A User-Oriented Multimodal-Interface Framework for General Content-Based Multimedia Retrieval

Jinchang Ren[†‡] Theodore Vlachos[†] Vasileios Argyriou[†]

[†] *Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U. K.*
{j.ren,t.vlachos,v.argyriou}@surrey.ac.uk
[‡] *School of Computer Science and Engineering, Northwestern Poly. Univ. China*

Abstract

A user-oriented multimodal interface (MMI) framework is proposed. Considering the complexities of media connotations and uncertainties of the user's demands, content-based retrieval has intrinsic requirements for MMI for effective media-content inter-actions. Through integration of knowledge based conduction, learning of semantic concepts, natural language processing and analysis of users' profiles, our framework can establish a solid basis for design and implementation of general CBR systems satisfying extensibility, condensability and inter-operability.

1. Introduction

With the explosion of digital multimedia and the rapid development of communication networks, efficient and effective storage and retrieval of multimedia data has become a challenging topic. Content-based retrieval (CBR) is a key tool for such applications. Compared with the traditional text-based retrieval (TBR), CBR can save manual effort and improve the efficiency and effectiveness by avoiding massive subjective and incomplete text annotations. As a consequence CBR has attracted a lot of attention in the research community [1-4].

The attributes of multimedia can be classified into three categories: The first is (fixed) media attributes, such as the data format, spatial-temporal dimensions and their linear or hierarchical representations, etc. The second is visual and aural features, such as colour, texture, shape, motion, tune and tone, etc. The third is perceptual semantics, which is directly associated with the human understanding of the media concerned.

Retrieval of multimedia needs more human-machine interactions (HMI) both to define user requirements and also provide appropriate responses, especially through a combination of multimodal information, to facilitate the efficient representation of visual, aural and text data. Multimodal interfaces can accurately capture users' requirements to improve the reliability and effectiveness of HMI in a more natural way [5-7]. A key problem is how to define and understand the demands of the users and also how to facilitate effective queries and represent the results in a user friendly way.

To address the above issues, a user-oriented multimodal interface framework is proposed. The framework involves conductive query, knowledge-based learning as well as the extraction and analysis of users' profiles, which can be applied for design and implementation of systems suitable for general multimedia retrieval.

2. Multimodal interface techniques in CBR of multimedia

2.1. Typical query methods

For fast and effective multimedia retrieval, a signify-cant number of HMI techniques have involved feature extraction, media representation and interactive query. Some typical query methods are given below [1-2]:

Query by browsing (QBB): This can be random browsing or category-based navigation (CBN). CBN is similar to operations on hypertext scripts in browsing of web pages, which is usually applied in prompt or hierarchical querying of semantics.

Query by example (QBE): Query results are determined by the similarity matching between user-specified example and candidate samples, and the similarity here are calculated by selected features, such as colour, texture or tone, etc.

Query by sketch (QBS): QBS allows users to submit their query by sketch, which is typically used for query of shapes. For example, an image query involves a the specification of a desirable shape, colour etc.

Query by text (QBT): The text used can be a manual annotations, but it also can be extracted through speech recognition, natural language understanding and image recognition. QBT is the extension of simple TBR method (namely QBT-S).

Query by feedback (QBF): Usually, QBF is used in combination with the query methods above. When the query results are delivered to users, QBF allows them to evaluate the results by assigning scores to each item of the results. The system will automatically analyze this feedback and adjust the weights of different features for more accurate retrieval.

0-7803-9332-5/05/$20.00 ©2005 IEEE

2.2. HMI techniques in CBR

In a HMI context, techniques may belong to at least one of four categories each occupying a different semantic level, namely: text-based user interface (TUI), graphic user interface (GUI), multi-media user interface (MUI) and multimodal interface (MMI). The properties of these four categories are summarized in Table 1 [10]. From Table 1 we can see that normally higher-level interfaces will inherit the input and output means from the lower ones, which makes HMI more user friendly with higher adaptability, stronger robustness and more intelligence.

Table 1. Comparison of different HMI techniques

Items	Input device	Output device	CBR method
TUI	keyboard, etc.	B/W monitor, printer, etc	QBT-S QBB
GUI	+mouse, etc.	+colour monitor, printer, plotter, etc	+ QBS
MUI	+microphone, scanner, digital camera, etc.	+speaker, multimedia mixer, etc.	+QBE
MMI	+ readers or recognizers for speech, lips, gesture, fingerprint, and expression, etc.	+VR instruments, controller, etc	+QBF, QBT

In comparative terms, MMI has apparent advantages in CBR of multimedia because media representation and understanding are associated directly with the semantic contents in CBR. General speaking, MUI is only GUI expanded with devices and functions for input and output of multimedia. Compared with MMI, MUI can only implement media interaction externally, while MMI places greater emphasis on interaction based on recognition and the understanding of media contents.

