A Survey on Effective Video Retrieval using Image Fuzzy Ontology

Jeeva S
PG Student, Dept of CSE,
Sri Vidya College of Engineering and Technology,
Virudhunagar, India

Saranya V
Assistant Professor, Dept of CSE,
Sri Vidya College of Engineering and Technology,
Virudhunagar, India.

Abstract--**In this paper a fuzzy basedapproach is used forEffective retrieval of video.Initially a video ispartitioned into framesand then we use fuzzylogic method in order to retrieve a video.Content based Video Retrieval (CBVR) involves the process for retrieving a set of alike video shots from a large database, with similar content as that of the query video shot. Two significant cues in this context are shape (of the moving object) and motion kinematics (movement pattern), which describe the lowlevel content of a video shot. This finds application in content analysis, video on demand, duplicate detection and incident analysis**

Keywords: **Video retrieval, fuzzy, Histogram, frames, image matching.**

I. INTRODUCTION

With the increasing proliferation of digital video contents [14], efficient techniques for analysis, indexing and retrieval [7], [15] of videos according to their contents have become ever more important.In recent years we have seen atremendous increase in multimedia data. Due to the rapid progress in capturing, acquiring and storing of audio-visualdata the traditional keyword annotation for accessing image or video has the drawback that,apart from large amount of developing annotation it is not efficient to characterize the rich characteristics of image or video using only textFurthermore, the performance of such a system heavily depends on the keywords.

A common first step for most content-based video analysis [14] techniques available is to segment a video into elementary shots, each comprising a continuous in time and space. These elementaryshots are composed to form a video sequence during video sorting or editing with either cut transitions or gradual transitions of visual effects. Shot boundaries are typically found by, computing an image-based distance between adjacent frames of the video and noting when this distance exceeds a certain threshold.

Zadeh at the University of California in the mid 1960s. However, it was not applied commercially until 1987 when the Matsushita Industrial Electric Co. used it to automatically optimize the wash cycle of a washing machine by sensing the load size, fabric mix, and quantity of detergent and has applications in the control of passenger elevators, household applications, and so forth.

Shot boundaries are typically found by, computing an image [9] based distance between adjacent frames of the video and noting when this distance exceeds a certain threshold.

Video indexing [15] is a process of tagging videos and organizing them in an effective manner for fast access and retrieval [11]. Automation of indexing can significantly reduce processing cost while eliminating tedious work [4]. The conventional features used in most of the existing video retrieval [2], [7] systems are the features such as colour, texture, shape, motion, object, face, audio, genre etc.

It is obvious that more the number of features used to represent the data, better the retrieval [11] accuracy. However, as the feature vector dimension increases with increasing number of features, there is a trade off between the retrieval accuracy and complexity. So it is essential to have minimal features representing the videos, compactly

