Content-Based video searching and Retrieval Systems- Traditional and Recent approaches

Waleed Y Khawagi
KING ABDULAZIZ UNIVERSITY
Faculty of Computing and Information Technology

Abstract— Video Contents Analysis (VCA) is a new research field that recently emerged .VCA is the processes that analyzes video to extract The desired knowledge and information .Contents may refer to motion ,color, background and objects such as a human face or a car, to mention a few. This thesis investigates the algorithms which were designed for video content analysis such as background, text, and speech and face detection. Those algorithms are on a process of continuous improvement. Algorithms related to background, text, speech and face detection has been analyzed. The most robust one has been identified. Convolutional Neural Network (CNN) which is one of the state-of-art technologies has been implemented in solving the face detection problems. A theoretical model has been designed for the process of face detection and an algorithm slightly modified from recent algorithm has been proposed.

Index Terms— search engines ,optimization, data,structures, Indexing, video retrieval.

1 Introduction

The unprecedented and rapid progress in data capturing, storages and communications technologies have resulted in availability of high volume of Video data [1]. Cisco expects that by year 2019 the video traffic will constitute 80% of internet traffic [2]. Surveillance cameras around the world daily record a huge amount of videos for different purposes. In brief, almost all fields contain huge amount of videos. It will be completely an inadequate trying to locate specific information or a feature in a video sequentially because it is time consuming and cost a lot. For example, the Federal Highway administration of U.S.A has led the Naturalistic Drive Study (NDS) research project to understand the many factors interactions involved in highway crashes. The result of the project was over 2 petabytes (2000 terabyte) of video data. To analyze such data using traditional way, it is found that it will take 600 technicians working a full year and 40 hours per week which is considered completely inadequate [3].

Video Content Analysis has emerged as a new field in computer vision and artificial intelligence. In general a traditional video content analysis and retrieval system consist of four primary processes namely feature extraction, structure analysis, abstraction (summarization) and indexing. Each of these processes involves challenges that should be handled. As a recent progress in machine and deep learning new approaches has also emerged that depend on deep understanding of the video content rather than on feature extraction. This research will investigate traditional and recent approaches in video content analysis.

2 Background on Traditional Content Based Video retrieval systems

Traditional Content-Based Video retrieval methods act by first extract the low-level features of a video (Visual/Acoustics) and then search the video data for similar characteristic match of the features. The mission of the researches in content-based video retrieval is to develop technologies that could automatically parse video, text and audio. The backbone of a video retrieval system is to develop an efficient parsing system that could extract structure and information content of a video. It should have the ability to index and represent content attribute of any video [1].Figure 1 illustrate the architecture of a general video content retrieval system. Parsing is basically consisting of segmentation of the video into units and extracting features from them. In the following sections we will highlight all the steps taken by video content analysis system in order to carry out it is mission.

2-1 Video Segmentation

Video segmentation is essentially the process by which a digital image is portioned into multiple segments or regions [4]. Segmentation could be performed based on scene change or shot detection [4].a shot is defined as an image sequence that present continuous actions [5].Key frames are extracted from video shots. Key frames are still images, extracted from original video data that best represent the content of shots in an abstract manner [1]. Key frames extraction play an important role in video retrieval systems. Many of the retrieval systems

used the first key in a shot as the Key-frame. Unless the video is static this method will not be efficient. Many algorithms have been developed for efficient key frame extractions.

Researchers in [15] have proposed an algorithm for key frame extraction. Their algorithm captures the first frame and compares it with the second one based on criteria such as similarity, color feature and motion. This process continues until the algorithm could capture the best key frame. Other algorithms also have been designed for the purpose of key frames extractions.

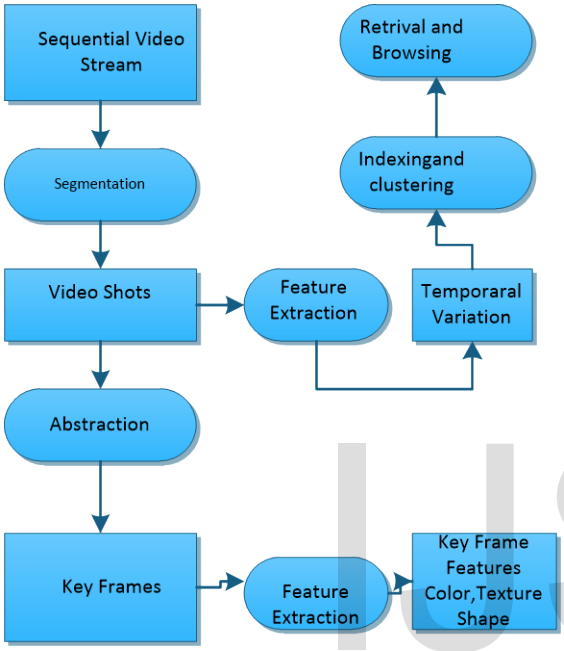

Figure 1: General Architecture of a video Content based retrieval system. Source [15]

