Multimodal Information Retrieval: Challenges and Future Trends

Mohammad Ubaidullah Bokhari
Department of Computer Science
Aligarh Muslim University, Aligarh

Faraz Hasan
Department of Computer Science
Aligarh Muslim University, Aligarh

ABSTRACT
Multimodal information retrieval is a research problem of great interest in all domains, due to the huge collections of multimedia data available in different contexts like text, image, audio and video. Researchers are trying to incorporate multimodal information retrieval using machine learning, support vector machines, neural network and neuroscience etc. to provide an efficient retrieval system that fulfills user need. This paper is an overview of multimodal information retrieval, challenges in the progress of multimodal information retrieval.

General Terms
Multi Modal Information Retrieval, Information Retrieval.

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

1. INTRODUCTION
The growth of digital content on web has reached impressive rates in the last decade, the convergence of web, mobile access and digital television has boosted the production of images, audio and video contents, making the web a truly multimedia platform. Nowadays, multimedia popularity demands intelligent and efficient maneuverings in order to manage with the large amount of multimedia data. Due to the fact that, every modality has its own retrieval models so, for better understanding of methodology this paper describes the work done in each modality one by one.

In the field of textual information retrieval there are broadly two modes of retrieval based on keyword (ad-hoc retrieval), categories (ontology) [1]. All these modes of retrievals have their pros and cons, therefore used in different types of applications. There is a need of a generic system that can work on various techniques and recognizes the application intelligently. There are broadly four traditional information retrieval applications: content searching, text classification/clustering, content management and question answering, most of these use statistical or machine learning approaches [2]. For improving the retrieval methods in faster mode, indexing techniques are used. Researchers also used query reformulation techniques for retrieving appropriate information required by the user. For better understanding of user's need, Lavrenko and Croft [3] introduced the concept of relevance feedback for obtaining the relevance of retrieved information with the user's need in the form of feedback interface or getting relevance on opened documents.

In the area of image retrieval, there are broadly four modes of retrieval viz retrieval through descriptors, texture or pattern recognition, feature based retrieval and retrieval through objects [4]. Early systems mainly used image descriptors based on color or with texture and shape [5]. Local feature extraction from image patches or segmented image regions based on feature matching is discussed in [6]. Recently, techniques used by the researchers are based on inverted files, Bag-Of-Visual words [7], Fisher Vectors [8] and Scale Invariant Feature Transform (SIFT) [9]. There are broadly two types of features discussed in past works i.e. local features and global features [10].

In the field of audio retrieval, most of the early systems use audio retrieval by metadata (artist, song title, album title etc.) [11], content based retrieval either via converting audio signals into text words or measuring similarity by rhythm and tempo. Researchers have also used annotation-based approaches for audio retrieval.

Multimodal information retrieval (MMIR) is about the search for information in any modality (text, image, audio and video) on web, using combinations of two or more retrieval models. A novel unified framework for multimodal search and retrieval by introducing a novel data representation for multimodal data in Content Object (CO) is described in [12]. Recent efforts in the field of multimodal retrieval systems have led to a growing research community and a number of academic and industrial projects. Besides focusing on single mode of retrieval systems, latest technologies target on multimodal retrieval engines. Current development explicitly focused on queries such as "Show me the video and related documents related to the input query" or "Give me all media (text, image, audio and video) that contains information about the mouth cancer" come into vogue. In order to support such challenging requests, researcher needs to work on several fronts; some of these are listed 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 paper has been organized in three sections. Section 2 gives an insight of various challenges faced in MMIR systems. In this section, several challenges have been discussed, the focus being semantic gap and Multimodal information fusion. Section 3 discusses future trends in MMIR systems; presents an idea for bridging the semantic gap and finally in section 4, conclusion summarizes the previous work done in MMIR system.

2. CHALLENGES IN MMIR SYSTEMS

MMIR systems still have a number of challenging problems, which will be discussed briefly here. One of the most critical problems is bridging the semantic gap between low-level features such as color, texture, shape, object motion, frequency etc. and high-level user's information need. The challenge is how to bridge the gap between these features used by contemporary multimedia systems and the high-level concepts required by the user like finding all shots containing accidents due to rash drive.

Repeatability, the ability to achieve the same performance when moving from the research test set to general-purpose video streams, is another outstanding challenge [13]. The ability to design these systems to intelligently help users to achieve their needs instead of overwhelming them with inappropriate results is another issue that needs further considerations [14]. Also, the high dimensionality of multimedia features makes the feature space very sparse, leading to generalization problems. A possible solution to the generalization problem is to use as much training data as possible [15].

Search speed is also an issue in particular when dealing with huge amount of multimedia data. It is a challenging task to ensure that the system performance will scale well even when dealing with projects such as the national digital libraries. A suggested solution to the scalability problem is distributed computing in particular in 3-tier architectures. Proper modeling is also essential because many of the current systems lack a sound theoretical basis.

