A Review of Audio-Visual Fusion with Machine Learning

Xiaoyu Song1, Hong Chen1*, Qing Wang1, Yunqiang Chen1, Mengxiao Tian1, Hui Tang1

1College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
*Corresponding author's e-mail: norman_chen@263.net

Abstract. For the study of single-modal recognition, for example, the research on speech signals, ECG signals, facial expressions, body postures and other physiological signals have made some progress. However, the diversity of human brain information sources and the uncertainty of single-modal recognition determine that the accuracy of single-modal recognition is not high. Therefore, building a multimodal recognition framework in combination with multiple modalities has become an effective means of improving performance. With the rise of multi-modal machine learning, multi-modal information fusion has become a research hotspot, and audio-visual fusion is the most widely used direction. The audio-visual fusion method has been successfully applied to various problems, such as emotion recognition and multimedia event detection, biometric and speech recognition applications. This paper firstly introduces multimodal machine learning briefly, and then summarizes the development and current situation of audio-visual fusion technology in some major areas, and finally puts forward the prospect for the future.

1. Introduction
In human real life, all kinds of perceptions are accomplished through the integration of multiple information in the brain. For example, reading a text, memorizing words, typing a keyboard, etc., require the brain to sense and integrate information at multiple stages. The brain integrates the information acquired by multiple sensory organs, including hearing, vision, touch, and smell, and finally forms various human perceptions of the outside world. So far, the research on information processing mechanism is mainly the processing of a certain kind of information, which is, mainly considering a single sensing system. In fact, the perception of the world is not only through a single sensing system.

In today’s society, economy and other aspects, the role of images and voice information generated by vision and hearing is becoming more and more prominent. The audio-visual fusion method has been successfully applied to various problems such as emotion recognition, laugh recognition, biometric recognition and speech recognition applications. The addition of visual modalities is particularly useful in noisy environments where the performance of audio classifiers is reduced. Therefore, visual information that is unaffected by acoustic noise can significantly improve the performance of audio-only classifiers in noisy environments [1].

Recently, some deep learning methods for audio-visual fusion have been proposed. This paper firstly introduces multimodal machine learning briefly, and then summarizes the development and current situation of audio-visual fusion technology in some fields, and finally puts forward the prospect for the future.
2. Multimodal machine learning
The source or form of each type of information can be called a modality. For example, people have tactile, auditory, visual, and olfactory; media of information, such as voice, video, text, etc.; a variety of sensors, such as radar, infrared, accelerometers, etc. Each of the above can be referred to as a modality. At the same time, the modal can also have a very broad definition. For example, we can treat two different languages as two modalities, and even the data sets collected in two different situations can be considered as two modalities.

Therefore, MultiModal Machine Learning (MMML) aims to realize the ability to process and understand multi-source modal information through machine learning. At present, the more popular research direction is multimodal learning between image, video, audio and semantics. Many scholars believe that multimodal learning is the real direction of artificial intelligence [2].

Multimodal learning started in the 1970s and went through several stages of development. After 2010, it fully entered the Deep Learning phase. Multimodal information fusion is an important part of multimodal machine learning. Among them, Audio-Visual Fusion aims to fuse image information and sound information in information. This has three advantages:

- Through multimodal Observing phenomena can make more robust predictions.
- Obtaining multiple forms may allow us to obtain supplemental information that is invisible in the individual model.
- Multimodal systems can still operate in the absence of one of the modalities, such as recognizing emotions from visual signals when people are not speaking.

3. Application of audio-visual fusion in emotion recognition
The rise of artificial intelligence has made the field of human-computer interaction become the focus of current research, and emotion recognition has broad application prospects in pattern recognition, and it is also one of the key technologies for human-computer interaction. For the study of single-modal emotion recognition, for example, the research on speech signals, ECG signals, facial expressions, body postures and other physiological signals has made some progress.

