Content-based Video Information Retrieval and Streaming System using Viewpoint Invariant Regions

박종안*

Jong-an Park*

ABSTRACT

This paper caters the need of acquiring the principal objects, characters, and scenes from a video in order to entertain the image based query. The movie frames are divided into frames with 2D representative images called "key frames". Various regions in a key frame are marked as key objects according to their textures and shapes. These key objects serve as a catalogue of regions to be searched and matched from rest of the movie, using viewpoint invariant regions calculation, providing the location, size, and orientation of all the objects occurring in the movie in the form of a set of structures collaborating as video profile. The profile provides information about occurrences of every single key object from every frame of the movie it exists in. This information can further ease streaming of objects over various network-based viewing qualities. Hence, the method provides an effective reduced profiling approach of automatic logging and viewing information through query by example (QBE) procedure, and deals with video streaming issues at the same time.

Key Word : Image and Video Retrieval, Profile, Multimedia Communication, Key frame, Key Object

Ⅰ. Introduction

With advances in visual technologies, scientists have greatly concentrated on feature retrieval from images and motion picture. The modern day techniques include all-embracing phenomenon for extracting feature sets from structural and color characteristics of images. 3D imagery and stereoscopy have also been the interests for image retrieval. The introduction of HDTV and high and low bandwidth streaming of video content over the internet have proved to be a promising power in steering the direction of research in the fields of computer vision, image processing, neural networks, data mining, and robotics.

One of the best review of CBIR till 2000 is provided by Arnold et al. [2], reviewing some 200 references in content based image retrieval. Content based video compression is very decently described by HongJiang Zhang

* 조선대학교 정보통신공학과(japark@chosun.ac.kr)

접수일자 : 2009.01.11
완료일자 : 2009.02.04
접수번호 : KIIECT2009-01-09

et al. [1] where they provide an idea about using key-objects individually from key-frames of a video. The idea of this paper is rooted in their research. The research in this paper may sound similar to Anil K. Jain et al. [3], which is also based on the research of Zhang et al. [1], but, it only uses their research for acquiring key- objects from the key frames, and rest of the process and the target achieved is entirely different. Similarly, Ling-Yu Duan et al. [4] have proposed a fast search based on index structure of objects in the key frames.

R. Lienhart [6] has given a very thorough account on reliable transition detection in videos. S. Agarwal et al. [7] researched for methods involving learning algorithm for a sparse representation for object detection. Mosaic based clustering of movie scenes algorithm, as devised by A. Aner et al. [8], also proved a major improvement in video retrieval by proposing a shot and scene clustering system. Content-Based Indexing algorithm, proposed by S. Eickeler [9], offer refined results for face detection and recognition systems by indexing faces existing in images and video. S. Lazebnik et al. [12] suggested an affine trans- form based method locating Affine-invariant local descriptors and neighborhood statistics for texture recognition. Similarly, B. Tseng et al. [13] devised a method for persona- lization and summarization of videos. T. Tuyte-laars et al. [14] have also worked on affine- invariant regions and stereo matching.

P. Hong et al. [10] had a good dealing with mining of inexact spatial patterns. R. Lienhart. [11] has produced a very reliable survey of transition detection in videos.

II. Information acquisition process

Amount of information in a movie depends upon the duration, frame size, and quality of the movie. The information retrieval process for the ease of the user can be segmented into two phases: Object recognition and Object Logging.

1. Object recognition

Keeping in mind the complexity of isolating and searching for a specific object, only those calculation methods can be applied which utilize least computation power time. Color based CBIR methods can be thought of for this reason, but, they do not carry information about changes in the orientation of any particular object. A unified solution for content based video retrieval, such as the one described by HongJiang Zhang, et al. [1], can be used in order to locate “key-objects”. A key object is an idea extended from the concept of “key-frame”. A key object consists of region within a key-frame that moves with similar motion.

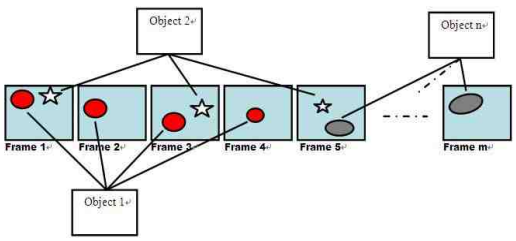


Figure 1. Identification of similar key object occurrences in various key frames.

