COLUMN: EIC MESSAGE

Embracing Multimodal Data in Multimedia Data Analysis

Shu-Ching Chen, *Florida International University, Miami, FL, 33199, USA*

Multimedia data analysis is at the heart of multimedia research, where the patterns, information, and knowledge are mined and recognized from the multimedia data for various applications. We have witnessed significant performance improvements in multimedia data analysis driven by the recent advances in artificial intelligence and deep learning. While most of the existing methods focus on analyzing unimodal data, many novel multimodal multimedia data analysis techniques have also been developed and achieved better performance than those using unimodal data. Multimodal multimedia data has proven to be effective to improve the performance and reliability of data analysis systems and applications. It is also important to match and retrieve data with similar contents across modalities. However, due to the highly complex nature of multimodal multimedia data, further research in multimodal multimedia data analysis is needed to address the challenges including missing modalities, lack of benchmark datasets, and so on.

Multimedia data analysis aims to recognize, mine, and extract patterns, information, and knowledge from the multimedia data using computational methods and is one of the most important directions in multimedia research. Due to the recent advances in artificial intelligence and deep learning, multimedia data analysis algorithms have been able to achieve the same or even better performance than human beings in many critical applications.[1] Most of the current algorithms and models process and analyze unimodal data while multimedia data such as videos can be naturally considered as multimodal data, which contains visual, auditory, and textual information. It has been shown that integrating multiple modalities of multimedia data is beneficial to understanding the semantics and extracting knowledge from the data.[2] Meanwhile, additional modalities can also be derived from and fused with the original multimedia data to enhance the recognition of specific aspects in the data. For example, optical flow is commonly used to improve the understanding of motions in the videos and the fusion of optical flow and raw visual streams can enhance the performance of video-based action recognition significantly.[3]

In addition to the natural cooccurrence of multiple modalities in multimedia data, multimodal data have become more popular due to the increased availability and accessibility of the advanced sensor technologies. Multiple heterogeneous sensors have been deployed to collect synchronized multimodal data. For example, it is common to deploy both camera and LiDAR sensors simultaneously on the autonomous driving cars. Recent studies have shown that multimodal multimedia data analysis can be key to enhancing the model performance and building robust and reliable systems for many critical applications like intelligent healthcare,[4] autonomous driving,[5] and so on, where complementary information carried by the data of various modalities could be integrated to improve the understanding about the situations and avoid the misinterpretations of a sole modality. To this end, attention architectures[6,7] and subspace analysis[8] have been proposed to effectively fuse information from various modalities to extract knowledge from the multimodal multimedia data.

While multimodal multimedia data fusion techniques explore the complementary information among all the modalities, another important problem in multimodal multimedia data analysis is to learn the joint

1070-986X © 2021 IEEE
Digital Object Identifier 10.1109/MMUL.2021.3104911
Date of current version 14 September 2021.

representations that reflect semantic and content similarity across modalities. Such joint representations are critical for matching information from different modalities and can be leveraged in cross-modal retrieval tasks such as *ad hoc* video search[9] and visual text comprehension.[10] Deep neural network models are trained with representation learning or contrastive learning strategies[10, 11] to ensure that the representations of data with similar semantic meaning to be close to each other regardless of the modalities.

Most of the multimodal multimedia data analysis techniques assume that all the modalities are present simultaneously. However, it is common that some sensors fail to operate, or some modalities are missing in practice. Therefore, it is important to allow the modal to handle missing modalities. The generative models[12] and joint representations[13] have been leveraged to tackle missing modalities in the testing data. Recently, Ma *et al.* proposed the SMIL algorithms to allow for missing modalities among training data to further enhance the feasibility of multimodal multimedia data analysis.[14] However, the performance degradation for data with missing modalities remains significant. Further research to improve the model performance when modalities are missing should be conducted.

Another challenge in multimodal multimedia data analysis is the lack of large-scale benchmark datasets, which limits the proper comparisons among various multimodal multimedia data analysis techniques and the monitoring of their advancements. Recently, many high-quality multimodal datasets have been developed and released to address the problem.[5,15,16] However, large-scale public benchmark datasets remain missing for many critical applications of multimodal multimedia data analysis. Moreover, it is more costly to build such multimodal multimedia datasets since all the modalities are required to be properly synchronized during data collection and the annotations and labels will need to be generated based on all the modalities. To mitigate this problem, advanced tools for multimodal data acquisition and annotations should be developed.