However, the diversity of human brain emotional information sources and the uncertainty of single-modal emotion recognition determine that the accuracy of single-modal emotion recognition is not high. Therefore, constructing a multimodal emotion recognition framework in combination with multiple modalities has become an effective means to improve the performance of the emotion recognition framework. The multi-modal fusion strategy in multi-modal emotion recognition research makes full use of the complementary information of multiple modes such as speech and facial expression, and also eliminates the instability of single mode, thus improving the emotion recognition accuracy rate to a certain extent. In many human emotion signals, speech signals and facial expression signals contain most of the emotional information, so most multimodal emotion recognition research focuses on speech facial expression bimodal emotion recognition [3].

With the improvement of computing power of computer hardware, it becomes possible to apply deep learning methods to multimodal emotion recognition research, such as Convolution Neural Network (CNN) and Long Short-Term Memory (LSTM). Various deep learning models such as LSTM are widely used in emotion recognition. In the multi-modal emotion recognition research, according to the stage of merging different modal data, the current multi-modal feature fusion strategy is mainly divided into feature layer fusion [4, 5] and decision layer fusion [6].

Deep learning models are used in both fusion strategies. The method of the decision layer fusion method has the advantages of not requiring strict timing synchronization between the voice signal and the facial expression signal, and solving the problems of different feature reliability of different modes. For example, Sahoo et al. [6] proposed an emotion recognition algorithm based on decision rules for audio-video dual-mode decision-making layer fusion. Firstly, single-modal recognition of speech and facial expressions is performed separately and tested, and then set. Decision rules to fuse audiovisual information at the decision-making level to identify emotions. Literature [7] uses Kernel Entropy Component Analysis (KECA) method and decision-making layer fusion method to study facial
expression and speech bimodal emotion recognition, and achieved high double in two commonly used emotion databases. Modal emotion recognition rate. In [8], the general background model plus the maximum posterior probability method combined with the feature extraction method of openSMILE feature extraction is used to extract the features of the audio and video information and make the decision layer fusion, and then verify it on the eNTERFACE database, and obtain 77.50% recognition rate. However, the decision-making layer fusion often has a large amount of data loss during the training process, which has an impact on the recognition accuracy. Thus, a method of merging some feature layers under the absolute synchronization of the speech signal and the facial expression signal timing is proposed. For example, in [9], the method of CNN plus openSMILE tool is proposed to extract the features of audio and video and some texts, and merge them in the feature layer, and verify the recognition rate of 76.85% on the IEMOCAP database. Tzirakis et al. proposed an end-to-end bimodal feature layer fusion algorithm in which speech and facial expression modalities use the CNN and 50-layer ResNet networks to extract emotional features, and then directly cascade the bimodal features. The two-layer LSTM model is used for timing modeling and then the system is trained end-to-end.

4. Application of audiovisual fusion in multimedia event detection

With the rapid development and popularity of computer science, network and multimedia. Video, due to its richness, intuition and vividness, has become an important multimedia information carrier in our daily life and has been widely used in all aspects of our lives. However, with the advent of the "Internet plus" era, the scale of video data has expanded dramatically. It takes a huge amount of manpower to manually analyze and manage massive amounts of video data. How to efficiently retrieve interesting events in massive video data, ie multimedia event detection (MED) has become a hot research topic in the field of computer vision and video retrieval in recent years.

Video, as a widely used multimedia carrier, is rich in content, and is composed of text, video, sound, image and other multimedia information files. It has its own characteristics of hierarchy, structure and complexity; content-based The retrieval method only uses the visual information of the multimedia, ignoring the effect of other information such as text, sound and the like on improving the video classification and retrieval tasks; and the text-based method relies too much on a large amount of manual annotation data. Based on this, many scholars combined the multimedia video visual, speech feature extraction methods, and feature coding methods to explore key technologies in multimedia event detection tasks based on multimodal features.