MMI and CBR of multimedia share the same task on recognition of text, image, speech, gesture, face, fingerprint, lip and expression, etc. To improve the effectiveness, CBR places more emphasis on intelligent MMI techniques, including natural language understanding (NLU) and many recognition-based techniques.

3. User-oriented multimodal interface

As multimedia has plenty of connotations, traditional text-based methods appear more and more difficult due to its incomplete and arbitrary annotations. Although feature-extraction based methods can partially satisfy the needs for multimedia retrieval, they are still far from ideal. Indeed the gap between low-level features and high-level semantics makes CBR only suitable for the more advanced users. In contrast, the proposed user-oriented MMI framework for effective CBR is suitable for a more general audience. The diagram of the proposed framework is given in Fig 1 with further descriptions in the following sections.

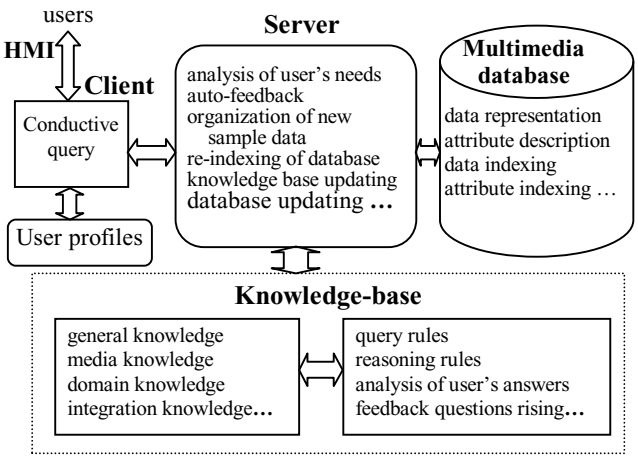

Figure 1. The diagram of user-oriented MMI for general CBR of multimedia.

3.1. Interactive conductive query

The above framework adopts an interactive conductive query approach. This provides a response to a user query via automatic analysis and processing of requirements that is carried out in a user-friendly way. The whole process can be described using natural language a simple example of which is given below:

User: *I'd like to find images with red flowers.*
System: *What is a "red flower"?*
User: **"***A red flower" is a flower in colour of red.*
System: *What is a "flower"?*
User: Define features of flowers, such as specify their shapes, or give examples.
System: *Then what is the colour of "red"?*
User: Use a colour palette to define the red colour or give examples.
System: *Ok, now I know what is the colour of red and what is a red flower, the query results are given as follows.*

In the example above, we assume that the system already knows the concept of "colour" and the usage or operations of it, i.e. colour can be used to qualify objects. If the system has defined the colour of "blue" and has the composite concept of "blue sky", it can automatically generate new concepts of "blue flower" and "red sky" through knowledge-based reasoning. If the system has no concept of "colour", then the user should define it as well as its operations and attributes. Such an approach has good correspondence with human as well as machine intelligence and reasoning and also generalizes and extends rather easily.

3.2. From low-level features to high-level semantic concepts

With user-defined objects and concepts, extensibility brings CBR more in accord with users' requirements and has good potential for different application domains. Three ways allowing users to define new concepts and objects are as follows:

The first is abstract definition. It is implemented by the creation of a certain level of hierarchical semantic net and descriptions. Subsequently it can be improved and expanded in further applications with support from relevant tools. This method has strong pertinence but requires a higher-level of user experience and therefore is only suitable for the more advanced user.

The second is definition by example. The entire media is segmented using visual/aural features. Subsequently objects and concepts are defined using specific spatial or temporal segments. For example, we can define entities in an image as a flower, a cat or a house and concepts as jumping, delight, etc. However, it is still very difficult to build robust mapping from low-level features to semantics.

The third is learning by feedback. The user will evaluate the retrieval results, and this feedback will be analyzed to adjust the combination of features for better query results. The system will learn from this process and encourage correspondences between low-level feature combinations and high-level user-specified concepts during the interactive dialogue stages. At present, this method seems the most practical one in CBR of multimedia.

