ISSN 2395-1621

Content Based Video Retrieval

[#1]Pooja Lahane, [#2]Rishabh Balyan, [#3]Kartik Dhimate, [#4]Prof. S.R.Todmal

[1]poojalahane121555@gmail.com
[2]balyan.rishabh@gmail.com
[3]kartikdhimate@gmail.com

[#123]Department of Information Technology
[#4]Prof. Department of Information Technology

JSPM's ICOER, Wagholi, Pune, Maharashtra, India

ABSTRACT

Today the video database is increasing rapidly and making it more difficult to search in that huge database and users are not satisfied with the traditional information retrieval techniques. To overcome this problem, in this paper, we have proposed a technique based on content-based video retrieval system (CBVR), which is an extension of content-based image retrieval system (CBIR), for searching and localizing objects using spatio-temporal localization, along with it how relevance feedback and ranking algorithm can improve the search results.

Keywords: CBVR, CBIR, Object searching, Spatio-Temporal localization, Relevance Feedback.

ARTICLE INFO

Article History

Received: 5th May 2016

Received in revised form :

5th May 2016

Accepted: 10th May 2016

Published online :

12th May 2016

I. INTRODUCTION

Content based Video Retrieval system (CBVR) [9] is the technique of searching and retrieving digital videos in huge databases. "Content-based" means that the search will analyse the actual content which is present in the video. The term 'Content' here refers to the colours, shapes or textures.

The traditional search engine uses the web crawlers to search and index the content. The web crawler uses the links in the web pages for crawling and uses the metadata to describe the content which is provided by the publisher. Then the search engine ranks and stores data in human-powered directories.

Content based video searching is done by following the below steps:

- Firstly, the key-frames [5] are extracted from the video. Key-frames are still images that represent the content of shots (sequence of images that represents continuous action) in an abstract manner. If extracted properly, key-frames are very useful for video browsing effectively and efficiently.
- After extracting the key-frames, the next step is to extract features from those key-frames. Feature can be colour, texture or shape. Then, the features are stored in the database.
- The final step is the similarity measurement between the features of query image and the stored features of the videos. Similarity measurement can be done by calculating the distance between the features. Euclidean distance is the most common metric for calculating the distance. Once the videos are retrieved, ranking algorithm can be used so that the user will get the most relevant videos first.

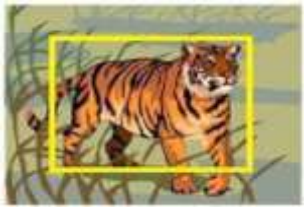
II. BASIC CONCEPT

A. IMAGE

A computer image is a matrix (a two-dimensional array) of pixels. The value of each pixel is proportional to the brightness of the corresponding point in the scene.

B. OBJECT

An object in image processing is a portion of an image that can be uniquely interpreted as a single unit. For example, in the image below the tiger is an object which can be interpreted as a single unit.



© 2015, IERJ All Rights Reserved

C. OBJECT SEARCHING

Object searching in an image/video is a process of identifying similar objects from pre-processed images/videos stored in a database. This is achieved by comparing different features of the query and the pre-processed images/videos.

D. HISTOGRAMS

The intensity histogram [6] shows how individual brightness levels are occupied in an image; the image contrast is measured by the range of brightness levels. The histogram plots the number of pixels with a particular brightness level against the brightness level. For 8-bit pixels, the brightness ranges from zero (black) to 255 (white).

E. VIDEO

Video is a visual multimedia source that combines a sequence of images to form a moving picture. Video usually have audio components that correspond with the pictures being shown on the screen.

III. LITERATURE SURVEY

In [1] 2009, M. Ryooand and J. Aggarwal, discussed about the basic idea to find the structural similarity between the two sets of feature points extracted from two videos and then calculate the spatial and temporal relationships between those points. By comparing those relationships, the spatio-temporal relationships match measures the amount of characteristic two videos contain in common, and exhibit an equivalent relation with each other. The main role of relationships match is to find weather feature points extracted from an activity video generate a particular spatio-temporal pattern in its 3-D XYT space or not. So it uses an extractor which detects each interest points in a video volume and constructs a descriptor which summarizes a small video patch around it. They have used a spatio – temporal relationship kernel which is a Histogram- based match kernel, it measures the similarity by constructing histograms and intersecting them. The major advantage of match kernel is its efficient consideration of spatio-temporal structures among feature points.