At present, the major methods mainly focus on extracting video features and audio features separately, then merging them in the feature layer, and finally making a “three-step” strategy for decision making [10]. For feature extraction, such as in [11], a large number of low-level audio and visual features and advanced information from object detection, speech and video text OCR are evaluated, multi-phase feature fusion strategies are used to combine multiple features, and multi-core learning is used. (MKL) performs feature-level early fusion, uses Bayesian model combination (BayCom) for scoring-level fusion, and uses video-specific weights for weighted average fusion. The results demonstrate that the fusion of low-level audio and video features has strong performance.

It is also an important topic to combine the audio and video features in MED to produce optimal performance. Usually, early fusion and late fusion are two popular combination strategies. The former combines features before classification, which can better capture features. The relationship is also very easy to overfit the training data. Late fusion combines the output of classifiers from different features to better handle overfitting problems, but does not allow the classifier to train all data at the same time. For example, [12] proposed a “dual fusion” scheme that simply combines early fusion and post-fusion to combine their advantages, and the results are reported on the TRECVID MED 2010, MED 2011, UCF50 and HMDB51 data sets. For the MED 2010 dataset, the average minimum normalized detection cost (MMNDC) was 0.49, which exceeded the state-of-the-art level at the time.

5. Application of audio-visual fusion in speech recognition
Audio-visual fusion originated from speech recognition, and speech recognition is also the most widely used field of audio-visual fusion technology. It is inspired by the McGraw effect, the interaction between hearing and vision in the process of speech perception. When a human subject hears a syllable / ba-ba/ while watching a person's lips say / ga-ga/, they will feel the third sound: / da-da/. These results have prompted many researchers from the language community to extend their methods with visual information. Given the salient features of the hidden Markov model (HMM) in the speech community at the time, many of the early models of AVSR were based on various HMM extensions [13], which is not surprising. Although the research on AVSR is not common afterwards, the deep learning community has once again attracted people's interest.

Although AVSR's original target was to improve speech recognition performance (e.g., word error rate) in all cases, experimental results indicate that the main advantage of visual information is when the speech signal is noisy. In other words, the interactions captured between patterns are complementary rather than complementary. Both capture the same information and improve the robustness of the multimodal model, but initially did not improve speech recognition performance in no-noise scenarios.

At present, only the audio information is used to recognize that the speech has reached an accuracy of more than 95%, but this is the precision that can be achieved in some quiet environments. In a noisy environment, the recognition accuracy of the current ASR system will drop significantly. It is necessary to use visual information to improve the robustness of recognition.

Understanding speech only through visual signals (lip reading) has also attracted decades of research interest. Many researchers are dedicated to extracting powerful visual features for accurate lip-reading models. These feature extraction methods include a top-down method and a bottom-up method. The former uses a priori lip representations in the model, such as Active Shape Model (ASM) and Active Look Model (AAM), and extracts model-based features. The latter directly estimates the visual features of the image by discrete cosine transform (DCT), principal component analysis (PCA) and discrete wavelet transform (DWT). Due to the development of neural networks in recent years, automatic learning of visual features by supervised training of convolutional neural networks (CNN) or LSTM has significantly improved performance [14]. Like the development of ASR, LipNet [15] uses CNN and LSTM and CTC loss to train the neural network to be fully end-to-end and to operate sentence-level lip reading, but because the data set used by LipNet is too small, the results are very limited. It has not been widely recognized. A lot of work has been done recently on the broader data set, [16], whose goal is to identify phrases and sentences spoken by people with conversations, with or without audio. Unlike previous works that focus on identifying a limited number of words or phrases, they use lip reading as an open world issue with only unconstrained natural language sentences and wild video. They proposed a novel dual attention mechanism that surpassed the performance of all previous standard lip-reading benchmark datasets and also demonstrated that visual information can help improve speech recognition performance even with audio.