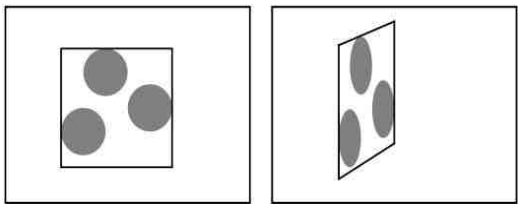


Figure 2. Spatial configuration: a square are centered at affine covariant region.

The key objects are used as the reference objects to be located throughout the movie key frames and are logged. Figure 1 shows the key objects being searched and marked from entire movie, which are logged according to their occurrences in the profile of table 1.

Our experiments show that lookup process for the next occurrences of the key object is remarkably performed by using the concept of viewpoint invariant regions, as described by Josef Sivic and Andrew Zisserman [5].

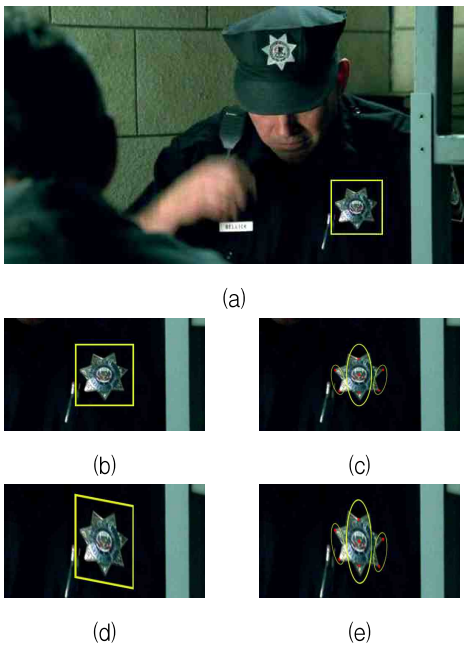


Figure 3. Example of scene from the TV series "Prison Break" (a) A key frame carrying a key object (b) Close-up of a key object from a key frame (c) elliptical region around encapsulating sub-key-objects (d) affine transformation of key object (e) elliptical region around encapsulating sub-key-objects after affine transformation

According to their method, the spatial configuration and extent have to be noted followed by any viewpoint invariant match across the frames of a motion picture. Hence, start from a detected elliptical region p encapsulating the key object in one frame and define it's neighborhood as all detected regions within an area A centered on p. The size of A determines the scale of the configuration, and the neighbors of p. The detected elliptical regions matching p are determined in rest of the frames, and a match between p and p' in a second frame also determines the 2D affine transformation between the

regions, which in turn can be used to map A surrounding p to its corresponding skewed area in the second frame. The neighbors of p', in the second frame, as those elliptical regions lying inside the skewed area are determined. When all the elliptical neighbors of p can be mapped onto correspond- ing elliptical neighbors of p' through affine tra- nsformation between the two neighborhood, it implies a match. This phenomenon is illustrated in figure 2.

The neighborhood of an elliptical region p is the convex hull of its N spatial nearest neigh- bors in the frame and the neighborhood of the matching region p' is the convex hull of its N spatial nearest neighbors. Hence, the two con- figurations are deemed matched if M of the neighbors also matches, where usually M is a small fraction of N.

Josef Sivic et al. [5] also devised a three stage algorithm to efficiently compute the fre- quency of occurrence of the neighborhoods de- fined as above.

1. Neighborhoods occurring in more than a minimum number of key frames are con- sidered for clustering.
2. Significant neighborhoods are matched by a progressive clustering algorithm.
3. Resulting clusters are merged based both on spatial and temporal overlap.

2. Object logging

This information is meant to be stored at the movie hosting or streaming servers as a separate database or in the form of profile bound AVI (movie) files (probably in the form of some new format). Any user who is searching for any particular object over the internet may need this information readily available. Similarly, key objects and key frames from various movies can be used to gather some really important information about occurrences involving similar object from the greatest movie database; the internet.

The information gathered in the previous section (2.1) will eventually prove as the useful information required by the multimedia application users looking for a specific object from one or many movies at

Objects	Occurrences				
O1	Frame	1	2	3	4
	Location	(x11,y11)	(x12,y12)	(x13,y13)	(x14,y14)
	Size	100%	100%	100%	65%
	Orientation	0°	0°	0°	0°
O2	Frame	1	3	5	
	Location	(x21,y21)	(x23,y23)	(x25,y25)	
	Size	100%	100%	70%	
	Orientation	0°	0°	0°	
...	...				
On	Frame	5	6	...	m
	Location	(xn5,yn5)	(xn6,yn6)	...	(xnm,ynm)
	Size	100%	...	...	130%
	Orientation	0°	...	...	18°

Table 1. Profile of object occurrences in figure 1

a time.

