Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 3, March -2017

A SURVEY ON CONTENT BASED VIDEO RETRIEVAL USING MPEG-7 VISUAL DESCRIPTORS

Kinjal Acharya[1], Prof. Tushar Ratanpara[2]

[1]Dharmsinh Desai University, Nadiad, India
[2]Department of Computer Engineering, Dharmsinh Desai University, Nadiad, India

Abstract — *The volume of the video content grows very fast and most of the video search systems are based on manual annotations or use text information. But such information is not always of high quality or lack precision. In the field of multimedia technology, video retrieval has become one of the fastest growing research areas. Several methods have been developed for retrieval of videos based on extracting their visual features automatically in recent years. The most commonly used low level visual features are colour, texture, shape, motion and spatial-temporal composition. The use of visual descriptors of MPEG 7 is to provide interoperability of system because of its standardization. This paper describes the applications, challenges, methods and limitations of content based video retrieval systems. It also explains the MPEG-7 visual descriptors with their application.*

Keywords- *Content based video retrieval, MPEG-7, shot detection, key frame extraction, visual descriptor.*

I. INTRODUCTION

Content-based video retrieval (CBVR) [1] [2] [3] is the solution to video retrieval problem which in turn is the problem of searching for digital videos in large databases. The meaning of "content-based" to search and analyzes the contents of the video rather than the textual metadata or descriptions associated with the video. The term "content" refers to the low level information such as colours, shapes, textures, or any other can be derived from the video itself. CBVR is desirable because earlier multimedia searches purely rely on metadata which in turn dependent on annotation quality and completeness. When humans manually annotate videos [4] [5] [6] [7] by entering keywords or metadata in a large database, it can become time consuming and may not capture the keywords desired to describe the video. The evaluation of the effectiveness of keyword video search is subjective and has not been well-defined. The large amount of the multimedia content information generates a great need for efficient techniques of finding, accessing, filtering and managing multimedia data. Before retrieving the video based on its content, some pre-processing has to be performed on video like video segmentation, key frame extraction and feature extraction. Multimedia database management systems use the query-by example paradigm to respond to user queries. Users are needed to formulate their queries by providing examples. One of the important issues to be considered in today's multimedia systems is interoperability: the ability of diverse systems and organizations to work together (interoperate) [1]. This is very critical for distributed architectures if the system is to be used by multiple clients. Therefore, MPEG-7 standard as the multimedia content description interface can be employed to address this issue. MPEG-7 is a multimedia content description standard. It was standardized in ISO/IEC 15938 (Multimedia content description interface) [8]. This description will be associated with the content itself, to allow fast and efficient searching for material that is of interest to the user. MPEG-7 is not a standard for encoding of moving pictures and audio. MPEG-7, formally named "Multimedia Content Description Interface," is the standard that describes multimedia content so users can search, browse, and retrieve that content more efficiently and effectively than today's mainly textually annotated search engines. It's a standard for describing the features of multimedia content. CBVR is useful in many applications like news broadcasting, advertising, searching music video clips, distant learning, video archiving, medical applications etc. CBVR systems provide the efficient and more accurate way to retrieve the videos. But it also faces some challenges listed below:

- Very large collection of video data still requires significant time to compute the features.
- Semantic information retrieval is also a major issue in CBVR.
- It is very challenging task to choose the feature that reflects the real human interest. It is difficult for CBVR systems to support multimodal query and allow flexibility for the user to specify its query parameters.
- The CBVR should also provide different search strategies adapted to the type of search to the user.

The CBVR systems are divided into mainly four steps as shown in following Figure 1: shot detection, key frame extraction, feature extraction and similarity measurement.