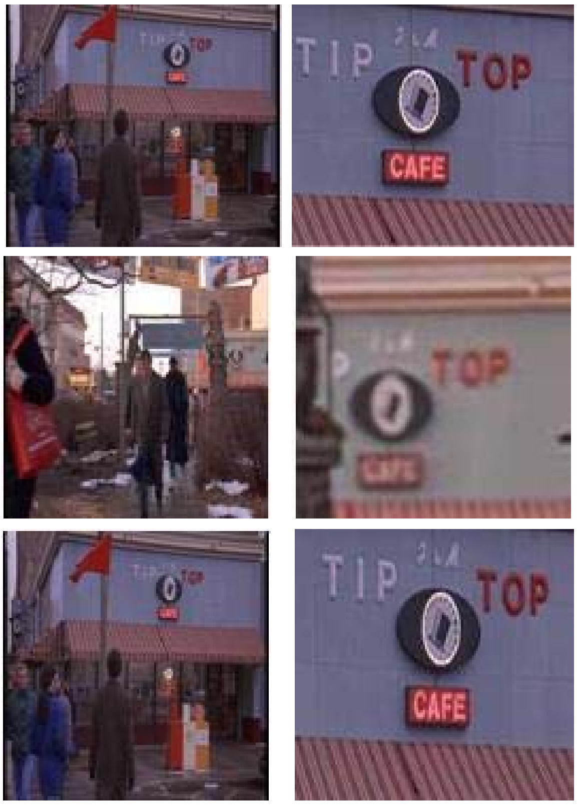

Figure. 1. Object query example I. (a) Top row: (left) a frame from the movie `Groundhog Day'with a query region in yellow and (right) a close-up of the query region delineating the object ofinterest. Bottom row: (left) all 1039 detected affine co-variant regions superimposed and (right) close-up of the query region. (b) (Left) two retrieved frames with detected region of interest inyellow and (right) a close-up of the images with affine co-variant regions superimposed.

These regions match to a subset of the regions shown in. Note the significant change in foreshorteningand scale between the query image of the object, and the object in the retrieved frames. Querying all the 5,640 keyframes of the entire movie took 0.36 seconds on a 2GHz Pentium.

For example, the principal actors will be mined because their face or clothes will appear often throughout a lm. Similarly, a particular set or scene that re-occurs (e.g. Rick'sbar in `Casablanca') will be ranked higher than those that only occur infrequently (e.g. a particular tree by the highway in a road movie).

There are a number of reasons why it is useful to have commonly occurring objects/characters/scenes. First, they provide entry points for visual search in videos andimage databases, or for generating a visual thesaurus. Second, they can be usedin forming video summaries [1, 4, 16].

A third application area is in detecting productplacements in a _lm. Where frequently occurring logos or labels will be prominent. The retrieval [11] and data mining methods will be illustrated for the feature length lms` Groundhog Day' [Ramis, 1993].

II. VIDEO INDEXING AND RETRIEVAL

A. Texture Features

Texture can be defined as the visual patterns that have properties of homogeneity that do not result from the presence of only a single colour or intensity.

Tamura et al (1978) proposed a texture feature extraction and description method based on psychological studies of human perceptions. The method consists of six statistical features, including coarseness, contrast, directionality, line-likeness, regularity and roughness, to describe various texture properties.

Gray co-occurrence matrix (GLC) is one of most elementary and important methods for texture feature extraction and description. Its original idea is first proposed in Julesz (1975). Julesz foundthrough his famous experiments on human visual perception of texture, that for a large class of textures no texture pair can be discriminated if they agree in their second-order statistics. Quantized index frame with GLC matrix is shown in figure 2.

III. COLOUR FEATURES

Colour is one of the most widely used visual features in multimedia context and image / video retrieval [2], in particular. To support communication over the Internet, the data should compress well and be suitable for heterogeneous environment with a variety of the user platforms and viewing devices, large scatter of the user's [5] machine power, and changing viewing conditions. The CBIR systems are not aware usually of the difference in original, encoded, and perceived colours, e.g., differences between the colorimetric and device colour data.

A. Colour Descriptors

Colour descriptors of images and video can be globaland local. Global descriptors specify the overall colour content [14] of the image but with no information about the spatial distribution of these colours.

Local descriptors relate to particularimage regions and, in conjunction with geometric properties of these latter, describe also the spatial arrangement of the colours. In particular, the MPEG-7 colour descriptors consist of a number of histogram descriptors, a dominant colour descriptor, and a colour layout descriptor (CLD)

IV. HIGH-LEVEL SEMANTIC FEATURES

Semantic Gap refers to the difference between the limited descriptive power of low-level index frame features and the richness of user semantics [5], [6].

To support query by high-level concepts, system should provide full support in bridging this'semantic gap' between numerical index frame features and the richness of human semantics [6]. In this survey we have considered the techniques used in reducing the semantic gap into five categories which are most widely used:

A. Using object ontology's to define high level concepts:

For databases with specifically collected images, simple semantics derived based on object-ontology may work fine, but with large collection of images, more powerful tools are required to learn the semantics [6].

B. Using machine learning tools to associate low-level features with query concepts:

The techniques mentioned are Machine Learning, Bayesian Classification [13], Neural Networks, etc. The disadvantage of these techniques is that they require a large collection of image database for learning the data.

