UNIVERSITY "ST. KLIMENT OHRIDSKI" - BITOLA
FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGIES - BITOLA
REPUBLIC OF NORTH MACEDONIA

Proceedings of the 13th International Conference on Applied Internet and Information Technologies

AIIT 2023

13 October, 2023, Bitola, Republic of North Macedonia

AIIT
International Conference

UNIVERSITY "ST. KLIMENT OHRIDSKI" - BITOLA
FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGIES - BITOLA
REPUBLIC OF NORTH MACEDONIA

13TH INTERNATIONAL CONFERENCE ON APPLIED INTERNET AND INFORMATION TECHNOLOGIES

AIIT 2023
PROCEEDINGS

Bitola, 2023

Proceedings publisher and organizer of the conference:
University “St. Kliment Ohridski” - Bitola, Faculty of Information and Communication Technologies - Bitola, Republic of North Macedonia

For publisher:
Blagoj Ristevski, PhD
Faculty of Information and Communication Technologies - Bitola
University “St. Kliment Ohridski” - Bitola, North Macedonia

Proceedings editors:
Kostandina Veljanovska, PhD
Eleonora Brtka, PhD

Conference Chairmans:
Kostandina Veljanovska, University “St. Kliment Ohridski”, Faculty of Information and Communication Technologies - Bitola, Republic of North Macedonia - chair
Eleonora Brtka, University of Novi Sad, Technical Faculty “Mihajlo Pupin”, Zrenjanin, Serbia – co-chair

Technical preparation of the proceedings:

Andrijana Bocevska, PhD

Kostandina Veljanovska, PhD

Cover design:

Andrijana Bocevska, PhD

e-Proceedings
ISBN 978-608-5003-03-7

CIP - Каталогизација во публикација
Национална и универзитетска библиотека "Св. Климент Охридски", Скопје

004-049.8(062)

INTERNATIONAL conference on applied internet and information technologies AIIT 2023 (13 ; 2023 ; Bitola, Republic of North Macedonia)
Proceedings / 13th International conference on applied internet and information technologies AIIT 2023, 13 October, 2023, Bitola, Republic of North Macedonia ; [editors Kostandina Veljanovska, Eleonora Brtka]. - Bitola : University "St. Kliment Ohridski", Bitola Faculty of information and communication technologies, 2023. - 411 стр. : илустр. ; 30 см
Библиографија кон трудовите
ISBN 978-608-5003-03-7

а) Информатичка технологија -- Примена -- Собири
COBISS.MK-ID 62559493

C O N T E N T

Latest Advances in Video Indexing and Retrieval

Nora Pireci Sejdiu[1], Zoran Kotevski[2], Blagoj Ristevski[3] and Kostandina Veljanovska[4]

[1,2,3,4] *"St. Kliment Ohridski" University - Bitola, Faculty of Information and Communication Technologies - Bitola, Studentska bb., 7000 Bitola, North Macedonia*

[1]*pireci.nora@uklo.edu.mk;* [2]*zoran.kotevski@uklo.edu.mk;* [3]*blagoj.ristevski@uklo.edu.mk;* [4]*kostandina.veljanovska@uklo.edu.mk*

Abstract:
This paper provides a review of recent advancements in the field of multimodal video indexing and retrieval. The presented papers cover various topics including multimodal fusion, keyframe extraction, semantic analysis, scalability, and evaluation. The reviewed studies propose innovative methods and techniques to address the challenges associated with these areas. However, several unresolved issues and opportunities for future research are identified. These include the effective integration of multimodal information, improvement of keyframe extraction accuracy, semantic understanding of video content, scalability of video retrieval systems, and standardized evaluation protocols. By addressing these challenges, future research can contribute to the development of more robust and efficient video retrieval systems, enabling better access and utilization of large-scale video collections across diverse domains.

Keywords:
Video indexing, video retrieval, semantic video indexing, keyframe extraction

1. Introduction

Multimodal video indexing and retrieval have emerged as crucial research areas due to the exponential growth of video data in various domains such as surveillance, entertainment, education and social media. With the increasing availability of multimodal content, including text, audio, and visual information, there is a pressing need to develop efficient and effective methods for indexing and retrieving videos based on their content. The ability to automatically analyze and annotate videos enables improved search capabilities, content recommendation, and information extraction.