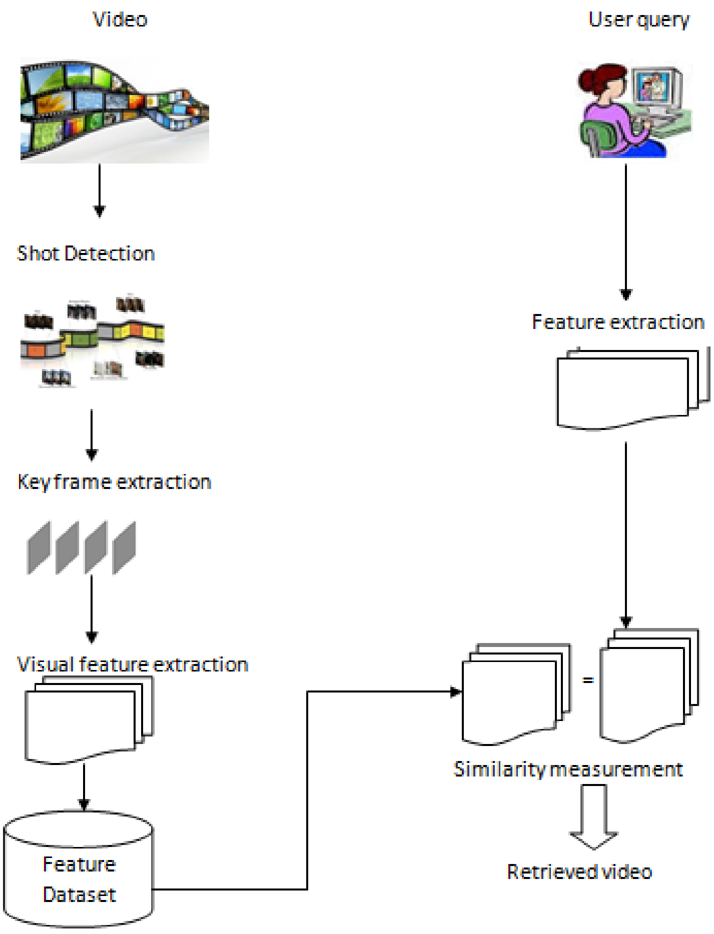

Figure 1. Processing of CBVR systems

II. SHOT DETECTION

It is very essential to segment the video by detecting shots and then extracting key frame from shots for effective video processing and content based video retrieval. A video is made up of number of shots each with boundary properties like cut, fade, dissolve etc [1]. A shot is consecutive frames from the start to end of the recording of video showing continuous action sequence. There are two types of shot boundaries: cut and dissolve [9]. A cut is an abrupt transition caused by video capturing process. A dissolve is a gradual transition caused by video editor where there is a partial overlapping between two consecutive shots and its detection is more difficult than cut. Shot detection method is made up of three steps: feature extraction, similarity measurement and detection. For shot detection, first the features of each frame are extracted, then similarities between the frames using extracted features are measured and finally the shot boundaries are detected between the frames which are not similar. Bin Liang, Wenbing Xiao and Xiang Liu [1] have proposed a CBVR system that consists of three parts: shot boundary detection, feature extraction and similarity measurement. The cut and dissolve have been detected using the histogram difference and skipping image difference, respectively.

A. FEATURE EXTRACTION

The most common features used for shot detection are colour histogram [10], edge change ratio [9], motion vectors etc. Colour histogram is more suitable when the camera motion is small but they cannot differentiate the shots within the same scene [10]. Edge features are more invariant to motion change and illumination change than colour histograms. Motion features can easily handle the object and camera motion. Bin Liang, Wenbing Xiao and Xiang Liu have used the colour structure descriptor and edge histogram descriptor for feature extraction for detecting shot boundaries [1].

B. SIMILARITY MEASUREMENT

The widely used similarity measures [11] are Euclidian distance, chi squared similarity, histogram intersection etc. The similarity can be measured pair wise and window wise. Pair wise similarity computes the distance between consecutive frames while window wise similarity computes the distance between frames within a window. In [1] dynamic weighted feature similarity calculation based on the success rate of the visual similarities of different features is used.

C. DETECTION OF SHOT BOUNDARIES

Shot boundaries can be detected using the measured similarity between frames. The methods of shot detection are of two types: Threshold based and statistical learning. In threshold based approach, the shot boundaries can be detected with predefined threshold. When a similarity is less than a threshold, the shot is detected [11]. The threshold can be global, adaptive or combination of both. The statistical learning method follows the classification task in which the frames are classified as shot change or not as per the features extracted from frame. Here supervised and unsupervised learning can be useful.

