# MULTIMODAL REPRESENTATION LEARNING: ADVANCES, TRENDS AND CHALLENGES

**SU-FANG ZHANG[1], JUN-HAI ZHAI[2], BO-JUN XIE[2], YAN ZHAN[2], XIN WANG[2]**

[1]Hebei Branch of China Meteorological Administration Training Center, China Meteorological Administration,
Baoding 071000, China
[2]Hebei Key Laboratory of Machine Learning and Computational Intelligence, College of Mathematics and
Information Science, Hebei University, Baoding, 071002, Hebei, China
E-MAIL: mczjh@126.com

**Abstract:**

**Representation learning is the base and crucial for consequential tasks, such as classification, regression, and recognition. The goal of representation learning is to automatically learning good features with deep models. Multimodal representation learning is a special representation learning, which automatically learns good features from multiple modalities, and these modalities are not independent, there are correlations and associations among modalities. Furthermore, multimodal data are usually heterogeneous. Due to the characteristics, multimodal representation learning poses many difficulties: how to combine multimodal data from heterogeneous sources; how to jointly learning features from multimodal data; how to effectively describe the correlations and associations, etc. These difficulties triggered great interest of researchers along with the upsurge of deep learning, many deep multimodal learning methods have been proposed by different researchers. In this paper, we present an overview of deep multimodal learning, especially the approaches proposed within the last decades. We provide potential readers with advances, trends and challenges, which can be very helpful to researchers in the field of machine, especially for the ones engaging in the study of multimodal deep machine learning.**

**Keywords:**

**Multimodal; Representation learning; Machine learning; deep learning; Multimodal deep learning**

## 1. Introduction

The goal of representation learning is to automatically learn good representations of raw data [1], which make it easier to extract useful information or features for building learning models, such as classifiers, regressor, and recognizer. Obviously, the representation learning is a crucial data preprocessing step of machine learning, it directly determines the performance of learning models. In the age of big data, the data used for describing a same event or phenomenon is usually come from multiple sources, different source data may be different modality, may be audio, video, text, and so on. Consequently, the representation learning of multimodal data is an interesting topic and is of new big challenges mainly due to the heterogeneities.

The general strategy for processing multimodal data is fusion [2, 3], multimodal data fusion can be made in data space, also can be made in feature space. The former is called raw data fusion or early data fusion, the latter is called feature fusion or intermediate data fusion. The early data fusion directly integrates the raw data from the sensor, which is quite challenging due to some factors [3], for instance, (1) it is difficult to determine an appropriate sampling rate between different sensors; (2) the synchronized data from multiple data sources might not be available. The intermediate data fusion can be viewed as an improvement of the raw data fusion, it is a hot research topic in deep learning due to wide applications, and has attracted much attention in recent years, many methods have been proposed, these approaches can be roughly classified two categories: joint representation approaches and coordinated representation approaches. Although there are a few review articles on the multimodal machine learning [2-4], the review papers are all based on the perspective of multimodal data fusion. In this paper, we attempt to provide the potential readers with a comprehensive survey of multimodal representation learning from two aspects: applications and algorithms (see figure 1). In addition to surveying the recent advances, we also present the trends and challenges of multimodal representation learning. This paper can provide researchers engaged in related works with very valuable help.

The paper is organized as follows. The applications which drive algorithm the designs for multimodal representation learning are reviewed in Section 2, the algorithms which address application problems are surveyed in Section 3. Section 4 concludes this paper, in addition, the trends and challenges of multimodal representation learning are also summarized in Section 4.

