Video Retrieval System for Bridging the Semantic Gap

Min Young JUNG\(^{a}\), Student Member and Sung Han PARK\(^{b}\), Nonmember

SUMMARY We propose a video ontology system to overcome semantic gap in video retrieval. The proposed video ontology is aimed at bridging of the gap between the semantic nature of user queries and raw video contents. Also, results of semantic retrieval shows not only the concept of topic keyword but also a sub-concept of the topic keyword using semantic query extension. Through this process, recall is likely to provide high accuracy results in our method. The experiments compared with keyframe-based indexing have demonstrated that this proposed scene-based indexing presents better results in several kinds of videos.

key words: video representation, semantic retrieval, ontology

1. Introduction

As a rapid development of computer communications and growth of multimedia information continues, multimedia data management becomes increasingly important. Contents-based research has steadily advanced after the development of query by image and video content (QBIC) systems that search multimedia data efficiently [1]. However, the early contents-based methods, such as QBIC systems, focused on automatic information extraction. The meta-data are generated in advance to accelerate the speed of retrieving the lower level features. As a result of extensive research on contents-based video indexing structures, standardization of multimedia feature expression was proposed. It led to the establishment of the standards for moving picture experts group layer-7 (MPEG-7). The information expression and storage method for retrieval was unified [2]. The MPEG-7 standard provides a multimedia description scheme (MDS) that describes both the low-level features and high-level features. However, the relationship between the low and high-level features has not been described. In addition, the overall scene content cannot be accessed because the scene is indexed with a single keyword or content.

To solve this problem, the semantic gap between the low and the high level features needs to be bridged. Therefore a multimedia retrieval system that employs background knowledge structured by ontology or a thesaurus is proposed and evaluated.

There have been several investigations into constructing ontologies using hierarchical tree structures of terms [3]. Structured ontologies can determine the meaning of certain terms and their actual meaning in relation to vocabulary. They can also infer alternate meanings of terms. Examples of these types of ontology are the medical subject headings (MeSH) ontology and gene ontology [5]. These are used mainly in a limited domain. In addition, non-domain specific knowledge structures such as WordNet [6] and Cyc [7] exist. Ontologies have also been used for video retrieval. Hoogs and Hollink describe low-level features and high-level features using extended WordNet, but because WordNet creates a structure that is expensive and complicated, the Hoogs method is not suitable for video retrieval [8], [9]. Moreover, the indexing structure describes only the keyframe image so that only the image, and not the video, is retrieved.

This paper is intended to propose a semantic indexing structure for all senses using an ontology. The remainder of the letter is structured as follows. In Sect. 2, the video ontology system for indexing the semantic scene is implemented. In Sect. 3, a simulation to compare the performance of the proposed method and conventional method is presented. Finally, conclusions are given in Sect. 4.

2. Proposed Video Ontology System

In this section, video ontology system to overcome semantic gap in video retrieval is proposed. As shown in Fig. 1, the proposed video retrieval ontology system is largely composed of four steps: semantic scene creation, video analysis, video ontology construction, and a semantic scene search.
2.1 Semantic Scene Creation

The video retrieval system needs to segment the semantic scene unit, in order to model and retrieve the video in the scene unit. We have defined the basic unit of retrieval. This is called ‘semantic scene’. The semantic scene can be created as follows. First of all, the video is automatically segmented in the shot unit. Segmented shots are automatically grouped using the scene change (SC) algorithm [11] to create video segments. The previous SC algorithm detects the point of the abrupt SC. However, this method do not takes into account gradual scene changes. For this reason, the video segments are semantically insufficient to be the semantic scene. Therefore, the video segments are needed a segment grouping through the user decision for the semantic scene unit when the following conditions applied. Adjacent video segments have same topic. The same topic is that is same object or same space. Using criteria such as user-defined rule, the video will be divided into semantic scene units. The subject of the semantic scene decides in user decision. Namely video is divided in syntactic units.