III. KEY FRAME EXTRACTION

Key frame extraction is essential in video processing to provide a summarization of video for video indexing, browsing and retrieval. The key frames are useful as they reduce the amount of data and time required for video processing. A key frame is a representative frame that shows the overall content and information of the whole shot. The methods of key frame extraction [11] is categorized as follow sequential comparison based, global comparison based, reference frame based, clustering based, curve simplification based and object/event based. Sanjoy Ghatak and Debotosh Bhattacharjee [12] have presented an approach for key frame extraction done in a totally automatic way without requiring that the user specifying the number of key frames to be extracted as a parameter. It is flexible enough to extract the variable number of key frames. The visually redundant frames i.e. non-significant frames were removed which have been termed as the. Variable numbers of key frames are generated depending on the size of the input video shot. The obtained results generated lesser number of key frames yet they are able to reflect the significant properties of the input video frames. The results are entirely based on the visual details of the input video. Khushboo Khurana and M. Chandak [13] have proposed the method for key frame extraction from video using edge difference. Only when the difference exceeds a threshold, one of the consecutive frames is considered as the key frame. The difference is chosen because the edge is content dependent.

IV. FEATURE EXTRACTION

MPEG-7 offers a comprehensive set of audiovisual description tools in the form of descriptors and description schemes that describe the multimedia data, forming a common basis for applications. The Description Definition Language is based on W3C XML with some MPEG-7-specific extensions, such as vectors and matrices. MPEG-7 documents are XML documents that conform to particular MPEG-7 schemas for describing multimedia content. Descriptors describe features, attributes, or groups of attributes of multimedia content. Description schemes describe entities or relationships pertaining to multimedia content. They specify the structure and semantics of their components, which may be description schemes, descriptors, or data types [14] [8]. MPEG-7 provides four visual descriptors: colour, texture, shape and motion.

A. COLOUR DESCRIPTOR

There are four colour descriptors: dominant colour descriptor, scalable colour descriptor, colour structure descriptor and colour layout descriptor [15]. The small numbers of representative colours of an image are characterized by dominant colour descriptor [16]. The principle clusters are formed by quantizing the pixel colours. The description is made up of the fraction of the image or region represented by each colour cluster and the variance of each one. A measure of overall spatial coherency of the clusters is also defined by dominant colour descriptor. A very compact description of the representative colours in an image is described by this descriptor. The scalable colour descriptor uses HSV colour space. In this descriptor a colour histogram in the HSV colour space is generated, which is encoded by a Haar transform. It has a scalable binary representation. The scalability is in terms of bin numbers and bit representation accuracy, over a broad range of granularity. So to balance the retrieval accuracy, descriptor size is useful. The spatial layout of colour images in a very compact form is represented by colour layout descriptor. In this descriptor, a tiny (8x8) thumbnail of an image is generated, which is encoded via DCT and quantized. This descriptor offers efficient visual matching, a quick way to visualize the appearance of an image, by reconstructing an approximation of the thumbnail, by inverting the DCT. Jian-Hua Li, Ming-Sheng Liu and Ping Song [17] have analyzed the color space used by MPEG-7 color layout descriptor. The HSV colour space is used instead of original color space. The result of HSV was also compared with HSL and YCbCr. The HSV performed better than the HSL and YCbCr. Ka-Man Wong, Lai-Man Po, and Kwok-Wai Cheung [6] have introduced a new mechanism known as Dominant Color Structure Descriptor (DCSD) for efficient representation by combining both the color and the spatial structure information with single descriptor. DCSD has combined the compactness of dominant color descriptor and the retrieval accuracy of colour structure descriptor. The new similarity measure algorithm was also developed based on matching similar colors to generate a common palette, instead of using a fixed histogram space. Experiment results show that DCSD gives better retrieval performance than the original dominant colour descriptor.

B. TEXTURE DESCRIPTOR

There are three texture descriptors: homogeneous texture, edge histogram and texture browsing. The properties of texture in an image (or region) are represented by homogeneous texture descriptor [15]. Here it is assumed that the texture is homogeneous which means that the visual properties of the texture are constant over the region. The texture browsing descriptor is representing homogeneous texture for browsing type applications. This descriptor is usually combined with the homogeneous texture descriptor [18], to provide a scalable solution to represent homogeneous texture regions in images. The edge histogram descriptor [1] [19] is used to represent the spatial distribution of five types of edges (four directional edges and one non-directional). This descriptor generates the bins of local histograms of these edge directions, which may be combined as global or semi-global histograms. When combining edge histogram descriptor with colour structure descriptor gives better result [20]. Hong Shao, Jun Ji, Yan Kang and Hong Zhao [18] have proposed a new texture feature known as wedge feature which has emphasized the texture direction. For CBVR, the homogeneous texture descriptor is more suitable. Aleksey Fadeev and Hichem Frigui [21] have designed a generic approach for representing image texture features approach, called dominant texture descriptor which is based on clustering the local texture features and identifying the dominant components and their spatial distribution. The enhanced version of the DTD (eDTD) was used that encodes the spatial distribution of the pixels within each dominant component. The result was also compared with two well-known descriptors, i.e MPEG-7 Edge Histogram, and Gabor texture. The database used was consists of 900 color images. In feature extraction part, two MPEG-7 visual descriptors, colour structure descriptor and edge histogram descriptor, are used to represent the colour feature and edge feature of the key frames [1].