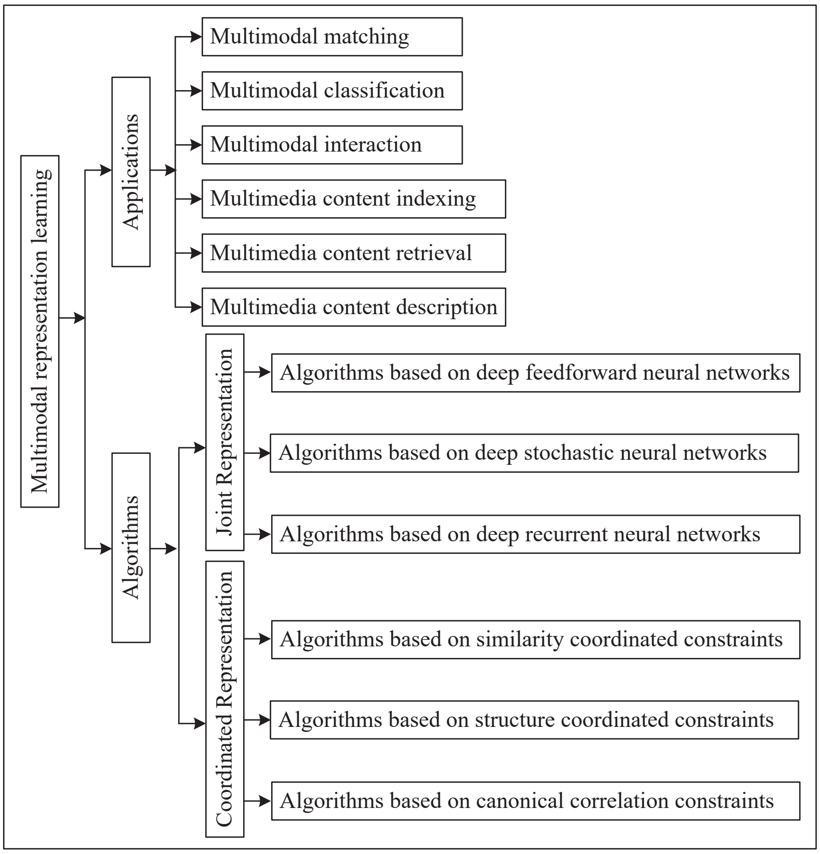
Multimodal representation learning
Applications
Multimodal matching
Multimodal classification
Multimodal interaction
Multimedia content indexing
Multimedia content retrieval
Multimedia content description
Algorithms
Joint Representation
Algorithms based on deep feedforward neural networks
Algorithms based on deep stochastic neural networks
Algorithms based on deep recurrent neural networks
Coordinated Representation
Algorithms based on similarity coordinated constraints
Algorithms based on structure coordinated constraints
Algorithms based on canonical correlation constraints

**FIGURE 1.** Applications and algorithms of multimodal representation learning

## 2. Applications-driving algorithm design

Applications drive the designs of various multimodal representation learning algorithms, the main applications include multimodal matching, multimodal classification, multimodal interaction, multimedia context indexing, multimedia context retrieval and multimedia context description (see figure 1).

The goal of multimodal matching is to establish relationship across different modalities such as image and text. Zhao et al. [5] propose a matching method named MRLS (Multimodality Robust Line Segment). The proposed method includes two steps: the first step generates the MRLS descriptors which is based on extracted highly equivalent corners and line segments for two multimodal images; the second step performs image matching by measuring the similarity of corresponding descriptors over two images is performed. Zhang et al. [6] proposed a multimodal matching approach which integrates the feature extraction and the matching in a unified framework using convolutional neural network (CNN). The CNN is trained for minimizing multiclass classification loss, each modal data is viewed as a class. In the study of cross-modal matching, Pitts et al. [7] found that even minor variations can lead to significant differences in match values and intraindividual match variability.

In the multimodal classification, Choi and Lee [8] proposed a deep learning-based multimodal fusion architecture for addressing classification problem, the proposed method have two advantages: (1) it has compatibility with any kind of learning models; (2) it can effectively solve the problem of modal loss. Bahrampour et al. [9] proposed a multimodal task-driven dictionary learning algorithm, which add joint sparsity constraint to enforce collaborations among multiple data. The appealing merit of this algorithm is that the multimodal dictionaries and their corresponding classifiers can be learned simultaneously. Jafari et al. Gomez-Chova et al. [10] provided a comprehensive survey for multimodal classification of remote sensing images. The most valuable content lies in illustrating the different approaches in seven challenging remote sensing applications, the interested readers may refer to the details of the seven challenges in [10].