In recent years, several studies have made significant contributions to multimodal video indexing and retrieval by proposing innovative approaches and techniques. These studies address various challenges and issues related to different aspects of video analysis. For instance, the fusion of multiple modalities plays a critical role in capturing the comprehensive representation of videos [15, 16]. By leveraging the complementary nature of textual, audio, and visual features, multimodal fusion techniques aim to improve the accuracy and robustness of indexing and retrieval systems.

Another important aspect is the extraction of keyframes, which are representative frames that capture the essence of video content. Keyframes play a vital role in summarizing and representing videos effectively [8, 13]. Several studies have proposed methods for automatic keyframe extraction using feature extraction, shot boundary detection, and clustering techniques. These approaches aim to identify salient frames that encapsulate the content of a video and facilitate efficient retrieval.

Semantic analysis and annotation of video content have also been addressed in the literature [12]. The challenge lies in bridging the semantic gap between low-level visual features and high-level semantic concepts associated with videos. By utilizing approaches such as correspondence-Latent Dirichlet Allocation (corr-LDA) and information extraction techniques, researchers aim to automatically annotate videos with textual descriptors, enabling more accurate and meaningful retrieval of video content.

Scalability and retrieval performance are significant concerns in the context of large-scale video databases [4, 5]. Efficient indexing structures, optimization algorithms, and parallel processing techniques are being explored to improve the scalability and retrieval speed of video retrieval systems. Additionally, the development of standardized evaluation protocols and benchmark datasets is essential for fair comparisons and objective assessments of different methods [3].

Despite the significant progress made in multimodal video indexing and retrieval, several challenges and unresolved problems persist. Future work should focus on addressing these challenges and exploring new directions for research. These include enhancing multimodal fusion techniques, improving keyframe extraction accuracy and adaptability, advancing semantic understanding of video content, enhancing scalability and retrieval performance, and establishing standardized evaluation frameworks.

In this paper, we provide a comprehensive review of recent advancements in multimodal video indexing and retrieval. We analyze and summarize the contributions of various studies in addressing the aforementioned challenges. By examining the proposed methods and techniques, we identify the key issues that need further exploration and discuss potential avenues for future research. Through this review, we aim to provide insights and directions for researchers in the field of multimodal video indexing and retrieval.

The rest of the paper is structured as follows. Section II describes the latest advances in video indexing and retrieval. Section III discusses the presented papers, which address various challenges and propose innovative solutions for multimodal video indexing and retrieval, including multimodal fusion, keyframe extraction, semantic analysis, scalability, and evaluation. Along with the discussion about their contribution, this section comments ongoing research that is needed to tackle the unresolved issues and improve the efficiency of video retrieval systems. Concluding remarks are highlighted in the last section.

2. Latest advances in video indexing and retrieval

The field of video indexing and retrieval has witnessed significant advances in recent years, driven by the exponential growth of video data and the increasing demand for efficient video search and analysis. Researchers have made notable progress in various aspects of video indexing and retrieval. In the domain of multimodal fusion, Podlesnaya and Podlesnyy have explored deep learning-based techniques for semantic video indexing and retrieval [15], while Hamroun et al. introduced a novel method employing machine learning and semi-automatic annotation on large video collections [16]. For keyframe extraction, Priya, Lakshmi, and Domnic have contributed to the field by developing shot-based keyframe extraction methods tailored for ecological video indexing and retrieval [13]. In the realm of semantic analysis, Safadi and Quénot proposed re-ranking by local re-scoring techniques to enhance video indexing and retrieval results [12]. Lee, Park, and Yoo addressed scalability challenges in surveillance video indexing and retrieval by introducing a data cube model [5]. Moreover, Wang, Huang, Wang, Zhang, and Tian have discussed video indexing and retrieval based on key frame extraction, providing insights into effective techniques for organizing and accessing video content [3]. Multimodal fusion techniques have been developed to effectively integrate and exploit the complementary information from different modalities such as text, audio, and visual features. Keyframe extraction methods have improved the accuracy of identifying representative frames that summarize video content. Semantic analysis approaches have aimed to bridge the gap between low-level visual features and high-level semantic concepts for more meaningful video annotation and retrieval. Scalability challenges have been addressed through the development of optimized indexing structures and parallel processing techniques. Furthermore, standardized evaluation protocols and benchmark datasets have been established to facilitate fair comparisons and objective assessments of video indexing and retrieval methods. These recent advancements lay a solid foundation for further research and innovation in this field, opening up new opportunities to enhance the efficiency and effectiveness of video indexing and retrieval systems.