C. SHAPE DESCRIPTOR

There are four shape descriptors available [8]: region shape, contour shape, shape 3D and perceptual 3D shape. The region shape descriptor is used to specify the region-based shape of an object [22]. The shape of an object is made up of a single region or a set or regions, as well as some holes in the object. Any complex shape can be described using this descriptor as it makes use of all pixels making up the shape. The region-based shape descriptor utilizes a set of ART (Angular Radial Transform) coefficients. A closed contour of a 2D object or region in an image or video sequence is specified by the contour shape descriptor. The object contour-based shape descriptor is based on the curvature scale space [22] representation of the contour. This representation of contour shape is compact on an average with the size below 14 bytes. The Shape 3D descriptor provides shape description for 3D mesh model using some local attributes of the 3D surface. A part-based 3D object representation is expressed as a graph using perceptual 3D shape descriptor to facilitate object description consistent with human perception. This descriptor supports functionalities like query by sketch and query by editing. So this descriptor is very useful for making a content-based retrieval system more interactive and efficient in querying and retrieving similar 3D objects. To extract information about shape and texture [23] feature are much more complex and costly tasks, usually performed after the initial filtering provided by color features.

D. MOTION DESCRIPTOR

There are four motion descriptors available [24]: camera motion, motion trajectory, parametric motion and motion activity. Camera motion descriptor is used to characterize 3D camera motion parameters information that can be automatically extracted or generated by capture devices. The camera motion descriptor supports the well-known basic camera operations like fixed, panning, tracking, tilting, booming, zooming, and rolling. The motion trajectory of an object is a simple, high-level feature which is localized in time and space. This descriptor should be used for content-based retrieval in object-oriented visual databases. The parametric model is associated with foreground or background objects which are defined as regions in the image over a specified time interval. Such description is very efficient for several types of motions, including simple translation, rotation and zoom, or more complex motions such as combinations of the above-mentioned elementary motions. The motion activity descriptor captures the intensity of action or pace of action in a video segment. This descriptor is should be used for video re-purposing, surveillance, fast browsing, dynamic video summarization, content-based querying, etc. Some researches aim to reducing the semantic gap between the visual features and the richness of human semantics [25]. In order to derive high-level semantic features for retrieval, object-ontology [26] was used to define high-level concepts. Supervised or unsupervised learning methods were used to associate low-level features with query concepts [27]. Relevance feedback was introduced into the retrieval loop for learning of users' intentions [5] and semantic templates [4] were generated to support high-level retrieval.

V. SIMILARITY MEASUREMENT

This is the method to retrieve the relevant video according to the query. The similarity is measured between the features extracted from query and the features extracted from the videos stored in the database. Then the videos having minimum distance from the query are retrieved as result. Euclidean distance and chi square distance are more commonly used for similarity measurement. Bin Liang, Wenbing Xiao and Xiang Liu [1] have calculated the similarity between key frames is using dynamic-weighted feature similarity calculation. The system is tested on three kinds of videos. Promising results are obtained in terms of both effectiveness and efficiency [1]. Table 1 shows the comparison of related research work performed in CBVR field.

