


METU-MMDS: An Intelligent Multimedia Database System for Multimodal Content Extraction and Querying

Adnan Yazici¹, Saeid Sattari¹, Turgay Yilmaz², Mustafa Sert³,
Murat Koyuncu⁴, and Elvan Gulen⁵

¹ Multimedia Database Laboratory, Department of Computer Engineering,
METU, Ankara, Turkey
yazici@ceng.metu.edu.tr

² Command Control and Combat Systems, HAVELSAN Inc., Ankara, Turkey

³ Department of Computer Engineering, Baskent University, Ankara, Turkey

⁴ Department of Information System Engineering, Atilim University, Ankara, Turkey

⁵ C + E Management, Microsoft Corporation, Redmond, WA, USA

Abstract. Managing a large volume of multimedia data, which contain various modalities (visual, audio, and text), reveals the need for a specialized multimedia database system (MMDS) to efficiently model, process, store and retrieve video shots based on their semantic content. This demo introduces METU-MMDS, an intelligent MMDS which employs both machine learning and database techniques. The system extracts semantic content automatically by using visual, audio and textual data, stores the extracted content in an appropriate format and uses this content to efficiently retrieve video shots. The system architecture supports various multimedia query types including unimodal querying, multimodal querying, query-by-concept, query-by-example, and utilizes a multimedia index structure for efficiently querying multi-dimensional multimedia data. We demonstrate METU-MMDS for semantic data extraction from videos and complex multimedia querying by considering content and concept-based queries containing all modalities.

1 Introduction

With the increasing amount of multimedia data production favored by technological advances and decreasing prices of digital devices, people are exposed to a very large volume of multimedia data in daily life. However, searching such a large volume of data efficiently is a real challenge. Although managing and retrieving textual content (metadata, tag) are relatively straightforward, users are usually interested in the semantic content (i.e., concept) such as objects and events, which are obtained from the various modalities of multimedia data [1]. In addition, complex properties of video data (i.e., multi-dimensional nature, uncertainties and temporal/spatial aspects) present the need for specialized MMDSs to efficiently access and retrieve videos based on their semantic content.

In this study, we develop a competent MMDS to address the following issues:

- Semantic content extraction: The semantic content in videos is very valuable for querying purposes. However, it is not feasible to extract the semantic content of large amounts of digital videos manually. Thus, the automatic extraction of semantic content is a vital issue.
- Multimodal processing: Multimedia data usually has a complex structure containing multimodal information (i.e., visual, audio and textual). During the extraction process, combining the information from different modalities is crucial to improve the extraction performance.
- Modeling & storing data: The data model for MMDS handles low-level and high-level features, information obtained from different modalities (multimodality), and relations (e.g. inheritance) between extracted objects. It is also important to handle the uncertainty existing in multimedia data.
- Powerful querying: In order to satisfy user requirements, the MMDS provides appropriate querying mechanisms like unimodal/multimodal querying, query-by-content (based on low-level features), query-by-concept (based on high-level features) and fuzzy querying (based on fuzzy measures).
- Efficient access structures: Since multimedia data contains a huge amount of information and exists in large sequential file formats, relatively slow responses to user queries are inevitable. Therefore, an efficient multi-dimensional access structure is used in MMDS. For this purpose, both low and high level contents are indexed for efficiently retrieving various types of queries.

Despite several studies exist [1–3], the need for a complete MMDS that combines all required capabilities in a single multimedia database system has not yet been fulfilled [4,5]. Hence, a well-defined system architecture for an integrated multimedia database system covering these requirements is vital.

In this demo paper, we present METU-MMDS to address all of the issues mentioned above. The rest of this paper gives an overview of METU-MMDS system architecture (Sect. 2) and a brief description of our demonstration (Sect. 3) and finally the conclusion (Sect. 4).

2 System Architecture

METU-MMDS is composed of two major components: (i) The semantic content extraction subsystem, and (ii) the storage and retrieval subsystem (Fig. 1).

