RETRIEVING, ADAPTING AND DELIVERING MULTIMEDIA CONTENT USING A MOBILE AGENT ARCHITECTURE

Nikolaos Papadakis, Anastasios Doulamis, Dimitrios Skoutas, Antonios Litke, Nikolaos Doulamis and Theodora Varvarigou

National Technical University of Athens
9, Heroon Polytechniou Str., 15773, Zografou, Athens, Greece
nkpap@telecom.ntua.gr

ABSTRACT

An integrated, reconfigurable, adaptable and open system for mining, indexing and retrieving multimedia information based on a mobile agent technology scheme is presented. The system consists of three integral subsystems, namely the acquisition, the transformation and the distribution modules. Innovative algorithms are used to extract information from web sources, transform it and deliver it to the end users. The system supports efficient content adaptation mechanisms, textual and visual summarization schemes (both sequential and hierarchical), automatic language translation, ontological representation, visual processing and web-based data mining. Experimental analysis on real-life web sites has been performed to test the efficiency of the proposed scheme and compare it with other approaches presented in the literature.

1. INTRODUCTION

Mining multimedia information in the web is in general an arduous task [1], due to the fact that, a) humans perceive media content using high level concepts [2], b) the subjective and vagueness of content interpretation [3], and c) the fact that relevant data are often hidden in a huge amount of irrelevant information. In addition, delivering and distributing the retrieved media information to a wide range of terminal devices of different properties over a wide range of networks to users of different preferences requires new tools and mechanisms for content transformation and adaptation. Other problems concern the language that the data are stored, which may not be the user's preferred language.

Several approaches for acquiring, managing and searching multimedia information exist. ICONS [4] is a web-based system for knowledge-based, multimedia content management relied on artificial intelligence. The MUMMY approach [5], which aims at enabling mobile, personalized knowledge management, provides significant results on content adaptation, annotation of vector graphics images and media integration. The MIND system [6] provides an end-to-end solution, including search in text, image and audio databases, as well as a variety of methods and tools for metadata generation of different media.

In this paper we present a framework for efficient mining, processing, retrieval and indexing of textual and visual data, content adaptation in terms of terminal capabilities, network characteristics and user preferences and multi-lingual content delivery. This integration comprises the main contribution of this paper. The system consists of three modules: acquisition, transformation and distribution and is designed to support two operation modes. The first, called "indirect query", covers the case of a schedule-based retrieval operation, where relevant information is sent to the users according to a service and/or user profile. The second, called "direct query", allows the user to submit a query (either in textual or visual form) and then the system locates and retrieves relevant information with respect to this query.

2. SYSTEM ARCHITECTURE

2.1. Acquisition Module (AM)

The AM searches and retrieves textual and visual data, using mobile agents to transfer the search engine to the data source and execute the search algorithms locally, thus achieving substantial saving in bandwidth. For retrieving textual information from web pages a new hierarchical clustering scheme is applied [7], while for image/video queries, algorithms for visual content description and similarity measures for multimedia data ranking have been incorporated [8]. The AM consists of the *Home Platform* (HP), which starts, monitors and manages the retrieval process, and one or more *Destination Platforms* (DP), which are related with one or more web data sources that potentially contain useful information.

2.2. Transformation Module (TM)

The TM transforms and adapts the retrieved content in various data types, by encapsulating functionalities of textual and visual summarization (both hierarchical and sequential), language identification and translation,

0-7803-9332-5/05/$20.00 ©2005 IEEE

descriptor extraction and query extension using ontological schemes. *Textual summarization* is achieved by assigning a relevant score to each sentence of the retrieved text and then selecting the sentences of the highest relevance score to construct the document summary. For *visual summarization* two different approaches are adopted: *sequential* and *hierarchical* summarization. In the sequential case, a small but meaningful abstract is extracted to represent the content of the entire image sequence ("video abstract" or trailer). In the hierarchical summarization case, the video is organized in a non-linear (non-sequential) way without discarding any visual information. For sequential summarization, a novel scheme is adopted based on a minimization of a cross correlation criterion [9]. Additionally, a new hierarchical content-based video decomposition scheme is proposed which organizes each video sequence at four content resolution levels of hierarchy (the shot representative, the shot, the frame representative and the frame level).