The application of contextual constraints enriches these systems with additional metadata that helps in improving the retrieval performance [16]. The use of context makes MMIR systems both content-based and context-based systems at the same tune. In addition, dynamic selection and fusion of the right modalities for interface is another hurdle that needs to be addressed.

This fusion should be done on the fly and with minimum user intervention. Measuring multimedia data similarity is crucial too. Many systems just use the standard metric system measures to evaluate the similarity of multimedia feature vectors, although humans use completely different models to assess multimedia data similarity [17]. One other challenge is the design of appropriate user interfaces that enable users to issue the query and browse through returned results [15, 18].

2.1 Semantic Gap

The semantic gap refers to the gap between low-level features and high-level semantic concepts. Many papers were surveyed for exploring problem of semantic gap [19]. Despite considerable research effort in this field over the last decade, there has not been any significant success for generic applications. Recently, the research community has accepted the fact that it will probably be impossible to retrieve multimedia content by semantic retrieval for several years [13]. Instead, researchers have started to use relevance feedback and machine learning techniques for bridging the semantic gap, whereas relevance feedback is an invaluable tool in this regard, which in itself has further implications on the query formulation process. Considering, low-level features do not directly express the user's high-level perception of the multimedia content, the query formulation process is even more difficult than in uni-modal retrieval systems [13].

Moreover, the user's search need is dynamic and evolving in the course of a search session. Most often, multimodal searching is explorative in nature, wherein searchers initiate a session and learn as they interact with the system. However, current MMIR systems fail to deal with the dynamic nature of search needs.

2.2 Multimodal Information Fusion

Many multimodal systems have been introduced after Bolt's pioneering work in voice and gesture [20]. The design of a multimodal system is critically dependent on the characteristics of the data as well as the requirement of the application. A set of key points that need to be considered include: information sources, feature extraction method, fusion level, fusion strategy, system architecture, and any background knowledge if at all, needs to be embedded [21]. The selection of information sources is determined by the requirements of the application. Practically, the selected modalities should be capable of providing discriminatory patterns for classification, can be collected quantitatively, and from a practical point of view, the system should be low cost and computationally efficient. Some applications may have specific qualifications, such as the universality, stability, and user-acceptability [22]. For example, a multimodal recognition system may utilize speech and video for security and consumer service applications. However, Fingerprint, face recognition and RFID techniques may be more appropriate for surveillance applications.

The effective integration of multimodal information is a challenging problem, with the major difficulties primary to the identification and extraction of complementary and discriminatory features, and the effective fusion of information from multiple channels. The extraction of features that can truly capture the universal characteristics of the intended perception, as well as offer distinctive representation against other perceptions, is of fundamental importance in a pattern recognition problem. In general, a preprocessing step is first applied to the original signal to reduce noise or detect the region of interest. For instance, noise reduction and echo cancellation for speech signals, and face detection for a face recognition system. Feature extraction techniques are then employed on the preprocessed signal to extract a compact representation. Note that depending on the specific problem, even for the same type of signal, different feature extractors are needed, such as prosodic features are usually believed to be the major indicators of human emotion in speech, while phonetic features are usually used for speech recognition. In addition, when a large number of features are extracted, a feature selection procedure might be necessary to reduce the dimensionality of the feature space, whilst selecting those most discriminatory ones.

The fusion of multimodal information is usually performed at three different levels: data/feature level, score level, and decision level [23]. Data/feature level fusion combines the original data or extracted features through certain fusion strategies. One major drawback of fusion at this level is due to the problem of 'curse of dimensionality', which is usually computationally expensive and requires a large set of training data [24]. Furthermore, the fusion problem may be difficult

due to the disparate characteristics of features extracted from different modalities, such as the minute points for fingerprints and principal component analysis (PCA) features for face images [25]. Fusion at score level combines the scores generated from multiple classifiers using multiple modalities through a rule based scheme. A score normalization process is usually required to scale the scores generated by different modalities in the same range, such that no single modality will over power the others, and the significance of each individual modality is leveraged in the final decision. Alternatively, the score level fusion can be conducted in a pattern classification sense in which the scores are taken as features into a classification algorithm. Fusion at decision level is rigid due to the limited information left. Moreover, it generates the final results based on the decision from multiple modalities or classifiers using methods such as majority voting. Fusion at score and decision level can be considered as special cases of fusion at data/feature level [21].

In general, the main challenges in multimodal information fusion include discriminatory feature extraction and selection, redundancy identification and elimination, information preserving fusion and computational complexity. In particular, the fusion strategy should be capable of taking full advantage of information collected from multiple sources and bearing a better description of the intended perception. An ill-designed multimodal system will possibly produce degraded performance and lowered feasibility.

