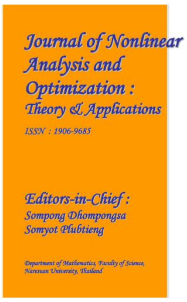
Journal of Nonlinear Analysis and Optimization
Vol. 15, Issue. 1, No.15 : 2024
ISSN : **1906-9685**



Journal of Nonlinear Analysis and Optimization : Theory & Applications
ISSN : 1906-9685
Editors-in-Chief :
Sompong Dhompongsa
Somyot Plubtieng
Department of Mathematics, Faculty of Science,
Naresuan University, Thailand

Paper ID: ICRTEM24_118 **ICRTEM-2024 Conference Paper**

EXPLORING THE CHALLENGES AND FUTURE TRENDS IN MULTIMODAL INFORMATION RETRIEVAL

#1CH. VAMSHI RAJ, ***Research Scholar,***
#2Dr. YOGESH KUMAR SHARMA, Associate ***Professor & Guide,***
#3Dr. M. ANJAN KUMAR, ***Professor & Co-Guide,***
Department of Computer Science & Engineering,
SHRI JAGDISHPRASAD JHABARMAL TIBREWALA UNIVERSITY, RAJASTHAN.

ABSTRACT: Multimodal information retrieval is a research topic that is of great interest in many fields. This is owing to the large amount of multimedia data available in a variety of settings, such as text, photos, audio, and video. To create an efficient and user-friendly retrieval system, researchers are incorporating multimodal information retrieval using a variety of techniques such as machine learning, support vector machines, neural networks, and neuroscience, among others. The goal of this study is to present an overview of multimodal information retrieval, as well as the current challenges associated with it.

Keywords: Multi Modal Information Retrieval, Information Retrieval, Machine Learning, SVM, Semantic Gap, Query Reformulation, Fusion Techniques.

1. INTRODUCTION

Over the last decade, there has been a huge growth in the amount of digital content available online. The combination of digital television, mobile connectivity, and the internet has resulted in an increase in the production of audio, video, and image content, transforming the web into a comprehensive multimedia platform. Given the ubiquitous use of multimedia, it is critical to develop intelligent and effective ways for managing the massive amounts of multimedia data. This study examines the work done in each modality independently to help readers understand the unique retrieval paradigms of each.

There are two primary methods for extracting textual information: categories (ontology) and keywords (ad hoc retrieval). Each retrieval method has benefits and disadvantages, thus they are utilized in a variety of applications. A flexible system capable of applying a variety of approaches to intelligently recognize the application is required. Classical information retrieval applications fall into four categories: content searching, text classification/clustering, content management, and question answering. The majority of these applications use statistical or machine learning methodologies. Indexing strategies improve the efficiency of retrieval procedures. Researchers employed query reformulation techniques to collect the data that users required. Lavrenko and Croft used relevant feedback to improve understanding of client needs.

This entails attaining relevance on open documents or through a feedback interface that evaluates the relevance of retrieved material to the user's needs.

There are four primary approaches for retrieving images: descriptor-based retrieval, texture or pattern recognition, feature-based retrieval, and object-based retrieval. Early systems relied heavily on color, texture, and shape as visual identifiers. The article addresses using feature matching to extract local features from segmented image sections or patches. The researchers used approaches such as Bag-Of-Visual terms, Scale Invariant Feature Transform (SIFT), inverted files, and Fisher Vectors. Previous study focused on two types of characteristics: local features and global features.

Early audio retrieval systems relied heavily on metadata such as the artist, song title, and album title. Content-based retrieval approaches include converting audio signals to text or determining rhythm and tempo similarities. Researchers have also used annotation-based techniques to audio retrieval. Multimodal information retrieval (MMIR) is the process of using two or more retrieval methods to search for information on the web in diverse formats. A new data format for multimodal data is introduced in Content Object (CO), providing a unified framework for multimodal search and retrieval. Recent advances in multimodal retrieval systems have resulted in the formation of various academic and business projects, as well as a burgeoning research community. Modern technologies prioritize multimodal retrieval engines alongside single-mode retrieval systems. Requests for video and related papers based on a specific inquiry, as well as information on mouth cancer, are becoming increasingly common. These demands are the core focus of current development. To achieve such stringent criteria, researchers must work on a variety of factors, as detailed below:

- ➢ New semantic models for combining individual media models.
- ➢ New retrieval engines for crossing the media boundary during search.
- ➢ New interfaces managing the input and presentation of various media data, etc.
- ➢ New retrieval models for retrieving relative information within the same media as well as in cross media.

This study is broken into three parts. Section 2 summarizes the numerous obstacles encountered by MMIR systems. This section has covered a variety of issues, with a focus on semantic gaps and multimodal information fusion. Section 4 concludes with a summary of previous research on MMIR systems. Section 3 examines upcoming trends in MMIR systems and makes a recommendation for closing the semantic gap.