Table 1 shows such a profile derived from information in figure 1. There are many frames in figure 1, where similar object is repeated in various sizes and orientations seen at various locations. All of this data and any other data can be stored in the profile.

Redundant objects may be removed from the objects list and placed as the containing frame information (frame number, location, size, and orientation). It will further reduce the size of the profile, and it will surely make it more web databases friendly due to the fact that it will further reduce the amount of data any database engine would require to browse and search during retrieval operation.

Our experiments show that logging such objects in the profile produces better results; however, other methods may be use for object recognition.

III. Streaming and support

The profile described in section 2.2, not only can ease object search from video(s), but also it serves helpful when it comes to dealing with streaming media over various network quality and speed issues.

The objects are sending to the client on priority basis. The priority is defined by taking into account, both, the size of object and its order of appearance over various frame sequences within a video. These objects are then steered remotely from the server only by sending their occurrence information described in table 1.

Another issue, dealing with the time of triggering the object download and steering signals, arises by considering the mentioned procedure. This issue is dealt with by under- standing the network quality and bandwidth, which can be specified by the user at the time of view. Otherwise, some information on the network bandwidth can help any server application, used for streaming the devised method of profile, to adjust automatically by reducing or retaining the streamed object quality to meet the viewing requirements.

This, hence, controls and reduces the network traffic generated due to streaming.

IV. Results

The results were generated both separately and collectively for the whole process. The key frame lookup that results in gathering key objects present a more efficient information retrieval system. At the same time, the precision/recall measurements as described by Josef Sivic et al. [5] were confirmed and drew affective results. The key-object retrieval phase (described in section 2.1) produced very promising results, hence, populating the profile representation with very distinct results. The redundancy check for repeating key-objects (described at the end of section 2.2) will also reduce the data size and refine the existence information for a particular object. Figure 4 shows the retrieval results for a particular key object from figure 3(a) when it is used as query by example (QBE) type searching.

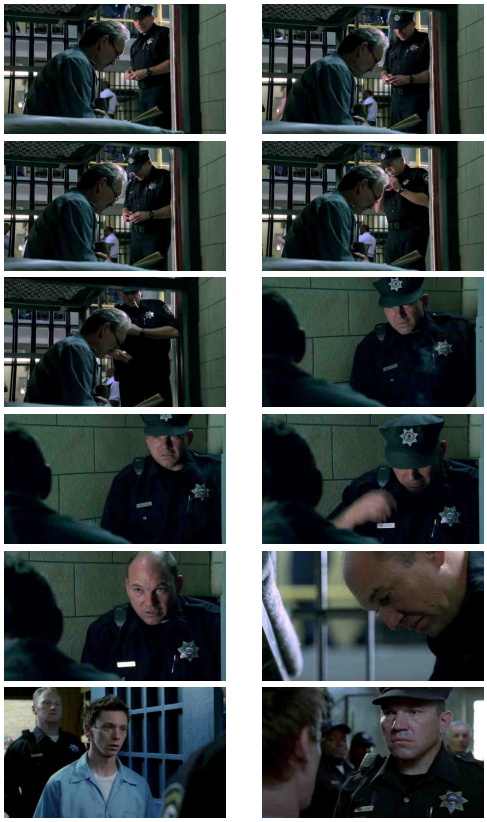


Figure 4. A bunch of frames retrieved using the key object from example in figure 3

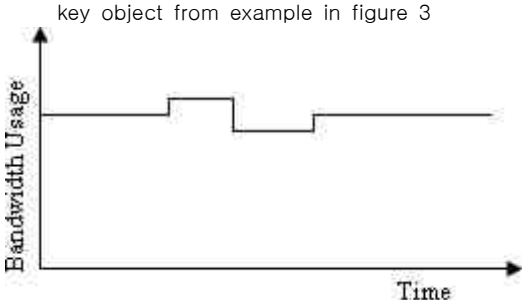


Figure 5. Bandwidth usage while streaming video content over a LAN using existing systems