In [2] Dec 2009 C.H. Lampert, M.B. Blaschko and T. Hofmann proposed a concept about Efficient Sub-window Search (ESS), a method for localizing objects, for finding the quality function over all possible sub -images in the possible candidate image, thus it return the same object locations that an exhaustive, complex sliding window approach would, It depends on Branch-and-bound scheme. There are few classifier evaluations which are required than there are candidate regions in the images often even less than the number of pixels—and it typically runs faster in linear time**.** The Branch-and-bound scheme and its optimization is used in computer vision for objectives of geometric matching and also to optimize more general object localization, including those based on quantized local features.

In [3] 2011 J. Yuan, Z. Liu, and Y. Wu present a new method called STBB search which is used to search the video space, this new method decomposes into two subspaces:

1) 4D spatial parameter space and
2) 2D temporal parameter space.

The objective here is to find the spatiotemporal sub-volume which has the maximum detection score. They have taken different search strategies in the two subspaces WW and TT and search alternately between them. First, if the spatial window W is determined; we can easily search for the optimal temporal segment in space TT. This relates to the max sub- vector problem, where given a real vector, the output is the contiguous sub-vector of the input that has the maximum sum.

L. Shao, S. Jones, and Xuelong Li [4] have used an alternative method for accessing videos which is known as content-based video retrieval(CBVR) it is an extension of content-based image retrieval (CBIR), CBVR can directly search the content in database's and can return result far more accuratelly than other search techniques. They emphasise to use temporal and spatial localization technique for the efficient search from large proportion of data, thus they have used an efficient algorithm known as Spatio-temporal localization for localizing object from large amount of data. For searching human actions in video database they use, given a query video containing a pre-localized human action, the system searches a video database for all instances of this human action. It spatio-temporally localizes and ranks these actions according to relevance, before returning them to the user. At this point, the user may mark results for relevance feedback and run the query again iteratively, until user is satisfied with the results.

B V Patel and BB Meshram in their paper [9] have selected content based video retrieval as a system for selecting extracted features regardless for videos attribute in it. In the image retrieval the problem is of searching for digital videos in large databases. They have used the term Content-based which symbolize the search of the actual content in the video. The term 'content' refer to colors, shape and textures. Search may rely only on image which is provided by user when video content cannot be examined practically. The first step towards the content based video search is to aim the segment like moving objects in video sequences which is known as Video segmentation. It initially segments the first image frame into some moving objects and then the evolution of the moving objects is tracked in the image frames. Then video sequence is partitioned into shots which are defined as an image in a sequence that presents a continuous action which is captured in a single operation of single camera. Shots are joined together in editing phase to form a complete sequence. The next step is to extract the Key-Frames from the original video data. Key-frames are used to supplement the text of the video log. Once key frames are extracted next step is to extract features. The features are extracted off-line for the efficient computation; there are two types of features low-level and high-level. Low-level features like color, shape, texture and loudness, pitch and object motion are extracted directly from video in the database. Features such as timbre, rhythm and instruments are high level feature which are also supposed to deal with semantic queries. They have also used the concept of Video indexing and Retrieval which a process of tagging videos and organizing them in an effective manner for fast access and retrieval.

© 2015, IERJ All Rights Reserved

IV. SPATIAL SCENE ANALYSIS

This section reviews visual document processing operations which are essential for extracting description of document. Aim of feature extraction is to characterize list of properties for components (pixel, frame) of a video. Some features used in video retrieval systems are color, texture and shape.

A. Texture Feature

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

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

The analysis of textures [11] is based on local neighborhood which corresponds to the basic texture pattern. The analysis can be done by mapping of textures based on the response of pre- defined filters against the image thus allowing the use of similarity measures between these feature vectors.

B. Color Feature

The most widely used visual features in multimedia and image/video retrieval is colour. Colour descriptors [7] of images and video can be *global* and *local*. Global descriptors specify the overall colour content of the image but with no information about the spatial distribution of these colours. Local descriptors relate to particular image regions and, in conjunction with geometric properties of these latter, describe also the spatial arrangement of the colours.

Various techniques which can be used for colour extraction are the conventional colour histogram, the fuzzy colour histogram, the colour correlogram, and a colour/shape-based method.

B1. The Conventional Colour Histogram

The conventional colour histogram (CCH) [6] of an image indicates every colour frequency occurrence of in the image. From the perspective of probability, it refers to the probability mass function of the image intensities. It captures the joint probabilities of the intensities of the colour channels. The CCH can be represented as $hA,B,C(a,b,c) = N.Prob(A=a, B=b, C=c)$, where *A, B* and *C* are the three colour channels and *N* is the number of pixels in the image.