Regarding multimodal interaction, just as Turk said that people naturally interact with the world multimodally [11]. With the rapid popularization of powerful mobile devices and various sensors in recent years, such as mobile phone, mobile pad and wearable device, multimodal interaction is growing in importance due to advances in hardware and software. The goal of multimodal interaction is to enable users to communicate with computer via text, audio, video, and other modalities. Vidakis et al. [12] presented a multimodal framework which facilitates deployment of a vast variety of modalities in blended learning environment. Mi et al. [13] proposed a deep CNN based architecture to learn the human-centered object affordance, and based on the affordance, a multimodal fusion framework is proposed to realize intended object grasping.

Multimodal indexing and multimodal retrieval are closely related, multimodal indexing is used to accelerate multimodal retrieval. Hu et al. [14] introduced the idea of generative adversarial networks (GANs) [15] for cross-modal retrieval and proposed a multimodal adversarial network (MAN) which consists of multiple modality-specific generators, a discriminator and a multimodal discriminant analysis loss. The technical route of MAN is to project the multimodal data into a common space wherein the similarities between different modalities can be directly computed by the same distance measurement. Shang et al. [16] combined GANs and dictionary learning for cross-modal retrieval, propose a novel framework termed DLA-CMR (Dictionary Learning Algorithm for Cross-Modal Retrieval). The adversarial learning mines the statistical characteristics for each modality, while dictionary learning serves as feature re-constructor to reconstruct discriminative features. Cao et al. [17] proposed an approach of hybrid representation learning for cross-modal retrieval, the proposed method consists three steps: (1) deep RBM (Restricted Boltzmann Machines) is employed to extract the modality-specific features; (2) a joint autoencoder and a feedforward neural net are used for learning hybrid representation; (3) stacked bimodal autoencoders are used to obtain the final shared representation for each modality. A unified framework for multimodal retrieval can be found in [18].

With the advent of depth generation model, multimodal description is a new application, some representative works summarize as follows. Park et al. [19] addressed the problem of personalized image captioning, and solved two post automation tasks in social networks, hashtag prediction and post generation. The former predicts a list of hashtags for an image, while the latter creates a natural text as the caption of the image. Niu et al. [20] investigated the problem of image annotation on two aspects: (1) multimodal feature representation for image annotation; (2) image annotation with the optimal number of class labels, and proposed a approach of multimodal multiscale deep learning for large-scale image annotation. Zhao et al. [21] proposed a multimodal fusion approach for image captioning. Some researchers extended image caption to video caption, for instance, Wu et al. [22] proposed a hierarchical attention-based multimodal fusion for video captioning, while Chou et al. [23] proposed a method of multimodal video-to-near-scene annotation.

In addition to the applications of multimodal representation learning summarized above, there are some other applications, such as the applications in multisensory systems, the ones in health, and the ones in medical image processing, etc. Due to the limitation of pages, the interested reader can refer the references [1] and [2].

## 3. Algorithms-addressing application problems

In the past decade, many multimodal representation learning algorithms have been proposed by different researchers. These algorithms can be roughly classified into two categories: joint representation algorithms and coordinated representation algorithms. Both categories can be further classified into three categories respectively (see figure 1).

Joint representation combines the unimodal data into the same representation space by projecting unimodal data together into a shared space, the diagram of joint representation learning is given in figure 2.

In the framework of multimodal representation learning based on deep feedforward neural networks (DFNN), the pioneering work was contributed by Ngiam et al. [24]. The DFNN model used in [24] is deep stacked autoencoder that is employed to automatically learning features from multimodal data. The idea of [24] is simple, which is diagramed in figure 3. Along this technical route, Shekhar et al. [25] proposed a joint sparse representation algorithm, and used for recognition of multimodal biometrics data. Based on multi-fusion deep neural networks, Gu et al. [26] proposed an approach for learning joint multimodal representation.

In the framework of multimodal representation learning based on deep stochastic neural networks (DSNN), the pioneering algorithm was proposed by Srivastava and

Salakhutdinov [27]. The DSNN in [27] is deep Boltzmann machines which is used for learning good features. Along this technical route, Sohn et al. [28] proposed an improved algorithm by variation of information, Amer et al. [29] proposed a hybrid approach for deep multimodal data fusion.