The semantic pathfinder, a generic approach for video indexing that aims to bridge the semantic gap in multimedia archives is presented in [1]. The approach is based on the understanding that videos are the result of an authoring process. The semantic pathfinder utilizes various detector types, multimodal analysis, hypothesis selection, and machine learning to select the most appropriate path through content analysis, style analysis, and context analysis. Experimental results using a lexicon of 32 semantic concepts demonstrate the effectiveness of the semantic pathfinder in enabling generic video indexing and validating the value of the authoring metaphor in the indexing process. However, the precision for

some concepts at 100 performances is still relatively low, which may be sufficient for tasks involving selecting illustrative footage but not accurate retrieval tasks. Further improvements are needed to enhance the accuracy of retrieval tasks.

Paper [16] introduces a novel method for multimodal indexation using machine learning and semi-automatic annotation. The authors propose a genetic algorithm-based approach to detect concepts from the text in videos. They enrich basic concepts using a method called DCM and utilize semantic and enriched concepts to improve multimodal indexation and construct an ontology. The method's effectiveness is evaluated on the TRECVID 2015 dataset.

Paper [15] focuses on deep learning-based approaches for semantic video indexing and retrieval. The system utilizes features extracted by convolutional neural networks (CNNs) to enable efficient indexing and retrieval. The paper provides implementation details and empirical evaluation results to showcase the effectiveness of the proposed approach.

In [14], the authors propose a framework for multimodal video indexing and retrieval using a similarity measure called shrinkage-optimized directed information assessment (SODA). The directed information captures the direction of information flow in videos and is applied to audio-visual features over successive frames. This approach leverages the natural characteristics of videos and aims to improve indexing and retrieval performance.

Lakshmi and Domnic in [13] present a new automatic shot-based keyframe extraction technique for video indexing and retrieval applications. They utilize feature extraction and continuity value construction steps of shot boundary detection to cluster frames into shots. Keyframes are then extracted based on inter-cluster similarity analysis. This method provides a more efficient and accurate way of representing video content for indexing and retrieval purposes.

A re-ranking method to enhance the performance of semantic video indexing and retrieval systems is proposed in [12]. The proposed method re-evaluates the scores of shots based on their homogeneity and their relationship with the overall video content. By considering the temporal sequence and the distribution of video shots, this method improves the ranking accuracy and provides a framework for re-ranking.

A specific approach for content-based video indexing and retrieval using the Correspondence-Latent Dirichlet Allocation (corr-LDA) probabilistic framework is presented in [11]. The method utilizes the semantic relations between video content and text to improve indexing accuracy. The authors focus on the audio components of video recordings and compare the results with a support vectore machines (SVM)-based approach.

In [10], a Knowledge Distillation framework called Distill-and-Select (DnS) for video indexing and retrieval is introduced. The framework involves training Student Networks with different retrieval performance and computational efficiency trade-offs. A Selector Network is also trained to direct samples to the appropriate student network at test time. This approach maintains high retrieval performance while ensuring computational efficiency.

Paper [9] presents a probabilistic Bayesian belief network (BBN) method for the automatic indexing of excitement clips in sports video sequences. The excitement clips are extracted using audio features, and multiple subclips corresponding to different events are detected and classified. The proposed method utilizes hierarchical classification and provides an efficient approach to indexing sports videos.

Paper [8] proposes a novel algorithm for shot detection and keyframe determination based on gradient fields. The algorithm aims to find a compact set of keyframes that accurately represents a video segment. By considering the structural similarity of frames, the proposed approach provides an effective method for content-based video indexing and retrieval.

Paper [7] presents an approach for automatic lecture video indexing based on video OCR technology. It introduces a video segmenter for slide video structure analysis and a weighted DCT-based text detector. The text occurrence information and analyzed text content are used for indexing. The proposed method leverages video OCR technology to automate the indexing process, improving the accessibility and searchability of lecture videos.