B2. The Colour Correlogram

The colour correlogram (CC) [7] expresses how the spatial correlation of pairs of colours changes with distance. A CC for an image is defined as a table indexed by colour pairs, where the *dth* entry at location *(i,j)* is computed by counting number of pixels of colour *j* at a distance *d* from a pixel of colour *i* in the image, divided by the total number of pixels in the image.

V. TEMPORAL ANALYSIS

The temporal dimension [11] of a video document contains information which is specific to that document. The analysis of the document requires partitioning of video document into basic elements.

The partitioning can operate in different level of granularity.

1. Frame level: Each frame is treated separately. There is no temporal analysis at this time.
2. Shot level: A shot is a set of contiguous frames all acquired through a continuous camera recording.
3. Scene level: A scene is a set of contiguous shots having a common semantic significance.
4. Video level: A complete video object is treated as a whole.

VI. DISTANCE MEASURE

The measure of similarity between the video sequences[12] is derived by distance which either uses temporal deformation of the video or use invariant features under temporal distortions thus frame to frame difference measure is not efficient in comparing generic video sequences.

In the paper the distance measure is calculated using Euclidean geometric [11] measures which calculate the feature based distance between the query media object (q) and the indexed media object (m).

Euclidean (dp):

This metric measure belongs to Minkowski family of distance.

$$d_p(A,B) = \left(\sum_{i=1}^{n} |a_i - b_i|^p\right)^{\frac{1}{p}}$$

The Minkowski distance is general form of the Euclidean (p=2). Here A= (a1, a2…a_n) and B= (b1, b2…b_n) are the query vector and test object vector respectively.

VII. EXISTING SYSTEM

Ling Shao, Simon Jones and Xuelong Li[5] gave an outline for efficiently searching and localizing human actions in large video database. Their system performs as:

- Firstly, they pre-processed the videos and stored them in the database.
- In pre-processing step, they extracted features from the video and represented in the form of a tuple *x=(x,y,t,c)* where *x,y* represents spatial location, *t* represents temporal location and *c* is the codeword of the feature.
- From the query image, the features are extracted and represented in the codeword.
- To perform efficient search, they first performed temporal localization to identify candidate region and then preformed spatial localization on the candidate region. They did this because it is observed that a human action occupy relatively large portion in spatial domain than in temporal domain.

© 2015, IERJ All Rights Reserved

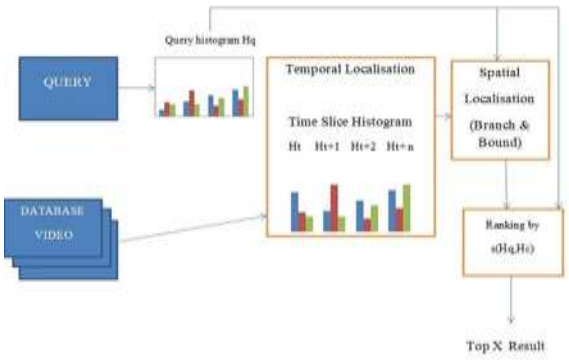


Fig 1. Existing System Architecture

In temporal localization, they compared the bag-of- words histogram Ht of the pre-processed with the bag- of-words histogram Hq of the query. For comparison they used following normalized histogram intersection Then they applied spatial localization [5] (SL) on the identified set of candidates. For this they used two approaches:

$$s(H_t, H_q) = \sum_{i=1}^{k} \frac{\min\ (H_q^i, H_t^i)}{H_n^i}$$

X-Y Separated SL: In this approach they linearly separated localization along X and Y spatial dimension to minimize computation complexity.

Branch-and-Bound Simultaneous SL: This approach is at the cost of higher computational complexity; localization is possible to perform simultaneously on both the dimensions and can achieve greater accuracy.

Once the results are retrieved, they are ranked on the basis of score assigned to them. The top *X* results are displayed to the user. If the user is not satisfied with the results displayed, then the user can enhance their result through relevance feedback.

VIII. PROPOSED SYSTEM

In our proposed system we are extracting multiple features of the videos and query image which provide more relevant result while comparing the query with the features of video stored in database.

