodal task. Auditory and visual information are connected in series with ASR system. The frame rates of audio and video are different to facilitate the use of modal attention mechanism. The modal attention mechanism is integrated into end-to-end attention AVSR. The attention weight can be automatically adjusted to a more stable and reliable state according to the quality of a single signal, with strong dryness resistance and good robustness. Based on AVSR system, a multi-modal signal sensing and recognition model is designed, which provides a theoretical basis for the research of multi-modal signal integrated sensing and selective attention to cross-modal migration between multi-sensory systems.

Acknowledgments
This work was supported by the National Key Research and Development Plan (2019YFB1406400)

References
[1] J. Kittler, Y. Li, J. Matas, and M. Sanchez, Combining evidence in multimodal personal identity recognition systems, in Proc. of Int. Conf. on Audio. and Video-Based Biometric Person Authentication, Crans-Montana, Switzerland, March, 1997, 1206: 327-334.
[2] Brunelli R, Falavigna D. Person identification using multiple cues [J]. IEEE Transactions Oil Pattern Analysis & Machine Intelligence, 1995, 17(10): 955-966.
[3] E Bigun, J Bigun, B Duc, S Fischer. Expert conciliation for multimodal person authentication systems using Bayesian Statistics[C]/The First International Conference on AVBPA, (Crans-Montana, Switzerland), 1997, 27(300): 143-149.
[4] A K Jain, S Prabhakar, S Chen. Combining multiple matchers for a high security fingerprint verification system[J]. Pattern Recognition Letters, 1999, 20: 1371-1379.

[5] A K Jain, S Prabhakar, L Hong. A multichannel approach to fingerprint classification[J]. IEEE Transactions on PAMI, 1999, 21(4): 348-359.

[6] Zhu L Q, Zhang S Y Multimodal biometric identification system based on finger geometry, knuckle print and palm print[J]. Pattern Recognition Letters, 2010, 31(12): 1641-1649.

[7] Kumar A, Zhou Y. Human Identification Using Finger Images[J]. IEEE Transactions on Image Processing, 2012, 21(4): 222-224.

[8] Prasanalakshmi B, Kannammal A Sridevi R. Multimodal Fingerprint Cryptosystem Involving Face, Fingerprint and Palm vein[J]. International Journal of Computer Science Issues, 2011, 8(44): 604-610.