The TM consists of three units: Processing, Transformation and Summarization. The Processing Unit is used only in the "direct query" mode. In case of textual queries, it uses an ontology component to extend the submitted textual query with other relevant information, increasing the retrieval, searching and mining efficiency. In case of visual queries, it extracts appropriate visual features (descriptors) from a query image that will be used to retrieve relevant information. The Transformation Unit converts the retrieved information to alternative data types to match user needs and preferences. For example it can translate textual content to achieve multi-lingual support or transform visual content from one form of representation to another. Finally the Summarization Unit adapts the retrieved content to terminal devices and network channels of different capabilities and characteristics or even different user preferences. Additionally, it provides an efficient framework for organizing, browsing, and managing the retrieved content.

2.3. Distribution Module (DM)

The DM delivers the extracted and adapted textual and visual content to the end users. In case of direct query, the delivery process is rather straightforward. In the indirect query mode, stochastic and deterministic algorithms are used to match the content retrieved from several web sources to end users according to their preferences and interests. To allow delivery of different types of media over a broad variety of network channels and to different types of terminal devices (e.g. through SMS, email, WAP portals etc.) the DM interacts with the TM to transform (i.e. convert and/or summarize) -if necessary- the content to the appropriate format.

3. VISUAL CONTENT MINING

A multimedia object is, in general, characterized by a) a set of features (descriptors) extracted to model its content (since raw visual pixels cannot provide a semantic representation of the visual information as the humans perceive) and b) the similarity metric used for determining how similar or dissimilar two multimedia objects are and ranking them according to their relevance [3], [10], [11]. Both aspects are addressed in our system for multimedia data retrieval.

The visual content is represented by two different types of descriptors. The first type refers to global frame properties. In particular, color, texture and motion histogram, estimated over all pixels of an image, are considered. Object-based descriptors are acquired by first partitioning the image into several objects (regions) by applying a segmentation algorithm. In particular, a multi-resolution implementation of the Recursive Shortest Spanning Tree algorithm (RSST) [12], called M-RSST [13], is used to perform the segmentation task.

The most commonly used similarity measure for multimedia data ranking is the Generalized Euclidean distance. However, the Euclidean distance does not directly express the similarity of two feature vectors and furthermore it is sensitive to scaling or translation. To overcome the aforementioned difficulties, the *cross correlation* of the feature vectors is used as similarity measure. By the term cross correlation we determine a standard method of estimating the degree to which two vectors are correlated indicating, thus, a metric of their content similarity. It is generally used when measuring information between two different data series. The value range of the cross correlation metric is from -1 to 1 such that the closer the cross correlation value is to 1, the more closely the information sets are.

4. MULTIMEDIA SUMMARIZATION SCHEMES

In the case of sequential visual summarization, minimization of a cross correlation criterion is adopted to perform the summarization task [9]. The concept of this approach is to extract as content representatives the most uncorrelated frames as they are described by the respective feature vectors. This is due to the fact that these frames represent the high variations of the content activities. On the other hand, hierarchical summarization organizes the visual content in a non-linear way without discarding any information as sequential summarization does. Hierarchical summarization is performed based on a tree structure scheme. The levels of the tree indicate the resolution to which the content is represented, while the nodes correspond to the segments that the respective resolution is partitioned to. More specifically, four semantically content resolution levels are adopted: the

shot representatives, the shots, the frame representatives and finally the frames (leaves of the tree). The number of the nodes of the tree is optimally estimated so that the total "entropy" measured as the difficulty for a user to find a video segment of his/her interest is minimized [14]. Shot and frame representatives are estimated based on a maximum discrimination criterion by applying a clustering algorithm to video shots or frames. Then, based on the shot or frame representatives, we construct the shot and frame classes at the shot and frame resolution level of the hierarchical tree. Based on the proposed hierarchical summarization scheme, the user can navigate throughout the visual content starting from the lowest (coarse) content hierarchy and ending to the highest (fine) content resolution. The hierarchical summary is described using the Description Definition language (DDL) of the MPEG-7 as an extension of XML Schema.

5. VISUAL CONTENT ADAPTATION

The hierarchical video summarization scheme is suitable for visual content adaptation. A measure is defined for each content resolution level of the tree as a metric of difficulty for transmitting and/or accessing information at this level. Only nodes of a content resolution that satisfy the information constraints for the available network channel are transmitted, whereas the remaining ones are ignored.