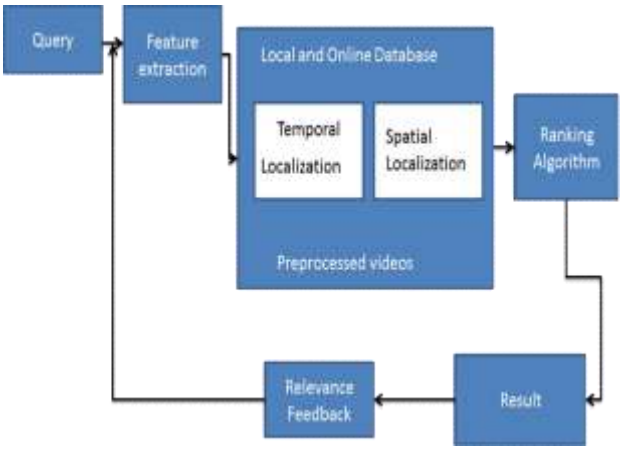


Fig 2. System Architecture

Histograms are created from the features which are extracted from the query, this query histogram (Hq) are then localized by temporal and spatial localization, after receiving the result we have applied Ranking Algorithm in which they are ranked on the basis of score assigned to them. The top *X* results are displayed to the user. If the user is not satisfied with the results displayed, then the user can enhance their result through Relevance Feedback.

IX. RESULT

Fig 3. Home Page

Fig 4: Relevant Results

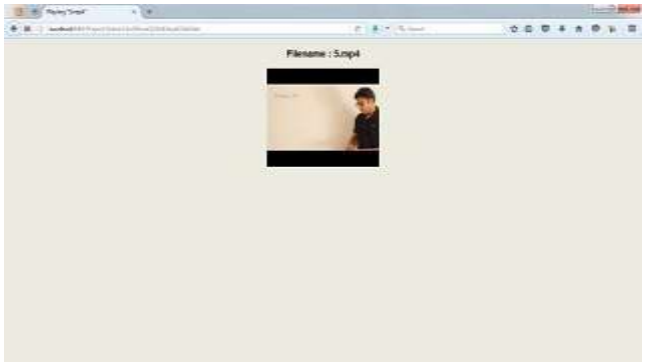


Fig 5. Video Playback

X. CONCLUSION AND FUTURE WORK

The CBVR technique and related algorithms are fast and can work well with online video database. The system uses localization which is based on histogram. The system takes the consideration of different file formats and large video files. By using the efficiency-first approach this has resulted in the creation of a fast permissive spatio-temporal

© 2015, IERJ All Rights Reserved

technique for localization of objects such as images followed by a more orthodox histogram ranking step, both of which can be assisted by ranking algorithm. The general principles of the system, such as batch pre-processing, spatio-temporal feature binning, and dimensionally sequential localization can be combined with a wide variety of existing human action recognition/localization techniques, and that concentrated efforts in investigating these technique may yield further improved performance.

REFERENCES

[1] M. Ryoo and J. Aggarwal "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in Proc. IEEE Int. Conf. Comput. Vision, 2009, pp. 1593–1600.

[2] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Efficient Subwindow search: A branch and bound framework for object localization," IEEE Trans. Pattern Anal. Mach. Intell, vol. 31, no. 12, pp. 2129–2142, Dec. 2009.

[3] J. Yuan, Z. Liu and Y. Wu, "Discriminative video pattern search for efficient action detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 9, pp. 1728-1743, Sep. 2011.

[4] S. Jones and L. Shao, "Rapid localization and retrieval of human actions with relevance feedback," in Proc. Int. Conf. Comput. Analysis Images Patterns, 2013, pp. 20–27.

[5] L. Shao, S. Jones, and Xuelong Li, "Efficient Search and Localization of Human Actions in Video Databases", IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, NO.3, March 2014

[6] Ivan Laptev, "Improving object detection with boosted histograms" June 2008

[7] Simon Jones, Ling Shao, "Content-based retrieval of human actions from realistic video databases" Feb 2013

[8] Simon Jones , Ling Shao , Jianguo Zhang , Yan Liu , "Relevance feedback for real-world human action retrieval", May 2011

[9] B V Patel and B B Meshram, "CONTENT BASED VIDEO RETRIEVAL SYSTEMS", International Journal of UbiComp (IJU),Vol.3, No.2, April 2012

[10] Xinmei Tian, Dacheng Tao, "Active Reranking for Web Image Search", IEEE Trans. On Image Processing, VOL. 19, NO. 3, March 2010

[11] Sr an Zagorac, Ainhoa Llorente, Suzanne Little, Haiming Liu, Stefan Ruger "Automated Content Based Video Retrieval"

[12] Stephan Marhand, Maillet Viperteam" Content based Video Retrieval"

© 2015, IERJ All Rights Reserved