2. CHALLENGES IN MMIR SYSTEMS

There are various difficult issues with MMIR systems, which will be briefly discussed here. Addressing the semantic discrepancy between user information requirements and low-level features such as color, texture, shape, object motion, frequency, and so on is an important task. The difficulty is to relate the features of modern multimedia systems to the high-level concepts that consumers want, such as identifying all images of accidents caused by reckless driving.

Another major concern is repeatability, which refers to the capacity to maintain constant performance when switching from the research test set to general-purpose video streams. Another essential aspect is how to strategically design these systems so that users can fulfill their needs without being bombarded with irrelevant results. Furthermore, the feature space is extremely sparse due to the high dimensionality of multimedia data, making generalization difficult. Maximizing the amount of training data used may help to address the issue of generalization.

Another factor to consider is search performance, particularly when working with large amounts of multimedia content. Ensuring optimal system scalability is difficult, especially for large programs like national digital libraries. Distributed computing, particularly in three-tier designs, is proposed as a solution to the scaling problem. Accurate modeling is critical since many existing systems lack a solid theoretical foundation.

Applying contextual limitations enhances these systems by adding more metadata, which boosts retrieval performance. Because context is included in MMIR systems, they are both content- and context-based. Another problem to overcome is the dynamic selection and integration of appropriate engagement modalities.

This fusion must be completed instantly with minimum user involvement. Evaluating the similarity of multimedia content is very important. Various models are used to investigate similarity in multimedia data, although most systems rely on classical metric system metrics to determine the similarity of multimedia feature vectors. Another challenge is to provide user interfaces that allow people to enter a query and then review the results.

Semantic Gap

The discrepancy between high-level semantic concepts and low-level features is known as the "semantic gap." Many papers were reviewed to investigate the semantic gap issue. Extensive study has been undertaken in this sector over the previous decade, although little progress has been made in terms of universal applications. The scientific community has agreed that multimodal content retrieval via semantic retrieval is unlikely to occur for several years. Researchers are now using relevance input as a tool in the question formulation process rather than relying entirely on machine learning methods to close the semantic gap. Queries are far more difficult to create than uni-modal retrieval methods since there is no unambiguous expression of the user's overall impression in the basic elements of multimedia content.

During a search session, the user's information requirements are dynamic and constantly changing. Typically, multimodal searching is experimental, with users starting a session and learning through interaction with the technology. Existing MMIR systems are unable to meet dynamic search requirements.

Multimodal Information Fusion

Since Bolt's seminal studies on voice and gesture, a number of multimodal systems have been implemented. The data specifications and application requirements are critical components in the creation of a multimodal system. Some important factors to consider include the information sources, feature extraction approach, fusion level, fusion strategy, system architecture, and any embedded previous knowledge that may be required. The application's needs dictate the information sources used. From a practical standpoint, the system should be both computationally efficient and affordable. Furthermore, the selected modalities must be capable of producing discriminatory patterns that can be measured for categorization purposes. Specific apps may be required to meet certain criteria, including user approval, universality, and stability. A multimodal recognition system, for example, could include visual and audio clues in customer service and security settings. In contrast, technologies such as RFID, facial recognition, and biometric technology may be more appropriate for surveillance purposes.

The key challenges in integrating multimodal information are the effective integration of information from several channels, as well as the identification and extraction of distinguishing and complimentary properties. When performing pattern recognition tasks, it is critical to extract features that faithfully capture the universal aspects of the intended perception while simultaneously distinguishing it from different perceptions. Preprocessing the raw signal is common in order to remove noise or identify a specific region of interest. Face detection in a face recognition system, for example, or speech input noise reduction and echo cancellation. To build a compact representation, feature extraction techniques are applied to the preprocessed signal. Even when dealing with the same sort of signal, numerous feature extractors may be required, depending on the specific task at hand. For example, while prosodic qualities are often used in voice recognition, phonetic properties are thought to be the primary indicators of human emotion in speech. Furthermore, when a large number of features are retrieved, a feature selection strategy may be required to determine which ones are most discriminatory while reducing the dimensionality of the feature space.

The fusion of multimodal information typically occurs in three stages: data/feature level, score level, and decision level. Data/feature level fusion is the process of merging extracted features or the original data using specialized fusion techniques. A significant disadvantage of fusion at this level is the "curse of dimensionality," which is frequently associated with high computing costs and necessitates a large amount of training data. Furthermore, the integration difficulty may be exacerbated by the different qualities of features collected from different modalities, such as minute points for fingerprints and principle component analysis (PCA) features for facial images.