2.2 Video Ontology Construction

The video ontology is indexing database for video retrieval. Therefore, the video ontology data must be implemented in an early stage for semantic retrieval. The video ontology system is largely composed of the scene name ontology (SNO) and the scene model ontology (SMO). The SNO is indexing for topic of the semantic scene. And the SMO is indexes the low-level features of the semantic scene. The two ontology using key of the main topic are connected each other. The SNO supports only text-based search. The other hand, the SMO can be used for image-based search and automatic scene indexing.

- Scene Name Ontology

The SNO is the dictionary of terms to be used for dealing with the contents of the object and event. Through the hierarchical structure of term, the semantic concepts are defined. Then, the new concept can be derived by defining the hierarchical relationships of these terms. The video indexing process takes advantage of the SNO indexes related to the subject of the semantic scene. For this reason, SNO must be implemented first than video indexing. The SNO internal structure provides the storage for the list of terms (subjects) to be used during the semantic scene indexing. Figure 2 (A) shows how an event relates to the main object. The left child node becomes an event of the main object whereas its right sibling is an object which relates to the main object. The shaded region labeled as (B) depicts the relationship between another object \( O_b \) and the main object. The left child of \( O_b \) stands for an event in relation to its parent node while its right child represents an event of object \( (O_b) \) or to node of object relate with object \( (O_b) \).

Given that the subject which defines the scene in the domain is formed by one term, an inclusion relation between terms is defined. For determining the subject of the semantic scene, the genre information as well as the information about the object and event is used. Indexing and retrieval can be performed within the range of genre domain in which the subject lies by means of the available indexed subject information. Figure 3 shows the example of that the subject of one semantic scene is determined in the SNO. By using ‘drama’, the genre information obtained from the semantic scene segmentation process, ‘Drama’ domain is determined. Then, it is indexed with the subject of ‘walking’ which is contains ‘object_c’ and ‘event_g’, the object and event information in the tree.

- Scene Model Ontology

The SMO is aimed at defining relationships between the visual features and the semantic concepts. The creation stage is pictured in Fig. 4. As I mentioned before, the semantic scene includes the shot-level information. Each shot extracts key frames. The low-level feature is drawn out of the key frame image, which contains the object and the event that defines the relation to key frame. But our system don’t describe visual feature of all object. We store only visual
feature of main object. Namely, main object and event information stored with SMO describes the low-level feature. The object itself is characterized by low-level features such as color, shape and texture. The event is described by the direction and motion of the object. The semantic scene is defined as main topic of scene model.

The metadata structure of SMO not only describes the relationship between the object and the event using MPEG-7 Semantic Ds but also stores low-level feature by means of MPEG-7 visual descriptor.

2.3 Semantic Retrieval

In this section, we present semantic retrieval processing. The user can be inputted a query of two types. One of the query types is text-based query. The other is image-based query. The text-based queries search by using the SNO. We search matching text of main-topic in video ontology. Those semantic scenes with a matching result include not only the concept of the topic keyword but also a sub-concept of it. After a user query is executed, a semantic query will not be created unless the retrieved results do not exist. This is more clearly explained in Table 1. A group of semantic queries give rise to an extended user query in the video ontology. Using user query, the semantic query is created in place of the standard query in the video ontology. If the child node exists in the video ontology, we create the semantic query that is able to retrieve it. In case of a leaf node, a semantic query for retrieving both the parent node and the sibling node is created. In addition, the SNO is composed of domains according to the video genre. The reason why the ontology is configured according to the genre is because the used terms can be perfectly different when the genre is different even though the text remains the same. Since SNO is categorized by genre, it is able to output the search results according to the selected genre.

The retrieval process based on the SMO is described. In this case, user can be inputted image query. Our system searches the semantic scene in the SMO by similarity distance measure [12]. A similarity value (based on color, texture, and motion) is associated with the key frames in the SMO. Moreover, when the number of retrieval results is less than 10 using the SNO, retrieval processing based on SMO execute in semantic retrieval system.