Paper [6] presents a novel framework for video surveillance indexing and retrieval. The proposed framework consists of three main modules: pre-processing, query processing, and retrieval processing. It aims to overcome the challenges of video surveillance data management and enhance the efficiency and effectiveness of video retrieval systems in surveillance applications.

Paper [5] proposes a data cube model, SurvCube, for the multi-dimensional indexing and retrieval of surveillance videos. The proposed method utilizes the data cube structure to provide multi-dimensional analysis of interesting objects in surveillance videos based on chronological view, events, and locations. It offers functionalities such as retrieval at different levels of abstraction, tracing object trajectories, and summarization of surveillance videos.

Paper [4] addresses the challenges of managing large-scale video databases by utilizing spatial indexing and querying of field-of-views (FOVs). The authors propose a class of new R-tree-based index structures that effectively harness FOVs' location, orientation, and view-distance properties. The proposed method optimizes filtering and query optimization by considering the geographical properties of FOVs, providing efficient indexing and retrieval of geo-tagged video databases.

Paper [3] presents a query-by-concept approach for video retrieval. The proposed method focuses on key frame extraction and achieves better results compared to existing methods. The evaluation shows a mean average precision (MAP) of 0.68 for the video retrieval model.

3. Discussion

The presented papers cover a wide range of approaches and techniques for multimodal video indexing and retrieval. These studies contribute to the advancement of video retrieval systems by addressing various challenges and proposing innovative solutions. However, there are still several issues and unresolved problems that need to be addressed in future research.

One of the challenges in video indexing and retrieval is the effective utilization of multimodal information, including text, audio, and visual content as show in Figure 1 [14]. Several papers propose methods for extracting and integrating these modalities to improve the indexing and retrieval process [14, 15, 16]. However, achieving a robust and accurate multimodal fusion remains an ongoing research problem.

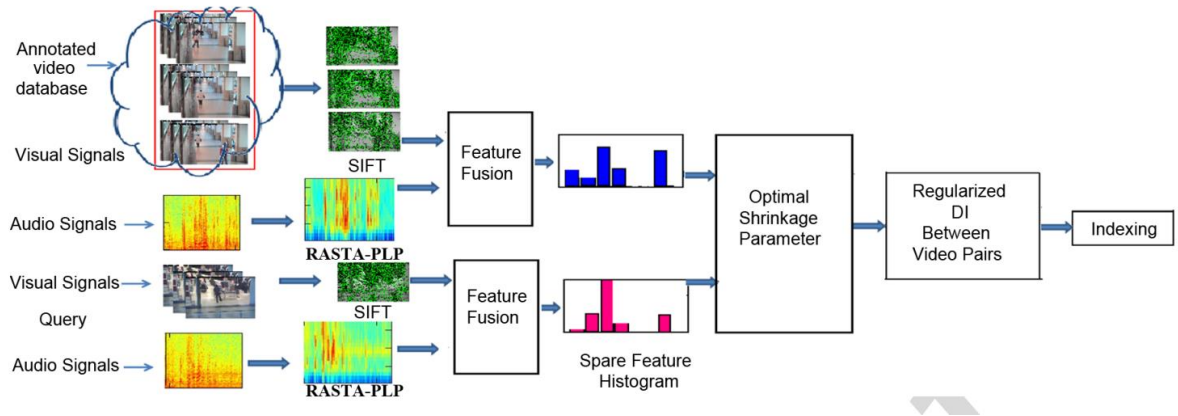

Figure 1: Block diagram of shrinkage optimized directed information (SODA) for fusion of audio and visual features for video indexing

Another area of research focuses on the extraction of keyframes and shots from videos as presented in Figure 2 [8]. Keyframe extraction plays a crucial role in representing the content of a video efficiently. While the proposed methods demonstrate advancements in automatic keyframe extraction, there is still room for improvement in terms of accuracy and adaptability to different video genres and styles [8, 13].

Semantic analysis and annotation of video content are also addressed in paper [5]. These approaches aim to bridge the semantic gap between the low-level visual features and the high-level semantic concepts associated with videos. However, achieving precise and comprehensive semantic understanding of video content remains a challenge due to the inherent complexity and subjectivity of video semantics.