Fusion at the score level combines the scores produced by different classifiers employing various modalities using a rule-based method. To ensure that the final decision takes into account the relative relevance of each modality and that no one modality dominates the others, a score normalization approach is used to scale the scores provided by various modalities within the same range. In contrast, score level fusion can be performed by pattern classification, in which the scores are used as features in a pattern classification method. The rigidity of decision-level fusion stems from the scarcity of remaining knowledge. Furthermore, the final results are formed using procedures such as majority voting, which aggregates the selections of different modalities or classifiers. For example, fusion at the decision and score levels are remarkable examples of data/feature fusion.

Multimodal information fusion normally faces four major challenges: preserving information during fusion, computational complexity, identifying and eliminating redundancy, and extracting and selecting discriminatory features. The fusion approach should be able to effectively leverage data acquired from multiple sources in order to provide a more precise depiction of the intended impression. A faulty design of a multimodal system might result in reduced performance and feasibility.

3. FUTURE TRENDS

The user can rate the returned results using a technique called relevance feedback. This rating is used by the system to enhance the outcomes in the next iteration [26]. Relevance feedback is one way to capitalize on the fact that MMIR systems are intelligent search instruments controlled by people. It is critical that every new system embraces the aforementioned notion, and web-based video search engines have the ability to transform this field. Because a spectator may be attracted by a part of a film rather than the complete sequence, the development of scalable Internet video streaming systems is a promising step. Furthermore, there is a shift toward more specialized vertical domains. The development of real-time interactive mobile gadgets is also promoting the formation of new forms of human contact.

Furthermore, extensive study is needed to overcome the privacy and intellectual property rights obstacles that limit data transmission, which is critical for the creation of effective benchmarking systems. Integrating information visualization strategies and Human Computer Interface (HCI) methodologies with user intelligence will take more effort. It is also necessary to provide tools for pasting and selecting document portions or objects so that users can access the material contained within the file. Furthermore, study should be performed to determine how to handle situations where a higher recall accuracy is required.

The possible benefits of reducing the semantic barrier that separates low-level and high-level features include multimodal image recognition (MMIR) systems, among others. These systems may include robotics, search engines, question-and-answer systems, and surveillance systems. We are attempting to bridge this semantic gap in order to improve MMIR systems by assessing the relationships between various objects in images.

Initiatives and standards that promote network interoperability, the incorporation of more relevant content analysis techniques into metadata-based systems, and the streamlined integration of methods for media sharing, annotation, search, and management are also priorities. An inquiry into the evolution of modern MMIR systems to improve their capacity for information perusal and summarization would be beneficial. A few insightful browsing algorithms use three-dimensional representations of search results as a final step to increase user visualization.

4. CONCLUSION

This article presents a succinct overview of the fundamental ideas that make up Multimodal Information Retrieval (MMIR) systems, with a focus on the current state of the art, future advances, and problems.

Furthermore, the report explains the rationale for the widespread use of MMIR systems and emphasizes their critical requirements. This study looked at the existing issues faced by MMIR systems, as well as potential remedies and future improvements in the field.

REFERENCES

1. G. Hubert and J. Mothe, “An adaptable search engine for multimodal information retrieval”. J. Am. Soc. Inf. Sci., 60: 1625–1634. doi: 10.1002/asi.21091, 2009.
2. E. H. Y. Lim et al., “Knowledge Seeker - Ontology Modelling for Information Search and Management”, Springer Berlin Heidelberg, Vol. 8, pp. 27-36, 2011.
3. V. Lavrenko, and W.B. Croft, “Relevance Models in Information Retrieval”, Language Modeling for Information Retrieval, W. Bruce Croft and John Lafferty, ed., pp. 11-56, Kluwer Academic Publishers, Boston, 2003.
4. R. S. Dubey, R. Choubey and J. Bhattacharjee, “Multi Feature Content Based Image Retrieval”, (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 06, 2010, 2145-2149.
5. E. Kasutani, “Image retrieval apparatus and image retrieving method”, US Patent application 2007.
6. Y. Chen and J. Z. Wang, “A Region-based fuzzy feature matching approach to content-based image retrieval”, IEEE Trans. On PAMI, 24(9):1252-1267, 2002.
7. G. Csurka et al., “Visual categorization with bags of keypoints”. In Proc. of the ECCV Workshop on Statistical Learning for Computer Vision 2004.
8. F. Perronnin et al., “Large-scale image retrieval with compressed fisher vectors”. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2010.
9. F. Alhwarin et al., "Improved SIFT-Features Matching for Object Recognition"; BCS International Academic Conference 2008 – Visions of Computer Science, pp. 179-190; 2008.
10. S. Sarin and W. Kameyama, “Joint Equal Contribution of Global and Local Features for Image Annotation”, CLEF working notes, 2009.