C. Introducing relevance feedback (RF) into retrieval loop for continuous learning of users' intention:

Most of the current RF-based systems use only the low-level key frames features to estimate the ideal query parameters and do not address the „semantic" content of the index frame.

D.Generating semantic template (ST) to support High level image retrieval

This technique improves the retrieval accuracy compared totraditional methods using colour histogram and texture features.

V QUANTIZED VIEWPOINT INVARIANT DESCRIPTORS

We build on work on viewpoint invariant descriptors which has been developed for wide baseline matching [12] object recognition, and image/video retrieval [2].The approach taken in all these cases is to represent an image by a set of overlappingregions, each represented by a vector computed from the region's appearance.

Theregion segmentation is designed so that the pre-image of the region corresponds to thesame surface region, i.e. their shape is not fixed, but automatically adapts based on theunderlying image intensities so as to always cover the same physical surface.

Note that the regions are computed independently in each image. In short, the segmentation commuteswith the viewpoint transformation between images, and such regions are knownas *affine covariant* (since the transformation is locally an affinity). Similar descriptorsare computed for all images, and region matchesbetween imagesare then obtained by for example, nearest neighbour matching [12] of the descriptor vectors, followed by disambiguating using local spatial coherence or global relationships (such as a homography transformation). This approach has proven very successful for lightly textured scenes with robustness up to a fivefold change in scale reported in .

A.Affine co-variant regions

In this work, two types of affine co-variant regions are computedfor each frame. The rest is constructed by elliptical shape adaptation about an interest point. The second type of regionis constructed using the maximally stable procedure of Mata s*et al.* where areasare selected from intensity watershed image segmentation. Both types of regions arerepresented by ellipses. These are computed at twice the originally detected region sizein order for the image appearance to be more discriminating. For a pixel videoframe the number of regions computed is typically between 1000-2000.Each elliptical affine covariant region is represented by a 128-dimensional vectorusing the SIFT descriptor developed by Lowe. Combining the SIFT descriptor withaffine covariant regions gives region description vectors which are invariant to affinetransformations of the image.

B. Vector quantized descriptors:

The SIFT descriptors are vector quantized using K-meansclustering. The clusters are computed from 474 frames of the video, with about6K clusters for Shape Adapted regions, and about 10K clusters for Maximally Stable regions.All the descriptors for each frame of the video are assigned to the nearest clustercentre to their SIFT descriptor.

Vector quantizing brings a huge computational advantagebecause descriptors in the same clusters are considered matched, and no further matching [12] on individual descriptors is then required. In an analogy with text retrievalthese vector quantized descriptors are termed

C. Visual words

They provide a vocabulary visual nounsfor representing an object or scene.

D. Stop list

The frequency of occurrence of single words across the whole video (database) is measured, and the top 5% are stopped. This step is inspired by a stop-list in text retrievalapplications where poorly discriminating very common words (such as `the') arediscarded. In the visual word case the large clusters often contain highlights that are distributed throughout the frames.

E.Final representation

The video is represented as a set of key frames, and each keyframe is represented by the visual words it contains and their position. This is the representationwe use from here on for retrieval and data mining. The original raw imagesare not used other than for displaying the results

VI PROBLEMS OF VIDEO RETRIEVAL

A. Internal metadata problem.

All video formats incorporate their own metadata. The title, description, coding quality or transcription of the content [14] is possible. To review these data exist programs like FLV Metadata Injector, Sorenson Squeeze or Cast fire. Each one has some utilities and special specifications.

Keep in mind that converting from one format to another can lose much of this data, so check that the new format information is correct. It is therefore advisable to have the video in lots of formats, so that all search robots will be able to find and index.

Normally a web based search engine bin crawls the web for the video content [14] and if there is pixel mismatch between the input and the related video then it can"t be retrieved.It is estimated that a video is 53 times more likely to end up on the front page of Google than a regular webpage. But you must optimize your video to take advantage of all this attention.