3. Experiment Results

In this section, we present some experimental results. We compared our proposed method, called the semantic-based video retrieval (SBVR) system with other well-known ontology-based video retrieval (OBVR) systems [9]. Conventional OBVR system indexes the low-level and the high-level features of the key frame. The simulated data used in our experiments were NIST Special Database 26 [10]. Our video retrieval experiments were conducted on 100 “use case” queries to uncover the relationship between the numbers of concepts, and retrieval performance. These 100 queries were used because truth annotation for relevant shots was available for each query on the same TRECVID 2005 video development collection, which had also been annotated with truth for the concept lists. The experiments were conducted in such a way that there is an increase in the number of concept. The metadata of the implemented ontology are parsed to be easily retrieved and then stored at the database with a structured table format. The effectiveness of retrieval is usually measured by the following two indicators, recall and precision. There are few differences in terms of precision and test results when the number of concept is not enough. But if constructed concepts are gathered more and more, the tests can yield a better precision favoring the semantic-based retrieval approach. This is due to the fact that the SBVR method defines not only a key frame but also a relationship of shots included in a scene. Hence the SBVR method is able to provide more precise results than the OBVR method as shown in Fig. 5.

The recall measure denotes the actually retrieved rate among all related experiment data concerning video retrieval. The OBVR algorithm is interesting in the sense that it indexes only the key frame while the proposed SBVR method describes the relation of frame in the scene. The SBVR method can retrieve related scenes which are constructed by way of the ontology. Therefore, the proposed method exhibits higher recall rates, as reported in Fig. 6. After adopting the SBVR, we observe that the text is matched precisely. Also, related scenes with sub concepts are correctly retrieved by the SNO structure. That is to say, even though a keyword of the semantic scene does not match, it
can be searched similarly provided it is included as one topic category. Put differently, though a keyword of the semantic scene does not match exactly, only if the scene to be probed is categorized into a single topic, it will be included into the search results.

4. Conclusions

We have proposed a system for ontology-based semantic indexing and retrieval of content. Also, the video can be semantically recognized as the scene unit via the scene unit ontology. If large video ontologies are constructed, videos will seem to be analyzed automatically. We believe that this approach is in the right direction for bridging the current semantic gap of content interpretation between humans and computers, which is the main hurdle for a wide expansion of multimedia in the Semantic Web.

Acknowledgement

This work was supported by National Research Foundation of Korea Grant funded by the Korean Government (R01-2006-000-10876-0).

References

[1] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, H. Qian, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, “Query by image and video content: The QBIC system,” Computer, vol.28, no.9, pp.23–32, Sept. 1995.
[2] J.M. Martinez, “Overview of the mpeg-7 standard,” Technical report 5.0, ISO/IEC, Singapore, 2001.
[3] G. Li, H. Sheng, and X. Fan, “Incorporating metadata into data mining with ontology,” IEICE Trans. Inf. & Syst., vol.E90-D, no.6, pp.983–985, June 2007.
[4] T. Morita, N. Fukuta, N. Izumi, and T. Yamaguchi, “DODDLE-OWL: Interactive domain ontology development with open source software in Java,” IEICE Trans. Inf. & Syst., vol.E91-D, no.4, pp.945–958, April 2008.
[5] H. Tveit, T. Mollestad, and A. Lægreid, “The alignment of the medical subject headings to the gene ontology and its application in gene annotation,” LNCS, vol.3066, pp.798–804, June 2004.
[6] C. Fellbaum, WordNet: An electronic lexical database, Bradford Book, 1998.
[7] D. Lenat and R.V. Guha, Building large knowledge-based systems: Representation and inference in the Cyc Project, Addison-Wesley, 1990.
[8] A. Hoogs, J. Rittscher, G. Stein, and J. Schmiederer, “Video content annotation using visual analysis and a large semantic knowledgebase,” Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol.2, pp.327–334, 2003.
[9] L. Hollink, M. Worring, and A.T. Schreiber, “Building a visual ontology for video retrieval,” Proc. ACM Multimedia, pp.479–482, Nov. 2005.
[10] Available: NIST Special Database 26: NIST Digital Video Volume 1<http://www.nist.gov/srd/nistdvd26.htm>
[11] C.L. Huang and B.Y. Liao, “A robust scene-change detection method for video segmentation,” IEEE Trans. Circuits Syst. Video Technol., vol.12, no.12, pp.1281–1288, Dec. 2001.
[12] L. Wang and L. Khan, Multimedia Data Mining and Knowledge Discovery, Chapter 3, pp.41–57, Springer, 2007.