Multimodal data have proven to be effective to enhance the performance and reliability of multimedia data analysis for various applications. Many novel techniques in multimodal multimedia data analysis have been proposed to effectively fuse data of various modalities. However, more research is still needed to address the challenges such as missing modalities and building large-scale multimodal dataset, which can further improve the feasibility and usability of multimodal multimedia data analysis in various applications.

REFERENCES

1. S. Pouyanfar *et al.*, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Comput. Surv.*, vol. 51, no. 5, 2018, Art. no. 92, doi: 10.1145/3234150.
2. H. Tian, Y. Tao, S. Pouyanfar, S.-C. Chen, and M.-L. Shyu, "Multimodal deep representation learning for video classification," *World Wide Web*, vol. 22, no. 3, pp. 1325–1341, 2019, doi: 10.1007/s11280-018-0548-3.
3. Z. Gao, L. Guo, T. Ren, A.-A. Liu, Z.-Y. Cheng, and S. Chen, "Pairwise two-stream convnets for cross-domain action recognition with small data," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, 2020, doi: 10.1109/TNNLS.2020.3041018.
4. Y. Wang *et al.*, "Robust content-adaptive global registration for multimodal retinal images using weakly supervised deep-learning framework," *IEEE Trans. Image Process.*, vol. 30, pp. 3167–3178, 2021, doi: 10.1109/TIP.2021.3058570.
5. H. Caesar *et al.*, "nuscenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11621–11631, doi: 10.1109/CVPR42600.2020.01164.
6. T. Kim and B. Lee, "Multi-attention multimodal sentiment analysis," in *Proc. Int. Conf. Multimedia Retrieval*, 2020, pp. 436–441, doi: 10.1145/3372278.3390698.
7. W. Guo, Y. Zhang, X. Cai, L. Meng, J. Yang, and X. Yuan, "LD-MAN: Layout-driven multimodal attention network for online news sentiment recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 1785–1798, Jun. 2020, doi: 10.1109/TMM.2020.3003648.
8. R. F. Silva, S. M. Plis, T. Adalı, M. S. Pattichis, and V. D. Calhoun, "Multidataset independent subspace analysis with application to multimodal fusion," *IEEE Trans. Image Process.*, vol. 30, pp. 588–602, Oct. 2020, doi: 10.1109/TIP.2020.3028452.
9. X. Li, C. Xu, G. Yang, Z. Chen, and J. Dong, "W2vv++ fully deep learning for ad-hoc video search," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1786–1794, doi: 10.1145/3343031.3350906.
10. Z. Yuan, S. Sun, L. Duan, C. Li, X. Wu, and C. Xu, "Adversarial multimodal network for movie story question answering," *IEEE Trans. Multimedia*, vol. 23, pp. 1744–1756, Jun. 2020, doi: 10.1109/TMM.2020.3002667.
11. L. Zhen, P. Hu, X. Peng, R. S. M. Goh, and J. T. Zhou, "Deep multimodal transfer learning for cross-modal retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, 2020, doi: 10.1109/TNNLS.2020.3029181.
12. W. Sun, F. Ma, Y. Li, S.-L. Huang, S. Ni, and L. Zhang, "Semi-supervised multimodal image translation for missing modality imputation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 4320–4324, doi: 10.1109/ICASSP39728.2021.9413461.

13. M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 5580–5590.
14. M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "SMIL: Multimodal learning with severely missing modality," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, pp. 2302–2310, 2021.
15. L. Cabrera-Quiros, D. M. J. Tax, and H. Hung, "Gestures in-the-wild: Detecting conversational hand gestures in crowded scenes using a multimodal fusion of bags of video trajectories and body worn acceleration," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 138–147, Jan. 2020, doi: 10.1109/TMM.2019.2922122.
16. J. Lei, L. Yu, T. L. Berg, and M. Bansal, "TVQA+: Spatio-temporal grounding for video question answering," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8211–8225, doi: 10.18653/v1/2020.acl-main.730.

SHU-CHING CHEN is a Professor with Florida International University, Miami, FL, USA. Contact him at chens@cs.fiu.edu.