2-2 Video Abstraction (Summarization)

Many videos may have Metadata and tags that are irrelevant to video content, accordingly this may be time consuming in the way that users will retrieve the video but find out that it has nothing to do with the metadata and tags given. One way to solve such a problem is through video summarization [6] .Video summarization aim at reducing the number of information that is to be search in order to retrieve the desired one. Video summarization generates a compact version of the original one. Video summarizations methods strive to achieve two important results: to include enough details from the original video that make it comprehensive and not to include redundant details [6].Approaches used to summarize videos depend on either low-level visual features or high-level visual features extraction[6].One of the challenges in Video summaarization is the definition of what is "Important" or "Essential" parts of the original video to iclude in the summarization[6].InGeneral, different type of approches take the types of the video into consideration ,for example in video related to sports the important parts will depend on the rules that govern that sport [6].

2-3 Features Extraction

Low features

Next step after extracting key frames is to extract law and high features of the video. Low level features such as object motion, colors, and shapes are extracted and registered into a database. The data base could answers queries such as "finding images with more than 30% blue and green colors" such a query could retrieve images of sky and grasses.

High level features or Semantics Features

High level features characterized each frame individually. Examples of high features are brightness, histogram and amount of certain objects such as cars or faces [2].High level features are supposed to answers queries such as: "find picture that contain sky". Extraction of high features is a complicated and hard task and more processing power and storage is needed [3].

2-4Text in video: Detection and Extraction

Text in video is the most reliable methods for users to locate their desired video content. If we could successfully extract text in a video it would help in video indexing, summarization and retrieval. Text detection and extraction is an important component in video search and retrieve. Video text extraction is a key point that helps in searching and indexing video. Videotext detection and recognition can be used in many applications, e.g. semantic video indexing, summarization, video surveillance and security, multilingual video information and education. Videotext could be categorized into two broad types: Graphic text and scene text. Graphic texts (Superimposed) which are added by the video creator or editor such as in educational video. Scene text which is the video text exists in the object and scene of the video such as street name or car plate [7]. The process of videotext extraction is consists of three main steps ,text detection ,text separation from background and in the last step production of binary image that contains the text and background .

2-5 Background subtraction

Detection of objects such as human and cars in a video scene require first to separate the moving objects called "foreground", for further processing, from the static information called "background". This method is used in many image processing and computer vision applications. Background subtraction is widely used for detecting moving objects in surveillance and static cameras [8].

There are different background representation model which could be classified into: basic models, statistical models, cluster models, neural network models and estimation models [9] .Many algorithms have been developed for background subtraction. The researchers in [9] have compared most of

background subtraction algorithm based on certain criteria such as CPU and memory requirement. They concluded that sophisticated algorithms and methods (CB, GMM and KDE) are not always the best one. Secondly, sophisticated algorithms need high computation capabilities and large memory, so they are not suitable for real-time applications such as surveillance cameras. Finally, there is no a win algorithm, but they depend on the environment on which they are being operated. An algorithm can perform better in a certain situation and bad on other one.

2-6 Face Recognition

Face recognition is consisting of two tasks: face verification and face identification. Face verification refers to the process of determining whether two faces belong to the same person while face identification is the process of identifying a face from a set of faces [10]. One of pioneer work in face detection have been carried out by Viola-Jones [11]. Viola-Jones algorithm is considered robust with very high detection rate and it works in real time but it only used for face detection. It consist of four stages Haar Feature which are digital image features used for object detection selection, creating an integral image, Adaboost training and cascading classifier [11].Another approach that is widely being used recently is HOG (Histogram of Oriented Gradients) as a result of work of the researchers Navneet Dalal and Bill Trigs [14].Practically this is carried out by dividing the image windows into spatial regions called cells. For each cell, the gradient intensity is measured for each pixel within the cell through 1-D histogram. When all histogram are combined this will form a representation of the shape. For better accuracy, the histogram could be normalized by calculating gradient intensity within numbers of cell called block and then normalize all cells within the block. This normalization will downplay the effects of shadowing and illumination. Each normalized descriptor blocks are called Histogram of Oriented Gradient (HOG). Tiling the detection window with a dense grid of HOG descriptors and using the combined feature vector in a conventional SVM based window classifier gives our human detection chain.

3-Query of Video database

When a user issued a query to search for a video, the query will be handled by comparing the features vector stored in the database and the query features. Whatever type of query is used weather it is a text, object, a face or still image, similarity will be computed. For example, similarity of two images could be determined by measuring Euclidean distance. A video clip is retrieved by finding key frames occurring sequentially in the video database which are similar to that of the query video. There are many ways to query a video database.