Given an information constraint limit, the target of visual adaptation is to estimate the most appropriate level of video hierarchy that should be transmitted. Starting from the lowest (first) tree level, all viewing elements of all nodes of the tree at a given resolution are transmitted only if the constraint information limit is satisfied. The viewing elements are transmitted at a full spatial quality (resolution). In case that the viewing elements of a given level cannot be transmitted at full spatial quality (due to the constraint limit), a spatial reduction is accomplished to fit the imposed constraint. All tree levels below are ignored. The adapted content is described using a DDL file format.

6. RESULTS

The presented system has been evaluated in case of textual queries, for selecting books from several on-line bookstores (e.g. www.randomhouse.com, www.buecher.de, www.barnesandnoble.com etc.) and, in case of visual queries, for retrieving images (or video frames) from a large image/video digital library. During the testing, 50 different web data sources were visited, acquiring more than 63.000 documents and more than 6.5 GB of media content.

Figure 1 presents the average precision-recall curve as obtained by submitting around 3,000 randomly selected queries to the image database. The respective values of the Average Normalized Modified Retrieval Rank (ANMRR) criterion [15] are also shown in Table I. As is observed, despite the complexity of the visual content (large database with high complicated content), the performance of the proposed scheme is quite well. In these figures, we also compare the performance of the presented scheme with the fuzzy color histogram method of [16], [17],[18] and the color histogram technique [18]. In all cases, the proposed scheme yields better performance than the compared ones.

We also examine the performance of the proposed video content decomposition scheme. In a linear video representation scheme, the average number of frames that should be transmitted for a user to access a frame of his/her interest is equal to the half of the total number of frames in the sequence, assuming that all frames present the same probability to be accessed. On the contrary, in any hierarchical video content decomposition and navigation algorithm, only the viewing elements of the selected nodes are transmitted. Consequently, the improvement ratio can be defined as follows:

$$\text{Improvemet Ratio} = \frac{\text{\# of frames transmitted in the sequential approach}}{\text{\# of frames transmitted in a hierachical decomposition approach}}$$

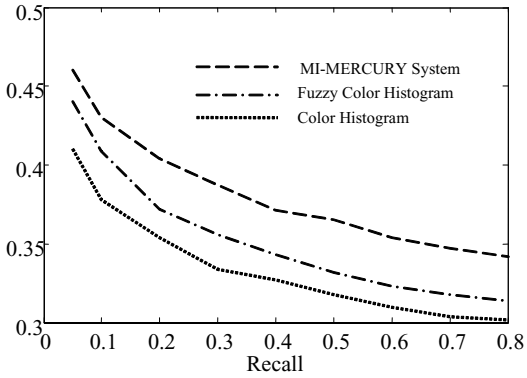

Figure 1: *The Precision-Recall curve obtained in the proposed system over a large image database.*

This ratio is used as an objective criterion for evaluating the performance of the proposed video decomposition scheme and comparing it with other non-linear video representation algorithms. A real user evaluation is considered in the following using three video sequences of total duration of 2 hours and 15 minutes. A randomly selected frame of the sequences is selected each time as user's interest and the transmitted information required for being accessed is calculated. The experiment is conducted by submitting 3,000 randomly selected frames of interest and then estimating the average

transmitted information. Table II presents the results. As is observed, the proposed video hierarchy provides a significant reduction of the difficulty in accessing frames of interest than to the sequential scanning (about 87 times). In this table, we have also compared the performance of the proposed algorithm with other hierarchical approaches for video content decomposition and navigation, such as the ones of [19],[20],[21],[22]. In all cases, the proposed video hierarchy outperforms the compared ones.

Method	ANMMR Values
The proposed Architecture	0.28
Fuzzy Color Histogram [17]	0.34
Color Histogram [18]	0.44

Table I: *The Average Normalized Modified Retrieval Rank (ANMRR) of several algorithms over the examined image database.*

Non-linear Video Representation Algorithms	Improvement Ratio
The Proposed Video Decomposition Algorithm	87.20
MDS Group [21] (MPEG-7)	67.42
VideoZoom [22]	42.30
Yeung-Yeo [19]	24.20
Hanjalic-Zhang [20]	26.13

Table II: *The improvement ratio of the proposed hierarchical video decomposition scheme compared with other approaches presented in the literature.*

7. CONCLUSIONS

An agent-based system for mining textual and visual information from the web has been presented and evaluated. The system incorporates a set of innovative algorithms for multimedia content search and retrieval, multilingual analysis, summarization and adaptation. The use of reconfigurable mobile agents and the integration of multiple components for data mining, and retrieval under a common framework consists of the main originality and contribution of this paper.