2.1 Semantic Content Extraction

The semantic content extraction subsystem enables extraction of the semantic contents of raw videos and passes them to the storage and retrieval subsystem. The extraction process starts with the shot-boundary detection of the given videos. The shot-boundary detection is performed by using a Canny Edge Detection algorithm based solution, which calculates Edge Change Ratio between

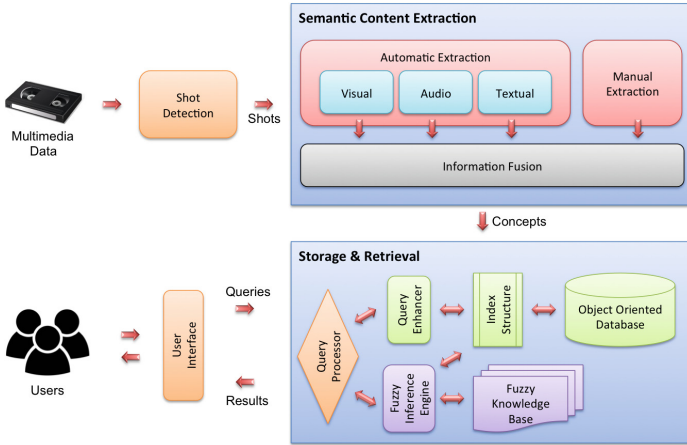


Fig. 1. The architecture of METU-MMDS

video frames [6]. After shot-boundary detection, available modalities (visual, audio and textual) are processed for extracting the semantic content. Once the processing of available modalities are completed, a multimodal fusion process is applied on the extracted content to improve the extraction accuracy. After performing semantic content extraction process, the extracted contents are stored in the storage and retrieval subsystem. For this purpose, the outputs of all modalities (visual, audio and text) and the fusion results are stored in the database separately to provide users with more flexibility in querying the system.

Visual Concept Extraction. The visual semantic content extraction is performed by employing a Class Specific Feature (CSF) [7] and a Support Vector Machine (SVM) based classification approach for object extraction. The object categories are defined with the CSF model and 8 different MPEG-7 descriptors (color, shape and texture based) are used to classify objects using image segments. Prior to the feature extraction and object classification steps, the key-frames of video are partitioned spatially into segments, by employing JSEG segmentation algorithm [8]. After extracting low-level features of these segments, the SVM classifier is utilized to classify these segments into potential objects. The classified objects are then passed to the storage & retrieval subsystem as the visual concepts, as well as the multimodal fusion module.

Audio Concept Extraction. The main task of this module is to extract meaningful audio content from the video data. The module first divides a given audio clip into 1 s consecutive segments and extract low-level audio features from each of these segments. Then, an acoustic classification procedure is carried out to label each of the segments into predefined classes. Then we apply two smoothing rules to reduce possible prediction errors. We use an SVM classifier and a joint

audio feature set due to their success as described in [9]. Finally, the classified segments are aligned with the corresponding video shots.

Textual Concept Extraction. With an intention to combine the advantages of the text modality, i.e., video subtitles, we execute a rule-based named entity recognizer presented in [10] on the video texts. The recognizer extracts named entities of the three basic types, namely, person, location, organization names. Assuming that the keywords meaningful for querying are the named entities occurring the video text, we accept the named entities as textual concepts.

Multimodal Information Fusion. After obtaining the visual, audio and textual concepts, SVM-based multimodal fusion process [11] is applied. The purpose of the fusion module is to integrate the observations belonging to the same class and to utilize the interactions and relations between the observations of different classes captured from independent modalities. This process results in increasing the detection accuracy of the concepts by obtaining new information from exploiting the relations of the concepts.

2.2 Storage and Retrieval

The storage and retrieval subsystem stores the extracted data and answers various types of queries. It consists of an object-oriented database, an indexing structure, a query enhancer, a fuzzy inference engine and a fuzzy knowledge-base. When a query is received by the system, it is firstly processed by the query processor. Based on the type of the query, it is forwarded to the query enhancer or to the fuzzy inference engine. If the query is a unimodal or multimodal query, based on the retrieval of particular concepts, the retrieval is performed by using the index structure and database.