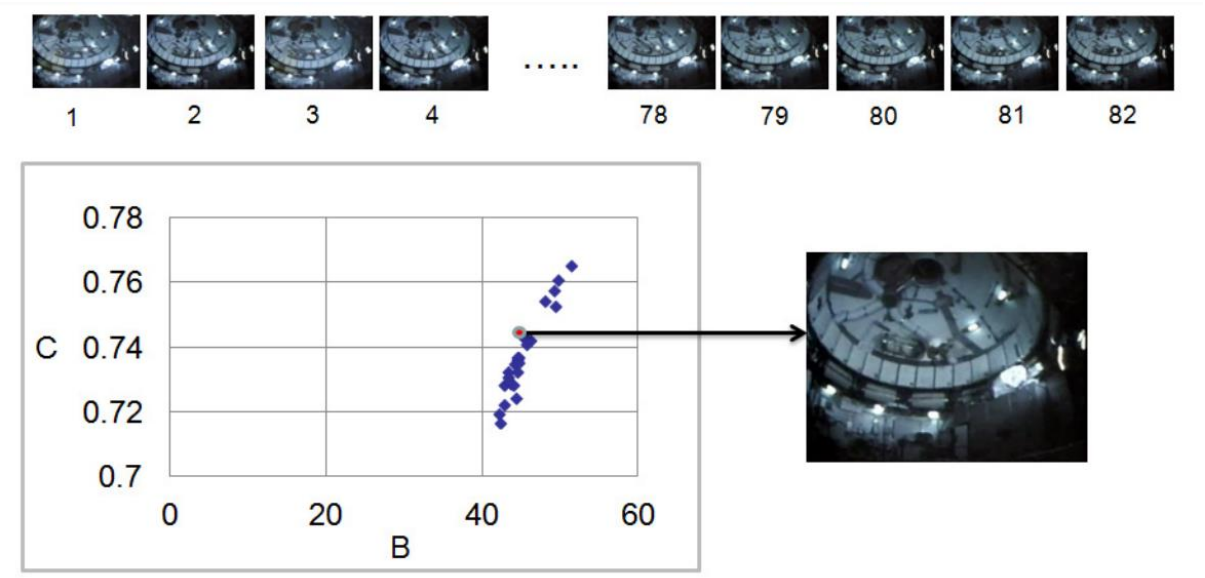

Figure 2: Illustration of key-frame

Efficient indexing and retrieval of large-scale video databases pose significant challenges [4, 5]. The scalability and retrieval performance of video retrieval systems need to be further improved to handle the ever-increasing volume of video data. Additionally, the integration of spatial and temporal information in video indexing and retrieval remains an area of research that requires further exploration.

Another important aspect is the evaluation and benchmarking of video retrieval systems. While some papers mention the evaluation of their proposed methods using specific datasets [16], there is a need for standardized evaluation protocols and benchmark datasets that cover diverse video genres and capture real-world scenarios.

In conclusion, the discussed papers contribute valuable insights and advancements in multimodal video indexing and retrieval. However, several challenges and issues, such as multimodal fusion, keyframe extraction, semantic analysis, scalability, and evaluation, require further attention and research efforts in order to develop more robust and efficient video retrieval systems.

4. Conclusions

In conclusion, the presented papers on multimodal video indexing and retrieval have made significant contributions to the field by proposing innovative methods and techniques. These studies address various challenges related to multimodal fusion, keyframe extraction, semantic analysis, scalability, and evaluation. However, there are still unresolved issues and opportunities for future work.

The challenges of effectively integrating multimodal information, including text, audio, and visual content, remain open research problems. Future work can focus on developing advanced fusion techniques that leverage the complementary nature of different modalities to improve the accuracy and robustness of video indexing and retrieval systems.

Keyframe extraction continues to be an important task for representing video content efficiently. Further research is needed to enhance the accuracy and adaptability of keyframe extraction methods across different video genres and styles. Additionally, exploring novel approaches that consider temporal dynamics and semantic relevance can lead to more comprehensive and meaningful keyframe representations.

Semantic analysis and annotation of video content remain challenging due to the complexity and subjectivity of video semantics. Future work can explore advanced machine learning and deep learning techniques to bridge the semantic gap and enable a more precise and comprehensive understanding of video content. Developing ontologies and knowledge graphs that capture the semantic relationships among video concepts can also contribute to more effective indexing and retrieval.