Optimization is simply the process of telling the search engines what your video is really about; labelling it so they can index it in a way that will make it very easy for a person to find.Image optimization is difficult than keyword optimization.

Because search engines are unable to index video in the same way that they index text, you need to tell them what the video is about. A video xml sitemap is simply a file that exists on your site for the search engines to see because they are not human and can"t watch video.

VII NEARDUPLICATE SHOT DETECTION USING EUCLIDEAN DISTANCE SIMILARITY MEASURE

Here, we extend the concept of near-duplicate image detection [10] to video shots. Given a query shot, the task is to find near-duplicate shots in the corpus that contain a large proportion of images that are near-duplicates of images in the query. We achieve this using a variant of the Hough transform. We first initialize a voting table whose size is the number of shots in the dataset. We take each frame from the query shot in turn and search for its near-duplicates, sorting them chronologically (the order they would have appeared in the original video).

Each returned image from the dataset acts like a voting permit and can be used to vote for a particular shot only once for all the images in the query shot. Once all the images in the query shot have been processed, we use an empirically derived threshold on the percentageof votes found in the voting table to return all near-duplicate shots. We could enforce a more stringent shot-level test [1] by examining the temporal ordering between the query and target shots, but empirically we find that this is not needed for accurate detection [10]. The aggregation of many votes makes near-duplicate shot detection quite robust to individual near-duplicate image false positives.

A normalized Euclidean distance (ai, aj) is used to measure the similarity between audio clips from query video shot $i$and database video shot j. $ai$and$aj$are the normalized feature vectors depicting the characteristics of clip $i$and j.

$$d\ (ai,\ aj) = \sqrt{}6\ \Sigma\ k{=}1\ (aik - ajk)/\ 2 \qquad (1)$$

We say $i$and $j$are dissimilar if and only if (ai, aj) >εaudio (a distance threshold) and the greater d, the greater dissimilarity between $ai$and aj.

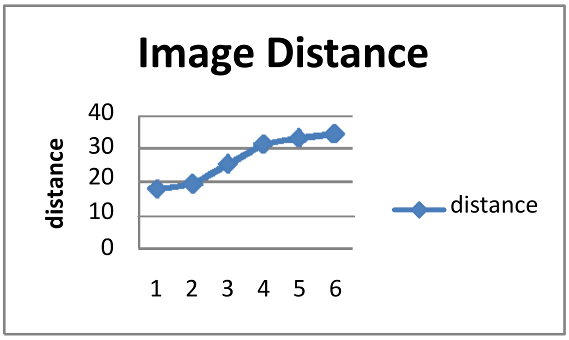

Figure 2. Image distance measure

VIII DATA SET

The data for this study comes from the various search engines, having different requirements on the videos associated with games, Food and Drinks , Education & Reference, computer & internet, travel , social culture , family and the news and events. We have created a pool of 1000 videos over 50 categories are considered as training data set among 300 videos.

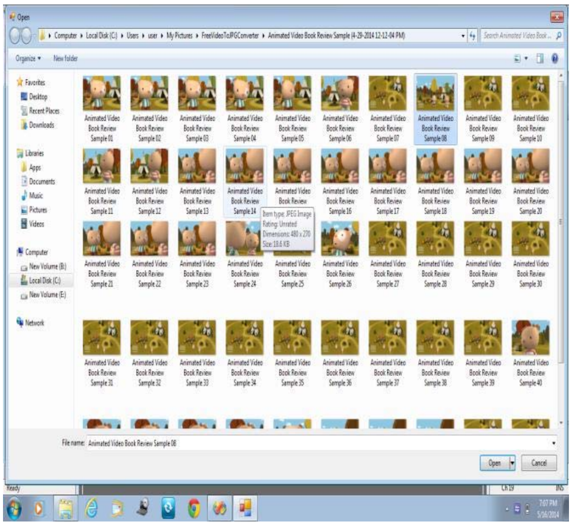
Open
Computer ▸ Local Disk (C:) ▸ Users ▸ user ▸ My Pictures ▸ FreeVideoToJPGConverter ▸ Animated Video Book Review Sample (4-29-2014 12-12-04 PM)
Search Animated Video Book...
Organize ▾
New folder
Favorites
Desktop
Recent Places
Downloads
Libraries
Apps
Documents
Music
Pictures
Videos
Computer
New Volume (B:)
Local Disk (C:)
New Volume (E:)
Network
Animated Video Book Review Sample 01
Animated Video Book Review Sample 02
Animated Video Book Review Sample 03
Animated Video Book Review Sample 04
Animated Video Book Review Sample 05
Animated Video Book Review Sample 06
Animated Video Book Review Sample 07
Animated Video Book Review Sample 08
Animated Video Book Review Sample 09
Animated Video Book Review Sample 10
File name: Animated Video Book Review Sample 08
Open
Cancel