3.3. Similarity learning

Neural network and clustering are two main methods for similarity learning. Moreover, fuzzy-set theory can be introduced in the learning process to enhance stability. Suppose we have two classes of features, namely colour and texture, and each has a group of features. To reduce the scope, we can use two kinds of weights in the learning process, namely inter-class and intra-class weights. Similarity learning is a process to adjust the weights between features of these two classes to achieve higher inter-class distance and lower intra-class distance by analyzing feedbacks of users.

Generally speaking, feedback from users is usually classified into five levels, namely: high correlative, comparatively correlative, no opinion, less correlative and no correlative. Different weights will be assigned to these levels with higher value to those with higher correlations for similarity learning and vice versa.

3.4. Analysis of users' profiles

User profiling can have a strong influence on the performance of a retrieval system. Such a profile may contain many useful items about the users, such as their expertise, subjective opinions and preferences, etc. These can be extracted using a questionnaire or through automatic analysis during the HMI process. This automatic analysis can improve considerably the performance of the CBR system.

Generally, users will only be concerned with a fraction of the contents of the database queried by the retrieval system. According to preferences and interests, simplified local mappings of the system can be established to facilitate querying. This will improve the performance and effectiveness of the CBR system considerably. Furthermore, we can extend the profiles to involve more information about users, such as concepts and methods defined by them as well as other parameters that will facilitate the optimization of weighting during the query. All the information can be stored in the system to improve efficiency and effectiveness.

4. Implementation and discussions

A general CBR system should meet three requirements, namely *inter-operability*, *extensible dynamic structure*, and *condensability* [9]. Inter-operability includes operability between different systems at different levels of media contents. Extensibility and dynamic structure are consistent with progressively building a general CBR system. Moreover, an extensible structure should be able to deal with a wider range of media formats as well as their features and the relevant methods to extract these features, etc. Condensability implies that the whole system can be simplified corresponding to different users by considering their profiles, thus it reflects the user-oriented property in the CBR system.