The scalability and retrieval performance of video retrieval systems require continuous improvement to handle the ever-increasing volume of video data. Future research can focus on developing efficient indexing structures, optimization algorithms, and parallel processing techniques to enhance the scalability and retrieval speed of large-scale video databases.

The evaluation and benchmarking of video retrieval systems need standardized protocols and benchmark datasets that cover diverse video genres and real-world scenarios. Future work can focus on developing comprehensive evaluation frameworks and datasets to enable fair comparisons among different methods and facilitate progress in the field.

References:

[1] Blanken, Henk M., et al., eds. “Multimedia retrieval.” Springer Science & Business Media. (2007).
[2] H. Yang, M. Siebert, P. Luhne, H. Sack and C. Meinel, "Lecture Video Indexing and Analysis Using Video OCR Technology," 2011 Seventh International Conference on Signal Image Technology & Internet-Based Systems, Dijon, France, pp. 54-61, doi: 10.1109/SITIS.2011.20. (2011).
[3] Wenshi Wang, Zhangqin Huang, Weidong Wang, Shuo Zhang, and Rui Tian. “Video Indexing and Retrieval based on Key Frame Extraction.” [J]. Int J Performability Eng, 14(8): 1824-1832. (2018).
[4] Lu, Y., Shahabi, C. and Kim, S.H. “Efficient indexing and retrieval of large-scale geo-tagged video databases.” Geoinformatica 20, 829–857. (2016).
[5] H. Lee, S. Park and J. -H. Yoo. "A data cube model for surveillance video indexing and retrieval." 2013 International Conference on Signal Processing and Multimedia Applications (SIGMAP), Reykjavik, Iceland, pp. 163-168. (2013).
[6] Fereshteh Falah Chamasemani, Lilly Suriani Affendey, Norwati Mustapha and Fatimah Khalid. “A Framework for Automatic Video Surveillance Indexing and Retrieval.” Research Journal of Applied Sciences, Engineering and Technology 10(11): 1316-1321. (2015).
[7] H. Yang, M. Siebert, P. Luhne, H. Sack and C. Meinel. "Automatic Lecture Video Indexing Using Video OCR Technology." *2011 IEEE International Symposium on Multimedia*, Dana Point, CA, USA, pp. 111-116, doi: 10.1109/ISM.2011.26. (2011).
[8] Asatryan, David, and Manuk Zakaryan. "Novel approach to content-based video indexing and retrieval by using a measure of structural similarity of frames." International Journal" Information Content and Processing 2.1, 71-81. (2015).
[9] Kolekar, Maheshkumar H. "Bayesian belief network based broadcast sports video indexing." Multimedia Tools and Applications 54, 27-54. (2011).
[10] Kordopatis-Zilos, Giorgos, et al. "DnS: Distill-and-Select for Efficient and Accurate Video Indexing and Retrieval." International Journal of Computer Vision 130.10, 2385-2407,(2022).
[11] Iyer, Rahul Radhakrishnan, et al. "Content-based video indexing and retrieval using corr-lda." arXiv preprint arXiv:1602.08581. (2016).
[12] Safadi, Bahjat, and Georges Quénot. "Re-ranking by local re-scoring for video indexing and retrieval." Proceedings of the 20th ACM international conference on Information and knowledge management. (2011).
[13] Priya, GG Lakshmi, and S. Domnic. "Shot based keyframe extraction for ecological video indexing and retrieval." Ecological informatics 23.107-117. (2014).
[14] Chen, Xu, Alfred O. Hero III, and Silvio Savarese. "Multimodal video indexing and retrieval using directed information." IEEE Transactions on Multimedia 14.1. 3-16. (2011).
[15] Podlesnaya, Anna, and Sergey Podlesnyy. "Deep learning based semantic video indexing and retrieval." Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016: Volume 2. Springer International Publishing. (2018).
[16] Hamroun, Mohamed, Karim Tamine, and Benoît Crespin. "Multimodal video indexing (mvi): A new method based on machine learning and semi-automatic annotation on large video collections." International Journal of Image and Graphics 22.02. 2250022. (2022).