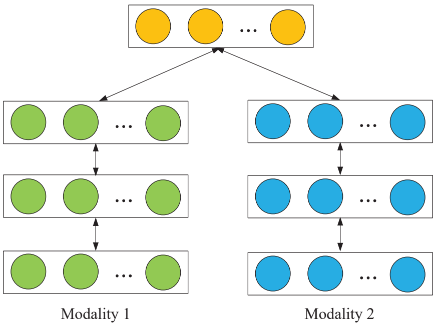

**FIGURE 2.** The diagram of joint representation learning

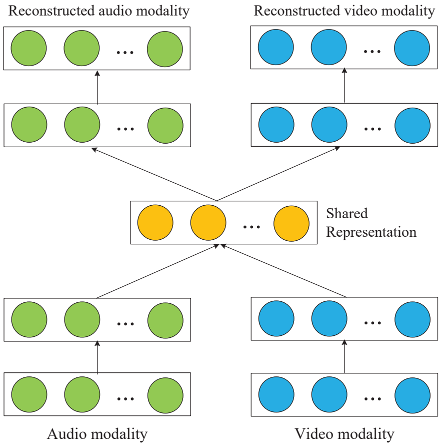

**FIGURE 3.** the idea of multimodal deep learning in [27]

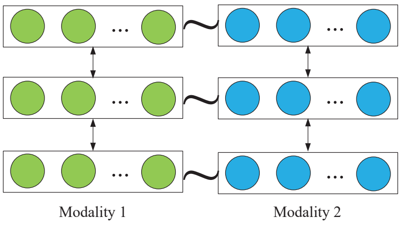

**FIGURE 4.** The diagram of coordinated representation learning

In the framework of multimodal representation learning based on deep recurrent neural networks (DRNN), the representative works include: Rajagopalan et al. [30] extended long short-term memory networks to multimodal representation learning; Feng et al. [31] employed multimodal recurrent neural networks to audio visual speech multimodal recognition; Abdulnabi et al. [32] proposed a multimodal recurrent neural networks with information transfer layers to label indoor scene.

Coordinated representation handles unimodal signals separately, but enforce certain similarity constraints on them to bring them to what we term a coordinated space [1], the diagram of coordinated representation learning is given in figure 4.

In the framework of multimodal representation learning based on similarity coordinated constraints. Generally, the similarity between coordinated spaces is measured by distance between modalities. The seminal work of this category comes from Weston et al. [33] on the WSABIE (web scale annotation by image embedding). In this work, a coordinated space constrained by similarity was constructed for images and their annotations. In deep learning framework, Frome et al. [34] proposed a model named DeViSE (Deep VisualSemantic Embedding). Kiros et al. [35] extended this model by using an LSTM model and a pairwise ranking loss to coordinate feature spaces. A similar model was also proposed by Pan et al. [36], the difference is image modality is replaced with video modality.

In the framework of multimodal representation learning based on structure coordinated constraints. The pioneering work come from Bronstein et al. [37]. In [37], the authors employed similarity-sensitive hashing to enforce the constraint of structure coordinate, but the hashing function is based on similarity-sensitive rather than based on global-learning. Kumar and Udupa [38] proposed an approach for learning hash function for cross-view similarity search, while Jiang and Li [39] extended this method to deep learning scenario.

In the framework of multimodal representation learning based on canonical coordinated constraints. This kind of multimodal representation learning methods use CCA (Canonical Correlation Analysis) to extract relevant and significant features for sequent tasks. For example, Mandal and Maji [40] proposed a feature extraction algorithm named FaRoC which integrates judiciously the merits of canonical correlation analysis (CCA) and rough sets. Yu et al. [41] proposed a category-based deep CCA for fine-grained venue discovery from multimodal data. Based on bimodal/multimodal hybrid centroid CCA, Elmadany et al. [42] proposed a multimodal feature learning approach, and applied to human action recognition.

## 4. Conclusions

This paper presented a comprehensive survey on multimodal representation learning from two aspects: applications and algorithms. The applications drive the studies of multimodal representation learning by different researchers engaged in related works. While various algorithms of multimodal representation learning are designed for solving practical application problems. The authors think that the trends and challenges of multimodal representation learning include the following five aspects:

(1) Handle the problem of modality miss via deep generative models, such as generative adversarial networks, variational auto-encoders.

(2) Multimodal similarity preserving hashing via deep representation learning.

(3) How to measure the correlation between different modalities?

