Deep learning approaches in content-based video indexing and retrieval

Newton Spolaôr[1], Huei D. Lee[1], Narco A. R. Maciejewski[1], Leandro A. Ensina[1], Weber S. R. Takaki[12], Cláudio S. R. Coy[2], Feng C. Wu[12]

[1] Western Paraná State University (UNIOESTE)
[2] University of Campinas (UNICAMP)

Introduction

Content-based Video Indexing and Retrieval (CBVIR) has been widely studied due to a large number of applications and research interest in exploring the rich video content [1]. Machine Learning (ML) methods are a usual choice to support feature extraction and other CBVIR tasks [2]. Recently, Deep Learning (DL) approaches such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) arose as promising alternatives to extract and analyze video information [3]. However, there is no comprehensive and replicable surveys on DL applications in CBVIR. This work reports a Systematic Review (SR) method to find and summarize recent CBVIR publications, focusing on the papers using DL concepts.

Method

An SR [4] was carried out in this work to survey the state-of-the-art of the deep learning methods for video retrieval applications. This search was based on the research protocol, available at (http://tiny.cc/eb3isy) and designed by us for a broader SR on CBVIR approaches that were published from 2011 to March 2017. First, we defined a research question: "what segmentation, feature extraction, dimensionality reduction and machine learning approaches have been applied for content-based video indexing and retrieval?". Based on this question, a search string structured as follows was proposed: "(indexing OR retrieval) AND (content-based) AND (video)". The final string with 35 terms was used in seven bibliographic databases. To select relevant papers from the search results, we specified 13 exclusion criteria. The methodological quality strategy consisted of applying four quality criteria as yes/no questions. Finally, 18 information items were extracted from the selected publications to verify their quality and allow us to summarize the papers using DL concepts.

Results

The method allowed us to find seven papers that use DL for video indexing. Most of them extract low or high-level features from video frames [5,6,7,8]. Other publications learn binary codes for compact video representation [7] that can even exploit temporal video properties [9]. Indexing based on automatic speech recognition and optical character recognition has also benefited from DL [10,11]. Regarding the DL approach, CNN was used to predict individual concepts, characters [10] and phonemes [11]. In [7], a multi-label CNN predicts multiple concepts simultaneously. CNN popularity can be associated with its ability to process data organized as several arrays, as illustrated by pixel matrices from frames [3]. RNN, an alternative approach effective in sequence modeling [3], considers the frame order to describe videos in [9]. It should be emphasized that most DL applications outperformed other methods, such as bag of visual words [7] and video hashing algorithms [9].

Conclusion

In this work, an SR was performed to find and summarize seven DL approaches recently applied for CBVIR. CNN has been the most usual architecture to predict concepts, characters and phonemes, while RNN was used to support temporal information modeling. Based on the findings, a potential future direction involves dealing with a DL challenge – the lack of large labeled datasets – by crowd-sourcing the video labeling task [12].

Acknowledgments

We would like to acknowledge the EurekaSD project - Enhancing University Research and Education in Areas Useful for Sustainable Development, the Araucária Foundation for the Support of the Scientific and Technological Development of Paraná through a Research and Technological Productivity Scholarship and the Coordination for the Improvement of Higher Education Personnel (CAPES) for the two scholarships granted.

References

[1] W. Hu, N. Xie, L. Li, X. Zeng and S. Maybank, A Survey on Visual Content-Based Video Indexing and Retrieval, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41:797-819, 2011.

[2] S. Sharma, J. Agrawal, S. Agarwal and S. Sharma, Machine Learning Techniques for Data Mining: A Survey, In: *IEEE International Conference on Computational Intelligence and Computing Research*, IEEE, Enathi, 2013, 1-6.

[3] Y. Lecun, Y. Bengio and G. Hinton, Deep Learning, *Nature*, 521:436-444, 2015.

[4] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner and M. Khalil, Lessons from Applying the Systematic Literature Review Process within the Software Engineering Domain, *Journal of Systems and Software*, 80(4):571-583, 2007.

[5] L. Jiang, S. Yu, D. Meng and Y. Yang, Fast and Accurate Content-based Semantic Search in 100M Internet Videos, In: *ACM International Conference on Multimedia*, ACM, Brisbane, 2015, 49-58.

[6] S. Yu, L. Jiang, Z. Xu, Y. Yang and A. G. Hauptmann, Content-Based Video Search over 1 Million Videos with 1 Core in 1 Second, In: *ACM on International Conference on Multimedia Retrieval*, ACM, Shangai, 2015, 419-426.

[7] M. Mühling, M. Meister, N. Korfhage, J. Wehling, A. Hörth, R. Ewerth and B. Freisleben, Content-Based Video Retrieval in Historical Collections of the German Broadcasting Archive, In: *International Conference on Theory and Practice of Digital Libraries*, Springer International Publishing, Hannover, 2016, 67-78.

[8] A. Agharwal, R. Kovvuri, R. Nevatia and C. G. M. Snoek, Tag-based Video Retrieval by Embedding Semantic Content in a Continuous Word Space, In: *IEEE Winter Conference on Applications of Computer Vision*, IEEE, Lake Placid, 2016, 1-8.

[9] H. Zhang, M. Wang, R. Hong and T. Chua, Play and Rewind: Optimizing Binary Representations of Videos by Self-Supervised Temporal Hashing, In: *ACM on Multimedia Conference*, ACM, Amsterdam, 2016, 781-790.

[10] W. Wattanarachothai and K. Patanukhom, Key Frame Extraction for Text Based Video Retrieval Using Maximally Stable Extremal Regions, *EAI Endorsed Transactions on e-Learning*, 15(7):1-9, 2015.

[11] T. H. Luong, N. M. Pham and Q. H. Vu, Vietnamese Multimedia Agricultural Information Retrieval System as an Info Service, In: *International Workshop on*

Worldwide Language Service Infrastructure, Springer International Publishing, Kyoto, 2016, 147-160.
[12] J. Ker, L. Wang, J. Rao and T. Lim, Deep Learning Applications in Medical Image Analysis, *IEEE Access*, 6:9375-9389, 2018.