With the enhancement of computing power, many methods of audio-visual fusion based on deep learning have been proposed. Most of them use a two-step approach, which first extracts features from audio and visual modalities and then inputs them into the classifier. [17] Apply principal component analysis (PCA) to the region of interest (ROI) and spectrogram, and train a depth autoencoder to extract bottleneck features. The features of the entire utterance are fed to a support vector machine (SVM), ignoring the temporal dynamics of the speech. A similar approach was used [18] where PCA was applied to the oral ROI and spectrogram, and the recurrence time multimode was trained to limit the Boltzmann machine to extract features fed to the SVM. [19] The connected MFCC and the scattering coefficients extracted from the port ROI are used to train a deep network with a bilinear softmax layer. [20] The Tobeck features were extracted from the lip image and the Mel map using a convolutional neural network, and these images were sent to the HMM. Obviously, none of the above work follows the end-to-end architecture.

Until [21] proposed an end-to-end audiovisual model based on bidirectional long-term and short-term memory (BLSTM) network, which is the first end-to-end audio-visual fusion model. It simultaneously learns to extract features directly from pixels and spectra, and Classification of speech
and non-verbal vocalization. The model consists of multiple identification streams, one for each modality, and features are extracted directly from the oral region and the spectrogram. The temporal dynamics in each stream/modality are modeled by BLSTM, and multiple stream/modal fusions are performed through another BLSTM.

The rise of the attention model has brought a new direction to AVSR. For example, [22] proposed a new audiovisual speech recognition method based on multimodal attention, which can automatically learn the fusion representation of the two methods according to its importance. The method is implemented using the state-of-the-art sequence-to-sequence (Seq2seq) architecture, and the key contribution is the use of an additional attention mechanism that allows the model to automatically adjust its modal attention to make it more reliable. The experimental results show that the end-to-end system equipped with modal attention can outperform the pure audio system and increase by 36% at 0dB SNR, showing the effectiveness of its proposal.

There are also ways to get progress from different data sets. In [23], there are two problems with AVSR. The first is that they are mostly based on 2D audiovisual (AV) corpus with lower video sampling rate. In order to solve this problem, They introduced a 3D AV dataset with a higher video sampling rate (up to 100 Hz). Another problem is the need for auditory and visual modes during system testing. In order to solve this problem, a dual-mode convolutional neural network (CNN) framework based on visual feature generation is proposed to construct an AVSR system with wider application. In this framework, the Long-Term Memory Recurrent Neural Network (LSTM-RNN) is used to generate visual modalities from auditory modalities, while CNN is used to integrate these two modalities. In the Mandarin Chinese far field speech recognition task, a significant average character error rate (CER) reduction of approximately 27% was obtained relative to the audio CNN baseline only when the visual modality was provided. When the visual modality is not available, the proposed AVSR system using visual feature generation techniques is superior to 18.51% of the audio-only CNN baseline versus CER.

In general, with the development of speech recognition, audio modality alone has been unable to obtain more robustness in noisy environments, and this is the direction in which speech recognition is currently moving, so visual information is used to improve recognition. Robustness is necessary.

6. Conclusion
In recent years, audio-visual fusion technology has received extensive attention from researchers. Driven by deep learning technology, audio-visual fusion technology has shown strong development momentum, which is greatly improved in speech recognition, emotion recognition and multimedia event analysis. However, it can be seen that there are still many problems in audio-visual fusion technology waiting for solution. From the feature fusion method, how to integrate, when to fuse and what to integrate these three issues need to do more researches. From the training model, a better model is needed to deal with the modal incompleteness, data processing Real-time, modal data imbalance; from the data set, a lot of work is required to provide a reliable and universal data set.

Although there are still many bottlenecks in audio-visual fusion technology, a large number of applied researches and analysis show that it has great application prospects. In the next few years, the application of audio-visual fusion technology, especially in speech recognition, will emerge in large numbers. Although its performance is not obvious in some environments compared with traditional single-mode speech recognition, it is highly popular.