The Database and Data Model. The low-level (feature) and the semantic (concepts) contents, which are extracted by the semantic content extraction subsystem, are stored in an object-oriented database. The object-oriented database design is based on a conceptual video data model developed by considering the special characteristics of multimedia data such as the relations between videos, shots, frames, concepts, low-level features [12]. The database is also capable of storing the content with uncertainty measures for handling fuzzy querying.

Fuzzy Inference and Knowledge Base. The fuzzy inference engine manages the interoperability between the object-oriented database and the fuzzy knowledge-base. The fuzzy knowledge-base is used to define and store knowledge about the video domains. Instead of storing some implicit data in the database, the fuzzy knowledge-base includes some (fuzzy) logical rules that infer new information from the existing data.

Multimedia Index Structure. We make use of a multi-dimensional index structure for efficiently retrieving multimedia data from the database. Our index structure indexes both content and concept descriptors of the multimedia data. This index structure is an adaptation of FOOD-Index [13] in which a spatial indexing technique is incorporated in order to handle both concept-based and content-based queries. The index structure uses a multi-dimensional scaling approach [14] that reduces the retrieval problem to a spatial-indexing task to lower the search space, and thus, querying the multimedia databases is much more efficient than using distance based indexing methods.

Multimodal Query Enhancer. The query enhancer improves the querying and retrieval capability of the system and helps increase the retrieval performance of the system by performing a query-level fusion process. The query-level fusion is based on the idea of exploiting the correlation between different concepts, and provides retrieval of the videos that have correlated concepts in addition to the queried ones [15]. The developed system enables partial match capability for retrieval by utilizing Vector Space Model (VSM). In this context, the query-level fusion exploits the co-occurrence of the VSM terms within a modality and across different modalities [16]. Co-occurrence metrics (calculated by using the Pearson correlation and Canonical Correlation Analysis methods) are used to update the weights of terms in the query.

3 Demonstration Details

Our demonstration allows conference attendees to use METU-MMDS for (i) extracting semantic content from videos, and (ii) querying multimedia data using various types of queries.

For semantic content extraction, the scenario includes uploading a video to the video server, performing shot-boundary detection on the video, key-frame extraction of all shots, visual segmentation of key-frames, visual concept recognition, audio segmentation, audio concept classification, textual concept extraction (named entity recognition) and multimodal information fusion. The defined process is usually performed as a batch job, yet a manual intervention on the visual/audio/textual concept extraction tasks, which enables the system to include human analysis when required and provides a hybrid (manual and automatic) extraction process. The hybrid annotation functionality enables interactive segmentation by defining or selecting automatically calculated minimum bounding rectangles for visual objects, automatic prediction of semantic concepts and reviewing the prediction results. After semantic content extraction, both the low-level features and the high-level semantic concepts are stored in the multimedia database, and indexed by the multi-dimensional index structure.

The demonstration of querying capabilities of the system basically includes queries based on (i) the semantic concepts (query-by-concept) and (ii) the low-level content (query-by-content or query-by-example). The system has an integrated architecture including components for different modalities. Thus, for the

query-by-concept, both unimodal querying (i.e. retrieving videos based on a concept related to a single modality like images, audio clips, or textual documents), and multimodal querying (i.e. retrieving videos based on multiple concepts, each of which belongs to a different modality, like a query of retrieve videos including cars and car horn) are enabled. In addition, for the video documents, queries can be made on fused annotations by using a single query concept. For the query-by-example query type, the query can be performed by giving an example multimedia document (image, audio clip, etc.). The system also supports querying by both specifying a concept and an example document.

Another important capability of the developed system is the improved querying capability provided by the multimodal query enhancer and the query-level fusion mechanism. Such a capability enables comparing the retrieval performance of the improved querying capability with the base multimodal querying functionality. This capability is also included in the demonstration.