3-1Query by object

In this method a sketch or an image is provides and If the object provided exist in the database it will successfully located.

3-2 Query by Text

The most popular method used for quarrying a video's database. The word or words used for queering the database is compared to all keywords and tags related to videos. Similar one will be displayed.

3-3Query by shot

Some systems allow query by the full shot instead of key frames. It could give better results but at high cost of computation.

3-4 Query by clip

A clip can be used for better performance of video retrieval as compared to the technique when a shot is used because a shot do not represents sufficient information about the whole context. All the clips which possess a significant similarity or relevancy with the query clip are retrieved [12].

3-5Query by Faces and Text

It is also possible to query by face and text. The key frame of the frame or clip used for query is first extracted. An algorithm is used to search and locate the required clip using the information obtained from the key frame.

3-6 Query by Example

Query by example will perform better if visual features of the query are used for content based video retrieval [13]. Low level features are obtained from key frames [9] of the query video and then they are compared to distinguish the similar videos using their key frames visual features [13].

4-Recent content-based Video Semantic Retrieval (CBVSR)

Due to progress and advance in object, faces, text and action detection, researchers started to try searching using complex queries called events. An event usually involves people engaged in a certain activity at specific place and time [14]. For example the event "Birth day" may include visual things such as "cake" ,"candles " and "kids" and audio concept such as "birthday songs". A concept can be regarded as a visual or acoustic semantic tag on people, objects, scenes and actions in the video content [14]. CBVSR is emerged as an advance in machine learning and neural networks. Machine learning aims to develop the computer algorithms which can learn experience from example inputs and make data-driven predictions on unknown test data. Deep learning aims to extract hierarchical representations from large-scale Data (e.g. images and videos) by using deep architecture models such as neural networks with multiple layers of non-linear transformations. Many new algorithms and methods has been developed using deep learning techniques. There are algorithms developed for object detection, image classification, face detection and object tracking. [10] has discussed many of these algorithms.
The challenge faced by deep learning decades ago was the

unavailability of big data so as to train classifiers. Now a day's huge libraries of free video data are available for researchers to use them. The techniques and algorithms used deep learning have a achieved promising results in video-based content analysis. The search in this method is solely based on content understanding rather than low-level features [14]. Jiang et.el [14] has proposed a new content base search and retrieval based on semantic queries for even detection. They proposed a new theory, which is inspired by cognitive process in human and animals, called self-paced Curriculum Learning (SPCL) to train robust content detectors.

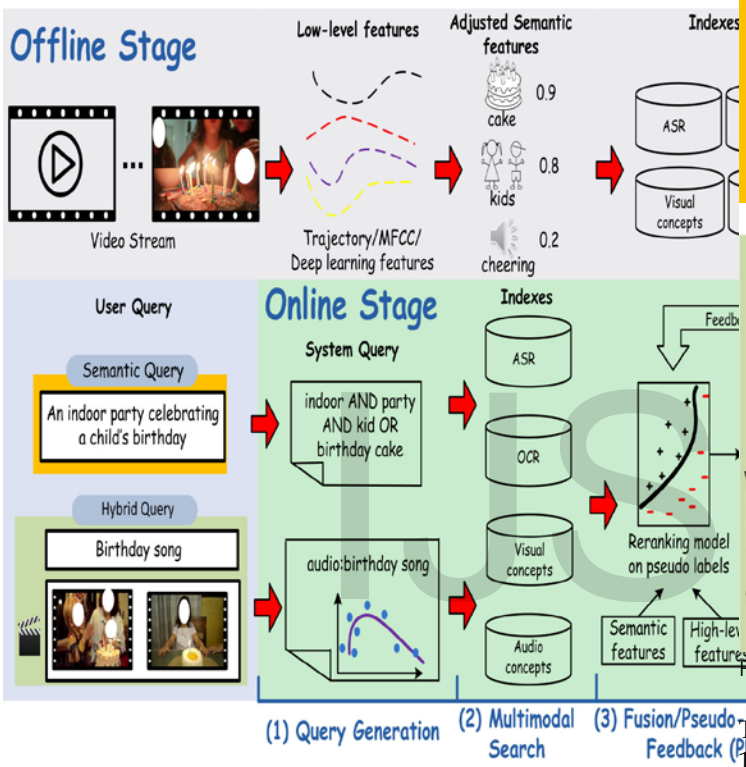

Figure 2 : Framework for the proposed system by [2].