In the future development, audio-visual fusion technology with more stable performance will create a more stable experience for human beings, and the performance of the trained model will be closer to the level of real humans. Future audio-visual fusion technology will greatly change the field of deep learning, which is an inevitable trend of technological development.

References
[1] Guanjun, S., Z. Shudong and W. Feigao.(2019)Research on audio and video bimodal emotion recognition fusion framework. Computer Engineering and Applications: 1-9(in Chinese).
[2] Baltrusaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: a survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1-1.

[3] Atrey, P. K., Hossain, M. A., Saddik, A. E., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. Multimedia Systems, 16(6):345-379.

[4] Sarkar, C., Bhatia, S., Agarwal, A., & Li, J. (2014). Feature Analysis for Computational Personality Recognition Using YouTube Personality Data set. In: Multi Media. pp.11-14.

[5] Poria, S., Cambria, E., Hussain, A., & Huang, G. B. (2015). Towards an intelligent framework for multimodal affective data analysis. Neural Networks, 63:104-116.

[6] Sahoo, S., & Routray, A. (2017). Emotion recognition from audio-visual data using rule based decision level fusion. In: Technology Symposium. IEEE.

[7] Xie, Z., & Guan, L. (2013). Multimodal Information Fusion of Audiovisual Emotion Recognition Using Novel Information Theoretic Tools. In: IEEE International Conference on Multimedia & Expo. IEEE.

[8] Dobrišek, S., et al., (2013) Towards Efficient Multi-Modal Emotion Recognition. International Journal of Advanced Robotic Systems., 10(1): 53.

[9] Poria, S., Chaturvedi, I., Cambria, E., & Hussain, A. (2017). Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. In: IEEE International Conference on Data Mining. IEEE.

[10] Shizhe, H., Multimedia event detection based on multimodal features. 2018, Beijing University of Posts and Telecommunications. p. 3-4 (in Chinese).

[11] Natarajan, P., Prasad, R., Park, U., Tsakalidis, S., Zhuang, X., & Vitaladevuni, S., et al. (2012). Multimodal feature fusion for robust event detection in web videos. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society.

[12] Lan, Z. Z., Bao, L., Yu, S. I., & Wei Liu. (2014). Multimedia classification and event detection using double fusion. Multimedia Tools and Applications, 71(1):333-347.

[13] Brand, M., N. Oliver and A. Pentland. (1997) Coupled hidden Markov models for complex action recognition. : IEEE.

[14] Tatulli, E., & Hueber, T. (2017). Feature extraction using multimodal convolutional neural networks for visual speech recognition. In: IEEE International Conference on Acoustics. IEEE.

[15] Assael, Y. M., Shillingford, B., Whiteson, S., & De Freitas, N. (2016). Lipnet: end-to-end sentence-level lipreading.

[16] Chung, J. S., & Zisserman, A. (2016). Lip Reading in the Wild. In: Asian Conference on Computer Vision. Springer, Cham.

[17] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. (2011) Multimodal Deep Learning. In: ICML.

[18] Hu, D., Li, X., & Lu, X. (2016). Temporal Multimodal Learning in Audiovisual Speech Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp.3574-3582.

[19] Mroueh, Y., Marcheret, E., & Goel, V. (2015). Deep multimodal learning for audio-visual speech recognition.

[20] Y. Takashima, R. Aihara, T. Takiguchi, Y. Ariki, N. Mitani, K. Omori, and K. Nakazono. (2016) Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss. In: Interspeech 2016, pp. 277–281.

[21] Petridis, Stavros & Wang, Yuji & Li, Zuwei & Pantic, Maja. (2017). End-to-End Audiovisual Fusion with LSTMs.

[22] Zhou, P., Yang, W., Chen, W., Wang, Y., & Jia, J. (2018). Modality attention for end-to-end audio-visual speech recognition.

[23] Su, R., Wang, L., & Liu, X. (2018). Multimodal learning using 3D audio-visual data for audio-visual speech recognition. In: 2017 International Conference on Asian Language Processing (IALP). IEEE.