8. REFERENCES

[1] O. Etzioni, "The World-Wide Web: quagmire or gold mine?" *Communications of the ACM*, Vol. 39, No. 11, pp. 65-68, 1996.

[2] N. Vasconcelos and A. Lippman, "Statistical Models of Video Structure for Content Analysis and Characterization," *IEEE Trans. on Image Processing*, Vol. 9, No. 1, pp. 3-19, January 2000.

[3] Y. Rui, T. S. Huang, M. Ortega and S. Mehrotra, "Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval," *IEEE Trans. Circuits. Systems for Video Technology*, Vol. 8, No. 5, pp. 644-655, Sept. 1998.

[4] "Intelligent Content Management System" (ICONS), IST Project, 2001-32429, http://www.icons.rodan.pl/

[5] "Mobile Knowledge Management" (MUMMY), IST Project, 2001-37365, http://mummy.felk.cvut.cz/

[6] "Multimedia International Digital Libraries" (MIND), IST project, 2000-26061, http://www.mind-project.org

[7] N. Papadakis, D. Skoutas, C. Raftopoulos and T. Varvarigou, "An Automatic Web Wrapper for Extracting Information from Web Sources, using Clustering Techniques", *International Symposium on Applications and the Internet 2005 (SAINT2005)*, Trento Italy

[8] B. Furht, O. Marques, and Borko Furht, *Handbook of Video Databases: Design and Applications*, CRC Press, Sept 2003.

[9] Y. Avrithis, N. Doulamis, A. Doulamis, and S. Kollias, "Optimization Methods for Key Frames and Scenes Extraction," *Journal of Computer Vision and Image Understanding*, Academic Press, Vol. 75, Nos 1/2, pp. 3-24, July/August 1999.

[10] J.H.Ahrens and G.Finke, "Merging and Sorting Applied to the Zero-One Knapsack Problem", *Operations Research 23,* No. 6, pp. 1099-1109, 1975

[11] Xiang Sean Zhou, T. S. Huang, "Small Sample Learning during Multimedia Retrieval using BiasMap", in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Hawaii, 2001.

[12] O. J. Morris, M. J. Lee and A. G. Constantinides, "Graph Theory for Image Analysis: an Approach based on the Shortest Spanning Tree," *IEE Proceedings*, Vol. 133, pp.146-152, 1986.

[13] A. D. Doulamis, N. D. Doulamis, and S. D. Kollias, "A Fuzzy Video Content Representation for Video Summarization and Content-Based Retrieval," *Signal Processing*, Elsevier Press, Vol. 80, pp. 1049-1067, June 2000.

[14] N. Doulamis and A. Doulamis, "Optimal Content-based Video Decomposition for Interactive Video Navigation over IP-based Networks," *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 4, No.6, pp.757-775, June 2004.

[15] "MPEG-7 Visual part of eXperimentation Model Version 2.0," MPEG-7 Output Document ISO/MPEG, Dec 1999.

[16] EPET-II, "An Intelligent System for Retrieval and Mining of Audiovisual Material," PANORAMA, Greek Ministry of Research and Development.

[17] C. Vertan and N. Boujemaa, "Using fuzzy histograms and distances for color image retrieval," In *Proc. Of CIR'2000*, Brighton, United Kingdom, 4-5 May 2000.

[18] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele and P. Yanker, "Query by Image and Video Content: the QBIC System," *IEEE Computer Magazine*, pp. 23-32, 1995.

[19] M. M. Yeung and B.-L. Yeo, "Video Visualization for Compact Presentation and Fast Browsing of Pictorial Content," *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 7, No. 5, pp. 771- 785, Oct. 1997.

[20] A. Hanjalic and H. Zhang, "An integrated scheme for automated abstraction based on unsupervised cluster-validity analysis," *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 9, No. 8, pp. 1280-1289, December 1999.

[21] ISO/IEC JTC 1/SC 29/WG 11/N3964,N3966, "Multimedia Description Schemes (MDS) Group", March 2001, Singapore.

[22] J. R. Smith, "VideoZoom: Spatio-temporal video browser," *IEEE Trans. on Multimedia*, vol. 1, No. 2, pp. 157-171, 1999.