3. FUTURE TRENDS

Relevance feedback is a technique that allows the user to associate his/her ranking of the returned results. This ranking is further used by the system to improve the retrieved results in the second round [26]. Relevance feedback is an example of exploiting the fact that MMIR systems are smart search tools operated by human users. The aforementioned concept needs to be a common feature in all-new system & Web-based Video search engines can be the killer application in this field [27]. The design of scalable techniques of streaming video over the Internet is a promising trend because the user might be interested only in part of the video not the entire sequence. Moving towards more specialized vertical areas are occurring too. Furthermore, real-time interactive mobile technologies are evolving introducing new ways for people to interact.

Overcoming the privacy and intellectual property rights that hinder sharing data required for designing effective benchmarking systems is another area that should be seriously searched, incorporating the user intelligence through Human-Computer Interface (HCI) techniques and information visualization strategies needs more work. Tools for selecting/pasting document parts or objects are also very important to access the inside-document content. Besides, handling the cases where greater recall accuracy is required should be investigated more.

Overcoming the semantic gap between low level features and high level features can enhance the MMIR systems and many other systems like search engines, question answering systems, surveillance systems, robotics etc. we are working to fill this semantic gap by analyzing the relativity between different objects in images to enhance the MMIR systems.

Other areas include initiatives and standards for inter-operability between networks, augmenting the metadata based systems with more content analysis techniques to be applied in a contextual manner and effective integration of media sharing, annotation, search, and management techniques. Improving the summarization and browsing capabilities of current MMIR systems is a trend that deserves further investigation. Finally, some promising browsing techniques use three-dimensional representation of the search results in a trial to improve the user visualization.

4. CONCLUSION

A brief introduction to the basic concepts of Multimodal Information Retrieval (MMIR) systems is presented in this paper with emphasis on state of the art, challenges and future trends.

The paper also illustrates the essential needs for MMIR systems and expounds the reasons behind their pervasive use. This paper along with possible solutions and promising future wends in the field surveyed numerous current challenges faced by MMIR systems.

5. 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.

[11] J. C. Haartsen et. al., "Adaptive Display for Enhancing Audio Playback.", U.S. Patent Application 12/209,300, filed September 12, 2008.

[12] A. Axenopoulos et al., "I-SEARCH: A Unified Framework for Multimodal Search and Retrieval", in Proc. Future Internet Assembly, 2012, pp.130-141.

[13] Urban, Jana, J. M. Jose, and C. J. V. Rijsbergen. "An adaptive technique for content-based image retrieval", Multimedia Tools and Applications 31 no. 1, 2006, 1-28.

[14] S-F. Chang, "Multimedia access and the state of the art and future directions", Proc. of ACM Multimedia, Nov. 1999, 443-445.

[15] A. Jaimes et al., "Multimedia Information Retrieval: What is it, and why isn't anyone using it?" ACM MIR-05, Singapore 2005.

[16] M. Davis. S. King. N. Good, and R. Sarvas, "From context to content: leveraging context to infer media metadata", Proc. of ACM Multimedia. Oct. 2004, 188-195.

[17] W. Farag and H. Abdel-Wahab, "A human-based technique for measuring video data similarity", Proc. of the 8th IEEE International Symposium on Computers and Communications, Turkey, 2003, 769-774.

[18] A. Hauptmann and M. Christel, "Successful approaches in the TREC video retrieval evaluations", Proc. of ACM Multimedia, Oct. 2004, 668-675.

[19] N. Aggarwal, Dr. Nupur Prakash and Dr. Sanjeev Sofat, "Mining Techniques for Integrated Multimedia Repositories: A Review", BVICAM'S International Journal of Information Technology, Issue 1, January-July, 2009 Vol.1 No.1.

[20] R.A. Bolt, "Put that there: voice and gesture at the graphic interface", Computer Graphics 1980, Vol. 14, No. 3, pp.262–270.

[21] M.M. Kokar, J.A. Tomasik, and J. Weyman, "Formalizing classes of information fusion systems", Information Fusion 2004, Vol. 5, No. 4, pp.189–202.

[22] A. K. Jain, and A. Ross, "Multibiometric systems", Communications of the ACM, Special Issue on Multimodal Interfaces 2004, Vol. 47, No. 1, pp.34–40.

[23] A. Ross, and A.K. Jain, "Multimodal biometrics: an overview", Proceedings of EUSIPCO, 2004, pp.1221–1224.

[24] J. Kludas, E. Bruno, and S.M. Maillet, "Information fusion in multimedia information Retrieval", Proceedings of 5th International Workshop on Adaptive Multimedia Retrieval: Retrieval, User, and Semantics, Paris, France, 2007, pp.147–159.

[25] Y. Wang and A. N. Venetsanopoulos, "Information Fusion for Multimodal Analysis and Recognition", Multimedia Image and Video Processing, Second Edition. Mar 2012, 153 -171.

[26] X. Zhou and T. Huang, "Relevance feedback in content-based image retrieval some recent Advances", Proc. Of the 6th Joint Conf. on Information Sciences. 2002. 15-18.

[27] R. Sarukkai, "Video search: opportunities and challenges", Proc. of the ACM workshop on Multimedia information Retrieval, 2005, 3-8.