(4) How to learn the joint distributions of different modalities?

(5) Given one modality, how to learn the conditional distribution of another modality?

## Acknowledgments

This research is supported by the natural science foundation of Hebei Province (F2017201026), by the advanced talents incubation program of the Hebei University, by the youth foundation of Hebei Province higher school science and technology research project (QN2018251), and by the postgraduate's innovation fund project of Hebei University (hbu2019ss077).

Corresponding author: JUN-HAI ZHAI

## References

[1] Y. Bengio, A. Courville, P. Vincent. Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(8):1798-1828.

[2] D. Lahat, T. Adali, C. Jutten. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. Proceedings of the IEEE, 2015, 103(9):1449-1477.

[3] Y. Zheng. Methodologies for Cross-Domain Data Fusion: An Overview. IEEE Transactions on Big Data, 2015, 1(1):1-14.

[4] D. Ramachandram, G. W. Taylo. Deep Multimodal learning: a survey on recent advances and trends. IEEE Signal Processing Magazine, 2017, 34(6):96-108.

[5] C. Zhao, H. Zhao, J. Lv, et al. Multimodal image matching based on Multimodality Robust Line Segment Descriptor. Neurocomputing, 2016, 177:290-303.

[6] Y. Zhang, Y. Gu, X. Gu. Two-Stream Convolutional Neural Network for Multimodal Matching. International Conference on Artificial Neural Networks (ICANN18), 2018, Pages:14-21.

[7] B. Pitts, S. L. Riggs, N. Sarter. Crossmodal Matching: A Critical but Neglected Step in Multimodal Research, IEEE Transactions on Human-Machine Systems, 2016, 46(3):445-450.

[8] J. H. Choi, J. S. Lee. EmbraceNet: A robust deep learning architecture for multimodal classification. Information Fusion, 2019, 51:259-270.

[9] S. Bahrampour, N. M. Nasrabadi, A. Ray, et al. Multimodal Task-Driven Dictionary Learning for Image Classification. IEEE Transactions on Image Processing, 2015, 25(1):24-38.

[10] L. Gomez-Chova, D. Tuia, G. Moser, et al. Multimodal Classification of Remote Sensing Images: A Review and Future Directions. Proceedings of the IEEE, 2015, 103(9):1-25.

[11] M. Turk. Multimodal interaction: A review. Pattern Recognition Letters, 2014, 36:189-195.

[12] N. Vidakis, K. Konstantinos, G. Triantafyllidis. A Multimodal Interaction Framework for Blended Learning. International Conference on Interactivity, Game Creation, Design, Learning, and Innovation, Pages:2016, 205-211.

[13] J. Mi, S. Tang, Z Deng, et al. Object affordance based multimodal fusion for natural Human-Robot interaction. Cognitive Systems Research, 2019, 54:128-137.

[14] P. Hu, D Peng, X. Wang, et al. Multimodal adversarial network for cross-modal retrieval. Knowledge-Based Systems, online first, 2019, https://doi.org/10.1016/j.knosys.2019.05.017.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. Generative adversarial nets. In: Proceedings of the 2014 Conference on Advances in Neural Information Processing Systems 27. Montreal, Canada: Curran Associates, Inc., 2014. 2672-2680.

[16] F. Shang, H. Zhang, L. Zhu, et al. Adversarial cross-modal retrieval based on dictionary learning. Neurocomputing, 2019, online first, https://doi.org/10.1016/j.neucom.2019.04.041.

[17] W. Cao, Q. Lin, Z. He, et al. Hybrid representation learning for cross-modal retrieval. Neurocomputing, 2019, 345:45-57.

[18] D. Rafailidis, S. Manolopoulou, P. Daras. A unified framework for multimodal retrieval. Pattern Recognition, 2013, 46:3358-3370.

[19] C. C. Park, B. Kim, G. Kim. Towards Personalized

Image Captioning via Multimodal Memory Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(4):999-1012.

[20] Y. Niu, Z. Lu, J. R. Wen, et al. Multi-Modal Multi-Scale Deep Learning for Large-Scale Image Annotation. IEEE Transactions on Image Processing, 2019, 28(4):1720-1731.

[21] D. Zhao, Z. Chang, S. Guo. A multimodal fusion approach for image captioning. Neurocomputing, 2019, 329:476-485.