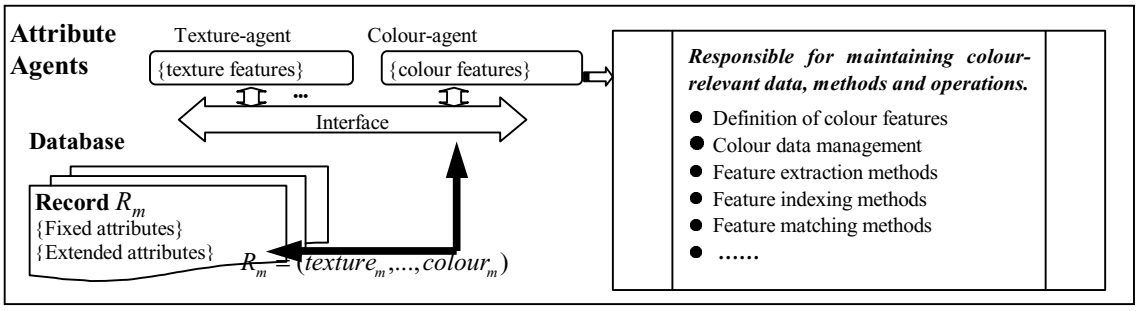

Figure 2. Extensible dynamic structure of ADM

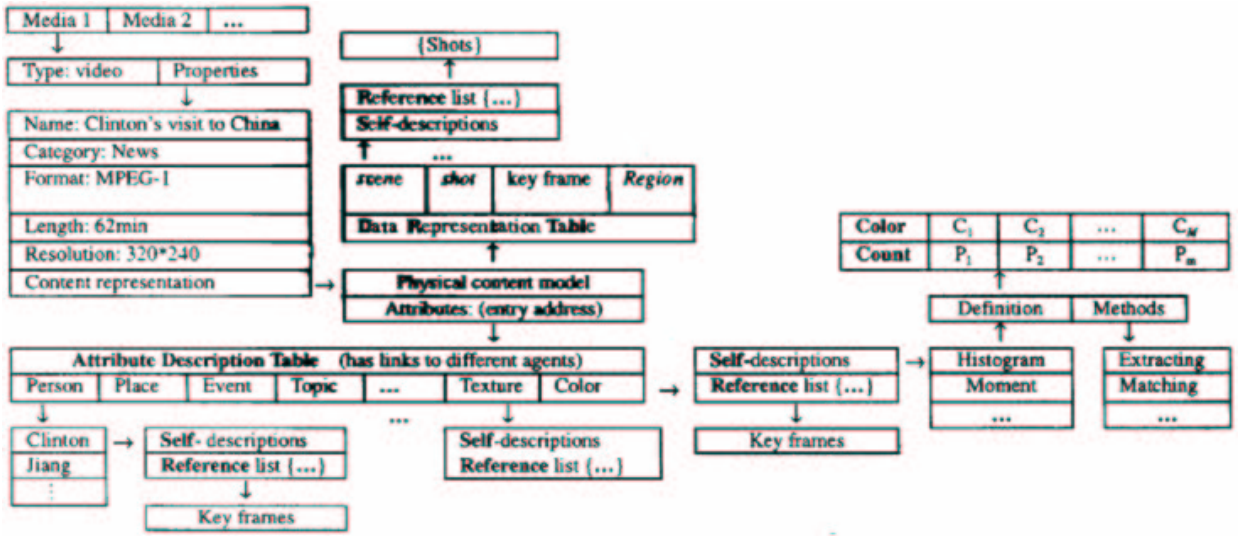

Figure 3. Example of video content description using ADM and SDRL

Inter-operability at the content level can be achieved using standard content-descriptions with the support of MPEG-7. As for extensibility and dynamic structure, an *attribute-dispersed model* (ADM) can be employed, in which different features and attributes are managed separately by corresponding agents for convenience in extension [9]. Fig 2 gives an example of the ADM illustrating how the contents and attributes in the multimedia database correspond to their agents. These are implemented as a set of plugins with standardized interfaces for further processing.

In Fig 2, new features and methods can be easily introduced in the system by adding new agents or new items in existing agents. Although each media has fixed and extended attributes, the extensible structure is only applied to attributes with variable extensions.

A self-description and reference list (SDRL) scheme can be further introduced. Self-descriptions are utilized to define the extended features, such as a scene is a group of similar shots, and thus the scene will be referenced by these shots it contained. Moreover, we can define new features in a CBR system. Fig 3 gives a comprehensive example to illustrate the SDRL scheme in the implementation of ADM.

For each media in Fig 3, its physical content model and extended attributes are defined. The former is for the hierarchical representation of the media, and the latter includes all relevant visual, oral and semantic attributes. Different methods for feature extraction and matching can also be defined. With ADM and SDRL, our proposed framework can be a practical proposition for the implementation of a general CBR system.

5. Summary

We proposed a user-oriented approach for effective and efficient multimedia content retrieval. By virtue of multimodal interactions with NLU, analysis of users' profiles and knowledge-based machine learning, our proposal offers a practical solution towards the implementation of general CBR systems for the non-specialist user and offers important advantages related to extensibility, condensability and inter-operability.

6. References

[1] Y. Rui, T. S. Huang and S. F. Chang. "Image Retrieval: Current Techniques, Promising Directions and Open Issues". J. of Visual Comm. and Image Repre., vol. 10, 1999, pp 1-23

[2] S. Chang, W. Meng, H. Sundaram and D. Zhong. "A Fully Automated Content-based Video Search Engine Supporting Spatial-Temp Acoustical Queries." IEEE Trans. on Cir. and Sys. for Video Tech., 8(5), 1998, pp 602-615

[3] A. Hauptmann and M. Witbrock. "Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval." Intelligent Multimedia Retrieval. Mark, T. Maybury, (ed.), AAAI Press, 1997, pp 213-223

[4] Y. N. Deng and B. S. Manjunath. "Content-based Search of Video Using Colour, texture and Motion." Proc. ICIP, vol 2, 1997, pp 534-537

[5] R. A. Bolt. "The Human Interface." California: Lifetime Learning Press, 1984

[6] H. Hartson, A. Siochi and D. Hix. "The UAN: A User-Oriented Representation for Direct Manipulation User Interfaces". ACM Trans. Infor. Sys. 8(3), 1990, pp 181-203

[7] A. G. Hauptmann and P. Mcavinney. "Gestures with Speech for Graphic Manipulation". Int. J. of Man-Machine Studies. 18(2), 1993

[8] G. Wiggens, E. Miranda, A. Smaill and M. Harris. "A Framework for the Evaluation of Music Representation Systems". J. Computer Music, 17(3), 1993, pp 31-42

[9] J. Ren, R. C. Zhao and D. D. Feng. "ADM: A Dynamic Model for General Multimedia Storage and Content-Based Retrieval". Proc. of ICSP, Beijing. vol II, 2000, pp 1309-12

[10] J. Ren, R. C. Zhao, D. D. Feng and W. C. Siu. "Multimodal Interface Techniques in Content-Based Multimedia Retrieval." Lecture Notes in Computer Science, Springer World, vol. 1948, 2000, pp 634-641