The researcher has divided the video retrieval system into two stages as illustrated in the upper figure. The off-line stage aims at extracting semantic features. It usually involves these steps are: Video clip will be represented by low-level features that capture features such as texture or acoustic. The researchers included the following features dense trajectories, Convolutional Neural Networks (CNN) feature and CNN feature for audio [14]. The low level features are then input to ready-made detectors to yield the semantic features. The semantic features are human intercept able features. The researchers considered the visual/audio concepts, Automatic Speech Recognition (ASR) and Optical Character Recognition (OCR) as the semantic features. After features are extracted they will be indexed for efficient on-line search. The second stage is called video search which implement processing user's queries. A user might process queries in Varity of ways which includes text descriptors or video examples [2].Semantic queries are executed by mapping the user's vocabulary concepts to the most relevant one in the system database. For hybrid queries, the example model will be trained.

Figure 3: Semantic and Hybrid queries in Source [2]

The experiments showed that the proposed system can search 1 million videos with 1core in less than 1 second while retaining 80% of accuracy of a state-of-the-art system.

CONCLUSION

In this research we have reviewed the traditional and recent content-based video retrieval system. We showed that these systems has been evolving and progressing. The latest approach being used is based on machine learning and Convolutional Neural Network. These new method have provided promising results that encourage researchers to pursue their researches based on. The draw backs of the traditional methods is their limited ability in processing high volume of data, While the recent methods have the ability to process millions of videos in a very short time.

ACKNOWLEDGMENT

To those who contributed to my arrival at the end road, to everyone who taught me something new. And fed my intellect with knowledge and knowledge, to all who stood beside me

and helped me in all the difficulties and obstacles. To my Father, who gave me strength, faith, love and tenderness, healing balm

To my mother, the most beautiful melody I am playing on the strings of my return, the sun shining in the night and the day, the most beautiful pulse beats the heart. To my wife, mate Darby, sun in the sky of my life, O light has covered my grief how wonderful your smile that gave me meaning for life.

To the first point in the line to who gave me the strength and opportunity and stood with me in each line to those who taught me success Dr. Ahmed Saeed Al Zahrani. To those who stood with me in every letter to those who supported me and lit my way Dr. Reza Mohammed Salama, Maher Ali Makhmakh, Dr. Anas Mohammed Fattouh

To Sindhi in the life of my beloved brothers, to the dream of life and the ambition of tomorrow and the daily brilliance of Laura and Lara

REFERENCES

[1] N. Dimitrova, et.el, "Applications of video content analysis and retrieval". IEEE Feature article2002

[2] Jiang, Lu Web-scale Multimedia Search for Internet Video Content, Carnegie Mellon University, PhD Thesis 2015

[3] Exploratory Advance Research Program Video Analytic Research Projects us department of Transportation

[4] Muthukrishnan.R et al Edge Detection Techniques for Image Segmentation, International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 6, Dec 2011

[5] Dueminis, Gregory et al Video index and search services based on content identification features IEEE 2005

[6] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, Naokazu Yokoya "Video Summarization using Deep Semantic Features" the 13th Asian Conference on Computer Vision (ACCV'16) 2016

[7] X. Liu and J. Samarabandu, "Multiscale edge-based text extraction from complex images," in Proc. Int. Conf. Multimedia and Expo(ICME), Jul. 2006

[8] https://en.wikipedia.org/wiki/Background_subtraction

[9] Yannick Benezeth, Pierre-Marc Jodoin, Bruno Emile, H_el_ene Laurent, Christophe Rosenberger. Comparative study of background subtraction algorithms, Journal of Electronic Imaging, Society of Photo-optical Instrumentation Engineers, 2010, 19

[10] Li Wang and Dennis Deep Learning Algorithms with Applications to Video Analytics for A Smart City: A Survey-IEEE December 2015

[11] P. Viola and M. J. Jones Robust real-time face detection, International Journal of Computer Vision, 57 (2004), pp. 137–154 http://dx.doi.org/10.1023/B:VISI.0000013087.49260.fb

[12] A. Anjulan and N.Canagarajah, "Unified framework for object retrieval and mining," IEEE Trans. Circuits Syst. Video Technol., vol. 19, no. 1, pp. 63–76, Jan. 2009.

[13] Weiming Hu, Nianhua Xie, Li, Xianglin Zeng, Maybank S., "A Survey on Visual Content-Based Video Indexing and Retrieval", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 41-6,797-819, 11/2011

[14] L, Jiange et.el "Fast and accurate content-based semantic search in 100m internet videos.", 2015.

[15] Hong Jiang Zhang, Jianhua Wu, Di Zhong, Stephen W. Smoliar, "An integrated system for content-based video retrieval and browsing", Pattern Recognition, Pattern Recognition Society, Published by Elsevier Science Ltd., Vol. 30, No. 4, pp. 643~658, 1997