Table 1. Comparison of CBVR methods

Author(s)	Title	Methods	Dataset	Result
Bin Liang, Wenbing Xiao and Xiang Liu [1]	Design of Video Retrieval System Using MPEG-7 Descriptors	Histogram Difference for Shot detection and key frame extraction, CSD and EHD descriptors for Feature Extraction and dynamic weighted feature similarity calculation	3 different videos of movie, news and sports	Precision and Recall between 0.75 to 0.80
Sanjoy Ghatak and Debotosh Bhattacharjee [12]	Extraction of Key Frames from News Video Using EDF, MDF AND HI Method for News Video Summarization	Colour Histogram, Edge Direction Histogram and Wavelet Statistics	-	-
YannisAvrithis, Nicolas Tsapatsoulis and StefanosKollias Yamada [28]	Broadcast News Parsing Using Visual Cues: A Robust Face Detection Approach	Face Detection using Color Segmentation, Skin –Tone Color Matching and Shape Processing	Six news broadcasts of duration 10 minutes each were recorded at 10frames per second with a resolution of 384×288×24bpp.	Anchorperson shots have the best classification rates. The classification rates for report/interview shots are smaller. Report/interview shots are usually misclassified as outdoor shots and vice versa.
Shafin Rahman, Sheikh MotaharNaim, Abdullah Al Farooq and Md. Monirul Islam [19]	Performance of MPEG-7 Edge Histogram Descriptor in Face Recognition Using Principal Component Analysis	EHD for Feature Extraction, PCA and Semi Supervised Learning for Face Recognition	ORL, Yale and Faces94 face Databases	-
Vojtech Zavrel, Michal Batko and Pavel Zezula [29]	Visual Video Retrieval System Using MPEG-7 Descriptors	Scalable colour, colour structure, colour layout, homogeneous texture and edge histogram descriptors are used.	The database consists of more than 2000 videos. The video resolution is mainly 720 x 576 pixels.	The whole process of extraction on a personal computer with Intel Core2 family CPU took 1275 hours of CPU time. The extraction of key frames took about 1120 hours of CPU time and the extraction of visual descriptors lasted about 85 hours.

VI. CONCLUSION

This paper comprehensively discussed the methods and future advancement in the field of CBVR. In this paper, the methods for content based video retrieval have been explained. The shot detection is performed to extract key frames from the video that represents the summary of whole video. The visual features of key frames can be extracted using MPEG-7 visual descriptors. The query input can be matched with the stored video using similarity measurements. It concludes that remarkable work is done in CBVR and MPEG-7.

But still some enhancements can be done in this area. Along with low level features, some high level features like face recognition can be employed for better retrieval. Still the methods of CBVR require the intervention of human for video annotation, which can be done automatically.

REFERENCES

[1] Bin Liang, Wenbing Xiao and Xiang Liu, “Design of Video Retrieval System Using MPEG-7 Descriptors”, International Workshop on Information and Electronics Engineering (IWIEE), Procedia Engineering, Vol. 29, pp. 2578-2582, 2012.

[2] Chandni Dhamsania, and Tushar Ratanpara. "Human Action Recognition Using Trajectory-Based Spatiotemporal Descriptors." Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications, Springer, pp. 1-9, 2017.

[3] T. V. Ratanpara and M. S. Bhatt,”A novel approach to retrieve video song using continuity of audio segments from bollywood movies”, third international conference on computational intelligence and information technology (CIIT), IET, pp. 87-92, 2013.

[4] Y. Zhuang, X. Liu and Y. Pan, “Apply semantic template to support content-based image retrieval”, Proceedings of the SPIE, Storage and Retrieval for Media Databases, vol. 3972, pp. 442–449, 1999.

[5] Y. Rui, T.S. Huang, M. Ortega and S. Mehrotra, “Relevance feedback: a power tool for interactive content-based image retrieval”, IEEE Transactions on Circuits and Systems for Video Technology 8 (5) , pp. 644–655, 1998.

[6] Tushar Ratanpara and Narendra Patel "Singer identification using perceptual features and cepstral coefficients of an audio signal from Indian video songs." EURASIP Journal on Audio, Speech, and Music Processing, Vol. 2015, Issue 1, 2015.

[7] Tushar Ratanpara and Narendra Patel “Singer Identification Using MFCC and LPC Coefficients from Indian Video Songs”, Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1, Vol. 337, pp. 275-282, 2015.

[8] www.cs.sfu.ca/CourseCentral/820/li/material/source/papers/ mpeg-7-introduction.pdf

[9] Jaspreet Kaur Mann and Navjot Kaur, “ Key Frame Extraction from a Video using Edge Change Ratio”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 5, pp. 1228-1233, May 2015.

[10] C. F. Lam, M. C. Lee, “Video segmentation using color difference histogram,” Lecture Notes in Computer Science, New York: Springer Press, pp. 159–174., 1998.

[11] Mei Huang, Ling Xia, Jin Zhang and Hui Dong,” An Integrated Scheme for Video Key Frame Extraction”, 2nd International Symposium on Computer, Communication, Control and Automation, pp. 258-261, 2013.