4 Conclusion

In this paper, we present METU-MMDS, an intelligent MMDS, for automatic semantic content extraction and querying multimedia data. We overview the architecture and the modules of the system, then briefly explain how the system extracts semantic content from different modalities. We also describe the types of queries supported by the system. We demonstrate the user interfaces to extract semantic content online and its ability to query multimedia data stored in the multimedia database, which is indexed using a multi-dimensional index structure for efficiently retrieving multimedia data from the database.

Acknowledgments. This work is supported by the research grant from TUBITAK with the grant number 114R0182. We also thank to all of the previous researchers of Multimedia Db. Lab. at METU who have contributed to this study.

References

1. Rashid, U., Bhatti, M.A.: Exploration and management of web based multimedia information resources. In: Elleithy, K. (ed.) *Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering*, pp. 500–506. Springer, The Netherlands (2008)
2. Brendan, J., Hongzhi, L., et al.: Structured exploration of who, what, when, and where in heterogeneous multimedia news sources. In: *ACM MM*, pp. 357–360 (2013)
3. Stefanidis, K., Koutrika, G., Pitoura, E.: A survey on representation, composition and application of preferences in database systems. *J. TODS* **36**, 19–45 (2011). ACM
4. Meng, T., Shyu, M.L.: Leveraging concept association network for multimedia rare concept mining and retrieval. In: *ICME*, pp. 860–865. IEEE, Melbourne (2012)
5. Smith, J.R.: Riding the multimedia big data wave. In: *SIGIR*, pp. 1–2. ACM (2013)

6. Aydinlilar, M., Yazici, A.: Semi-automatic semantic video annotation tool. In: Gelenbe, E., Lent, R. (eds.) *International Symposium on Computer and Information Sciences*, pp. 303–310. Springer, Paris (2012)
7. Yilmaz, T., Yazici, A., Yildirim, Y.: Exploiting class-specific features in multi-feature dissimilarity space for efficient querying of images. In: Christiansen, H., De Tré, G., Yazici, A., Zadrozny, S., Andreassen, T., Larsen, H.L. (eds.) *FQAS 2011. LNCS*, vol. 7022, pp. 149–161. Springer, Heidelberg (2011)
8. Deng, Y., Manjunath, B.S.: Unsupervised segmentation of color-texture regions in images and video. *IEEE J. TPAMI* **23**(8), 800–810 (2001)
9. Okuyucu, C., Sert, M., Yazici, A.: Audio feature and classifier analysis for efficient recognition of environmental sounds. In: *ISM*, pp. 125–132. IEEE, USA (2013)
10. Kucuk, D., Yazici, A.: Exploiting information extraction techniques for automatic semantic video indexing with an application to Turkish news videos. *J. Knowl.-Based Sys.* **25**(6), 844–857 (2011)
11. Gulen, E., Yilmaz, T., Yazici, A.: Multimodal information fusion for semantic video analysis. *J. IJMDEM* **3**(4), 52–74 (2012)
12. Kucuk, D., Ozgur, N.B., Yazici, A., Koyuncu, M.: A fuzzy conceptual model for multimedia data with a text-based automatic annotation scheme. *J. IJUFKS* **17**(1), 135–152 (2009)
13. Yazici, A., Ince, C., Koyuncu, M.: Food index: a multidimensional index structure for similarity-based fuzzy object-oriented database models. *J. IEEE Trans. Fuzzy Sys.* **16**(4), 942–957 (2008). IEEE
14. Arslan, S., Yazici, A., Sacan, A., Toroslu, I.H., Acar, E.: Comparison of feature-based and image registration-based retrieval of image data using multidimensional data access methods. *J. TKDE* **86**, 124–145 (2013). Elsevier
15. Safadi, B., Sahuguet, M., Huet, B.: When textual and visual information join forces for multimedia retrieval. In: *ICMR*, pp. 265–272. ACM (2014)
16. Yu, J., Cong, Y., Qin, Z., Wan, T.: Cross-modal topic correlations for multimedia retrieval. In: *International Conference on Pattern Recognition*, pp. 246–249. IEEE, Japan (2012)