[22] C. Wu, Y. Wei, X. Chu, et al. Hierarchical attention-based multimodal fusion for video captioning. Neurocomputing, 2018, 315:362-370.

[23] C. L. Chou, H. T. Chen, S. Y. Lee. Multimodal Video-to-Near-Scene Annotation. IEEE Transactions on Multimedia, 2017, 19(2):354-366.

[24] J. Ngiam, A. Khosla, M. Kim, et al. Multimodal deep learning. in Proc. 28th Int. Conf. Machine Learning (ICML-11), 2011, pp. 689-696.

[25] S. Shekhar, V. M. Patel, N. M. Nasrabadi, et al. Joint Sparse Representation for Robust Multimodal Biometrics Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(1):113-126.

[26] Z. Gu, B. Lang, T. yue, et al. Learning Joint Multimodal Representation Based on Multi-fusion Deep Neural Networks. International Conference on Neural Information Processing, 2017, pp.276-285.

[27] N. Srivastava, R. R. Salakhutdinov. Multimodal learning with deep Boltzmann machines. in Proc. Advances in Neural Inform. Processing Syst., 2012, pp. 2222-2230.

[28] K. Sohn, W. Shang, H. Lee. Improved multimodal deep learning with variation of information. in Proc. Advances in Neural Information Processing Systems., 2014, pp. 2141-2149.

[29] M. R. Amer, T. Shields, B. Siddiquie, et al. Deep Multimodal Fusion: A Hybrid Approach. International Journal of Computer Vision, 2018, 126(2-4):440-456.

[30] S. S. Rajagopalan, L. P. Morency, T. Baltrusaitis, et al. Extending long short-term memory for multi-view structured learning. in Proc. Eur. Conf. Comput. Vis., 2016, pp. 338-353.

[31] W. Feng, N. Guan, Y. Li, et al. Audio visual speech recognition with multimodal recurrent neural networks. 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14-19 May 2017, 681-688.

[32] A. H. Abdulnabi, B. Shuai, Z. Zuo, et al. Multimodal Recurrent Neural Networks with Information Transfer Layers for Indoor Scene Labeling. IEEE Transactions on Multimedia, 2018, 20(7):1656-1671.

[33] J. Weston, S. Bengio, N. Usunier. WSABIE: Scaling up to large vocabulary image annotation. in Proc. 7th Int. Joint Conf. Artif. Intell., 2011, pp. 2764-2770.

[34] A. Frome, G. Corrado, J. Shlens. DeViSE: A deep visualsemantic embedding model. in Proc. 28th Int. Conf. Neural Inf. Process. Syst., 2013, pp. 2121-2129.

[35] R. Kiros, R. Salakhutdinov, R. S. Zemel. Unifying visualsemantic embeddings with multimodal neural language models. Trans. Assoc. Comput. Linguistics, 2015, pp. 1-13.

[36] Y. Pan, T. Mei, T. Yao, et al. Jointly modeling embedding and translation to bridge video and language. in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 4594-4602.

[37] M. M. Bronstein, A. M. Bronstein, F. Michel, et al. Data fusion through cross-modality metric learning using similarity-sensitive hashing. in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2010, pp. 3594-3601.

[38] S. Kumar, R. Udupa. Learning hash functions for cross-view similarity search. in Proc. 7th Int. Joint Conf. Artif. Intell., 2011, pp. 1360-1365.

[39] Q. Y. Jiang, W. J. Li. Deep cross-modal hashing. in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 3270-3278.

[40] A. Mandal, P. Maji. FaRoC: Fast and Robust Supervised Canonical Correlation Analysis for Multimodal Omics Data. IEEE Transactions on Cybernetics, 2018, 48(4):1229-1241.

[41] Y. Yu, S. Tang, K. Aizawa, et al. Category-Based Deep CCA for Fine-Grained Venue Discovery from Multimodal Data. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(4):1250-1258.

[42] N. E. D. Elmadany, Y. He, L. Guan. Multimodal Learning for Human Action Recognition Via Bimodal/Multimodal Hybrid Centroid Canonical Correlation Analysis. IEEE Transactions on Multimedia, 2019, 21(5):1317-1331.