Figure 3 Sample Data set

IX HUMAN JUDGMENT

To complement the user/seeker ratings [5] the human judgments are obtained from users of search engines. Here Cohen„s kappa score is used to evaluate human judgment. Surprisingly our proposed methodology provided amazing results than normal video search engines. For that we had chosen the users of Search engines from YouTube and Google shown in table 9.1. And the comparison chart has been produced based on the evaluation results.

IX EXPERIMENT RESULTS

The proposed fuzzy [9] representation of visual content has been evaluated, using a large database consisting of MPEG coded video sequences and several images compressed in JPEG format. In Figure an image of a space shuttle is submitted as user's query [5]. The retrieval results are displayed in table 9.2 and figure 4 using the fuzzy [9] parameters selected in the previous section. In the same figure a comparison of the proposed method with two other methods is also presented; with a binary classification [13] and the traditional method of colour histogram.

Table 9.1 Results obtained in you tube video search

Users	No of Queries	Retrieved results	You tube (by text)
U1	5	4	0.8
U2	10	9	0.9
U3	15	10	0.67
U4	10	8	0.8
U5	15	13	0.87
U6	20	15	0.75
U7	15	12	0.8
U8	15	12	0.8
U9	10	7	0.7
U10	20	15	0.75
U11	15	11	0.733
U12	10	8	0.8
U13	5	4	0.8
U14	10	8	0.8
U15	15	11	0.73

Table 9.2 Results obtained in Fuzzy logic video search

Users	No of Queries	Retrieved results	Fuzzy logic (by image)
U1	5	5	1
U2	10	9	0.9
U3	15	12	0.8
U4	10	9	0.9
U5	15	14	0.93
U6	20	18	0.9
U7	15	13	0.87
U8	15	14	0.93
U9	10	8	0.8
U10	20	17	0.85
U11	15	12	0.8
U12	10	9	0.85
U13	5	4	0.8
U14	10	8	0.8
U15	15	14	0.93

X RESULT EVALUATION

Cohen's kappa measures the agreement between the two ratters, who classify results into two mutually exclusive categories (satisfied and unsatisfied) Kappa score is defined by,

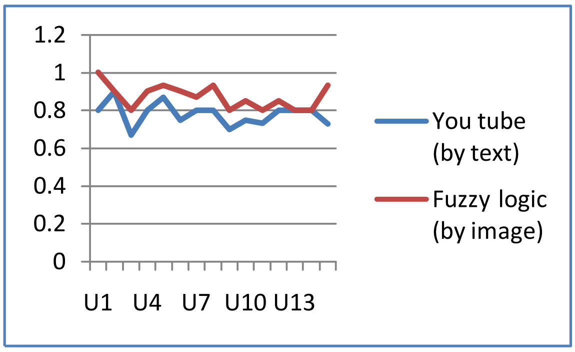

Figure 4 You tube video search & fuzzy logic video search comparison

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \qquad (2)$$

Where Pr (a) is the relative observed agreement among ratters, and Pr(e) is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then κ = 1. If there is no agreement among the raters (other than what would be expected by chance) the score is ≤ 0. Our proposed method Video retrieval using fuzzy logic highly correlated but not exceeding with the human judgments which is shown in table 9.3.

It is very difficult to predict what the user exactly thinks [5] in his mind and also the taste of the human continuously changes based on the environmental factors.