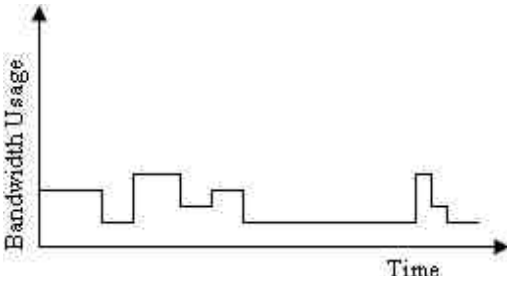


Figure 6. Bandwidth usage while streaming video content over a LAN using proposed algorithm

According to readings generated through our network simulation software for our algorithm, we found out that the method may be equally useful for high definition (HD) and low bit rates. Figures, 5 and 6, show the remarkable difference between the bandwidth consumption through classic video streaming systems against the algorithm proposed in this research.

V. Conclusion and discussion

This paper provides another effective application of QBE systems. The process devised in this paper in order to retrieve images from motion picture is based on very authentic research from some of the most adroit scientists. It can be clearly noticed that all of the potential key objects were mined. The search process is biased towards lightly textured regions. It may seem like the system carries a drawback of big monochrome objects cannot be logged due to lack of texture properties. However, it is noticed that occurrences may be saved from the shape perspective, which does provide reasonable amount of information about monochrome regions.

Also, the streaming mechanism described in this paper can serve as a food for though for systems, where bandwidth bottlenecks are needed to be avoided or reduced. Thus, we can say that this paper introduces a reduced profiling mechanism which simultaneously deals with streaming video quality and computation costs.

Hence, the proposed system can prove fruitful for low bandwidth requiring and consuming multimedia searching and streaming systems, as well as HD systems with less computation latency due to high bit rates and frame sizes.

References

[1] Hong Jiang Zhang, J.Y.A. Wang, Y. Altunbasak. Content-based video retrieval and compression: a unified solution. International Conference on Image Processing (ICIP'97) - Volume 1. p. 13. 1997

[2] Arnold W.M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain, "Content-based image retrieval at the end of the early years," IEEE Transactions of Pattern Analysis and Machine Intelligence, vol. 22, No. 12, pp. 1349-1380, December 2000.

[3] Anil K. Jain , Aditya Vailaya , Xiong Wei, Query by video clip, Multimedia Systems, v.7 n.5, p.369-384, September 1999

[4] Ling-Yu Duan , Jun-Song Yuan , Qi Tian , Chang-Sheng Xu, Fast and robust video clip search using index structure, Proceedings of the 12th annual ACM international conference on Multimedia, October 10-16, 2004, New York, NY, USA

[5] Josef Sivic and Andrew Zisserman. Video Data mining using Configuration of View point Invariant Regions.

[6] R. Lienhart. Reliable transition detection in videos: A survey and practitioner's guide. International Journal of Image and Graphics, 2001.

[7] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In Proc. ECCV, pages 113-130, 2002.

[8] A. Aner and J. R. Kender. Video summaries through mosaic-based shot and scene clustering. In Proc. ECCV. Springer-Verlag, 2002.

[9] S. Eickeler, F. Wallhoff, U. Iurgel, and G. Rigoll. Content-Based Indexing of Images and Video Using Face Detection and Recognition Methods. In ICASSP, 2001.

[10] P. Hong and T. Huang. Mining inexact spatial patterns. In Workshop and Discrete Mathematics and Data Mining, 2002.

[11] R. Lienhart. Reliable transition detection in videos: A survey and practitioner's guide. International Journal of Image and Graphics, 2001.

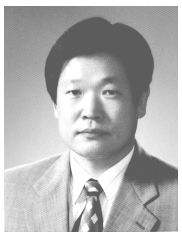
[12] S. Lazebnik, C. Schmid, and J. Ponce. Affine-invariant local descriptors and neighborhood statistics for texture recognition. In Proc. ICCV, 2003.

[13] B. Tseng, C.-Y. Lin, and J. R. Smith. Video personalization and summarization system. In MMSP, 2002.

[14] T. Tuytelaars and L. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In Proc. BMVC., pages 412-425, 2000.

저자약력

박 종 안(Jong-An Park)



1975년 조선대학교 전자공학과 공학사
1978년 조선대학교 전기공학과 공학석사
1986년 조선대학교 전기공학과 공학박사

<관심분야> 정보통신, 디지털신호처리, 멀티미디어 영상처리