[12] Sanjoy Ghatak and Debotosh Bhattacharjee, “Extraction of Key Frames from News Video Using EDF, MDF AND HI Method for News Video Summarization”, International Journal of Engineering and Innovative Technology (IJEIT) Vol 2, Issue 12, pp. 188-194, 2013.

[13] Khushboo Khurana and M. Chandak, “Keyframe Extraction Methodology For Video Annotation”, International Journal of Computer Engineering and Technology (IJCET), Vol 4, Issue 2, pp 221-228, 2013.

[14] Muhammet Bastan and Hayati Cam, “BilVideo-7: An MPEG-7-Compatible Video Indexing and Retrieval System”, IEEE Computer Society, pp. 62-73, 2010.

[15] B. S. Manjunath, Jens-Rainer Ohm, Vinod V. Vasudevan, Akio Yamada, “Colour and Texture Descriptors”, IEEE Transactions On Circuits And Systems For Video Technology, VOL. 11, NO. 6, pp. 703-715, 2001.

[16] Ka-Man Wong, Lai-Man Po and Kwok-Wai Cheung,” Dominant Color Structure Descriptor For Image Retrieval”, IEEE Conference, 2007.

[17] Jian-Hua Li, Ming-Sheng Liu and Ping Song,” An novel modified extraction method of MPEG-7 visual descriptor for image retrieval”, IEEE Conference, 2012.

[18] Hong Shao, Jun Ji, Yan Kang and Hong Zhao,” Application Research of Homogeneous Texture Descriptor in Content-Based Image Retrieval”, IEEE Conference, 2009.

[19] Shafin Rahman, Sheikh MotaharNaim, Abdullah Al Farooq, Md. Monirul Islam, “Performance of MPEG-7 Edge Histogram Descriptor in Face Recognition Using Principal Component Analysis”, IEEE Computer and Information Technology (ICCIT), 13th International Conference, pp: 476 - 481, 2010.

[20] Swapnalini Pattanaik and D.G. Bhalke, “Efficient Content based Image Retrieval System using Mpeg-7 Features”, International Journal of Computer Applications (0975 – 8887) Volume 53– No.5, pp. 19-24, September 2012.

[21] Aleksey Fadeev and Hichem Frigui, “Dominant Texture Descriptors For Image Classification and Retrieval”, IEEE Conference, 2008.

[22] Miroslaw Bober, “MPEG-7 visual shape descriptor”, IEEE transactions on circuits and systems for video technology, vol 11, No. 6, pp. 716-719, 2001.

[23] Sarfraz and M. Ridha “Content-Based Image Retrieval Using Multiple Shape Descriptors”, IEEE/ACS International Conference On Computer Systems and Applications, pp. 730-737, 2007.

[24] Ajay Divakaran, “An overview of MPEG-7 motion descriptors and their descriptors”, Lecture notes in computer science, Springer, vol. 2124, pp. 29-40, 2001.

[25] Y. Lu, C. Hu, X. Zhu, H. Zhang and Q. Yang, “A unified framework for semantics and feature based relevance feedback in image retrieval systems”, Proceedings of the ACM International Conference on Multimedia, 2000.

[26] V. Mezaris, I. Kompatsiaris and M.G. Strintzis, “An ontology approach to object-based image retrieval”, Proceedings of the ICIP, vol. II, pp. 511–514, 2003.

[27] I.K. Sethi and I.L. Coman, “Mining association rules between low-level image features and high-level concepts”, Proceedings of the SPIE Data Mining and Knowledge Discovery, vol. III, pp. 279–290, 2001.

[28] Yannis Avrithis, Nicolas Tsapatsoulis and Stefanos Kollias, “Broadcast News Parsing Using Visual Cues: A Robust Face Detection Approach”, Multimedia and Expo, ICME 2000, IEEE International Conference, Vol. 3, pp 1469 – 1472, 2000.

[29] Vojtech Zavrel, Michal Batko and Pavel Zezula, “Visual Video Retrieval System Using MPEG-7 Descriptors”, SISAP, pp. 125-126, 2010.

[30] Stefan Eickeler, Frank Wallhoff, Uri Iurgel and Gerhard Rigoll, “Content Based Indexing Of Images And Video Using Face Detection And Recognition Methods”, Acoustics, Speech, and Signal Processing, IEEE International Conference, vol.3, pp: 1505 – 1508, 2001.