Table 9.3 Human Judgment for proposed methodologies

Method	Video retrieval using fuzzy logic	Video retrieval using text in normal video search engines
Kappa	0.9	0.8

The results for binary classification [13] are illustrated in, where it can be seen that the average performance error in higher in all cases compared to the fuzzy approach. As it can be seen by comparing these values with that presented in the colour histogram performance is worse for any partition number and membership function since only the global image characteristics are taken into consideration.

The computational cost for binary classification is very small and the total cost is mainly affected by the segmentation load, resulting in a similar cost to the fuzzy approach. Instead the colour histogram method demands smaller computational load compared to segmentation. It is observed that the highest cost is for segmentation while the load for fuzzy representation is very small and independent of the image size. Colour histogram requires the lowest cost but yields no sufficient performance for the retrieval.

IX CONCLUSION

In this paper a method to retrieve Video based on Fuzzy approach is proposed on visual content description [14], similarity/distance measures, user interaction and system performance evaluation. It emphasis is on a technique which uses visual contents to search images from large scale image databases according to users' interests, visual feature description technique**.** The computational cost for binary classification [13] is very small and the totalcost is mainly affected by the segmentation load, resulting in a similar cost to the fuzzy approach.

Instead, the colour histogram method demands smaller computational load compared to segmentation. It is observed that the highest cost is for segmentation while the load for fuzzyrepresentation is very small and independent of the image size. Colour histogram requires the lowest cost but yields no sufficient performance for the retrieval.

X. REFERENCES

[1]"Audio-Visual-Based Query by Example Video Retrieval" Sujuan Hou1,2 and Shangbo Zhou1,2

[2]"Content based Video retrieval systems", B V Patel and B B Meshram (2012), International journal of Ubicomp (IJU) vol .3, No.2.

[3] "A Fast Search Algorithm for a Large Fuzzy Database" FengHao, John Daugman, and PiotrZielin´ski

[4] "Comscore"s Qsearch 2.0 Service,"comScore"s Report Article, www.comscore.com, 2007.

[5] B.J. Jansen, A. Spink, and T. Saracevic, "Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web,"Information Processing and Manage-ment, vol. 36, no. 2, pp. 207-227-2009.

[6] K. Barnard and D. Forsyth, "Learning the Semantics [6] of Word and Pictures," Proc. Int"l Conf. Computer Vision, vol. 2, pp. 408-415

[7] Y. Rui and T.S. Huang, "Image Retrieval: Current Techniques, Promising Directions and Open Issues," Journal of visual communication and image representation.

[8]“Fuzzy Block Matching Motion Estimation For Video Compression ”S.M.R. Soroushmehr , ,S. Samavi, , M. Saraee*Isfahan University of Technology*.

[9] “A Fast Search Algorithm for a Large Fuzzy Database” FengHao, John Daugman, and PiotrZielin´ski.

[10] Ondrej Chum, James Philbin , Michael Isard , Andrew Zisserman ”Scalable Near Identical Image and Shot Detection” 1Department of Engineering Science, University of Oxford2Microsoft Research, Silicon Valley.

[11] **“**Content-Based Image Retrieval - Approaches and Trends of the New Age”RitendraDatta, Jia Li, and James Z. Wang

[12]”FCBIR: A Fuzzy Matching Technique for Content-Based Image Retrieval**“**Vincent. S. Tseng, Ja-Hwung Su, and Wei-Jyun Huang.

[13] **“I**nteractive Content-Based Retrieval in Video Databases Using Fuzzy Classification and Relevance Feedback” Anastasios D. Doulamis, Yannis S. Avrithis.

[14] “Applications of Video content analysis and retrieval” Nevenka Dimitrova, Hong-Jiang Zhang, Behzad Shahraray and Ibrahim Sezan. (2002),IEEE.

[15] “A Survey on Visual Content-Based Video Indexing and Retrieval” Weiming Hu and Nianhua Xie, IEEE transactions on systems, man, and cybernetics, vol.41, no.6, 2011.