A Survey on Video Annotations Different Techniques

Archana.V.Potnurwar
Research Scholar Computer Science &Engg Department
P.I.E.T Nagpur

Mohammad Atique, Ph.D
Associate Professor P.G.Department Of Computer Science[2]S.G.B.A.U,Amravati

ABSTRACT

The insufficiency of labeled training data for representing the distribution of the entire dataset is a major obstacle in automatic semantic annotation of large-scale video database. Semi-supervised learning algorithms, which attempt to learn from both labeled and unlabeled data, are promising to solve this problem. In this paper **,r**etrieving videos using key words requires obtaining the semantic features of the videos. Most work reported in the literature focuses on annotating a video shot with a fixed number of key words, no matter how much information is contained in the video shot.

Keywords

Videoannotations, multimodal, automatic video annotations, domain specific.

1. INTRODUCTION

Image annotation is an active field of research that serves as a precursor to video annotation in numerous ways. Video features are often inspired and sometimes directly borrowed from image techniques and many methods for image indexing are also easily applied to video. Here we survey some of the most relevant static image annotation literature including modern trends in the field and adaptations of techniques for static image annotation to video. In the following literature the covered topics include emerging and state of the art feature extraction techniques specifically designed for video. We review image features, indexing techniques, and scalable designs that are particularly useful for working with web-scale video collections

The annotation is the basis for the detection of video's semantic concepts and the construction of semantic indices for videos.
The following are the approaches for video annotation

a) Statistic-based approach

b) Rule or knowledge-based approach

c) Machine learning-based approach

Video annotation is very important for video management, such as video retrieval. Despite continuous efforts in inventing new annotation algorithms, the annotation performance is usually unsatisfactory, and the annotation vocabulary is still limited due to the use of a small scale training set. The effectiveness of proposed method is analyzed by valuating the precision-recall of test videos.

Most of the current existing video annotation systems are video scenario based. Notes can be added to the time segments on a video timeline. A user can also view the video clip, mark a time segment, playback the segment, or attach his/her written notes to the segment. All of the annotation information is in the video level and will be mixed together, which makes it very difficult on semantic video retrieval. That is, the users cannot effectively and easily get what they want. Resolving this problem is our main objective. In addition, a semantic video annotation tool at least should support the following functionality:

- Divide a video into a number of scenes
- Divide a scene into a number of frames;
- Develop a unified schema for video annotation
- Annotate a scene and a frame solely

The annotation set is not limited to words that have training data or for which models have been created. It is limited only by the words in the collective annotation vocabulary of all the database documents. Different types of modality issues to be considered while performing annotations i.e Textual Modality , Visual Modality, Auditory Modality. Learning-based video annotation is a promising approach to facilitating video retrieval and it can avoid the intensive labor costs of pure manual annotation. But it frequently encounters several difficulties, such as insufficiency of training data and the curse of dimensionality

1.1 Challenges found in video annotation

- Training data insufficiency
- Curse of Dimensionality
- Choice of distance Function
- Neglect of temporal consistency

2. BACKGROUND

There are three types of image annotation approaches available: manual, automatic and semi automatic. Following table describes difference we can go for semi automatic as per the review.

table describes difference we can go for semi automatic as per the review.

. Table 1 Annotation Techniques

Annotation techniques	Manual	Semi Automatic	Automatic
Initial Human Interaction	Enter some descriptive keyword	Provide initial query at the beginning	No interaction
Machine task	Provide storage for annotation to be	Parse Human's query and extract	Detect labels semantic keywords automatically

	saved such as disk space or database	semantic information to perform annotation	using recognition technology

Table 2 Advantages and Disadvantages

Annotation techniques	Manual	Semi Automatic	Automatic
Advantages	The most accurate annotation	Quality of the annotation improves in the interactive manner after correction	The most efficient the least time consuming
Disadvantages	Time consuming , expensive, difficult, subjective, inconsistency	Less time than automatic greater time than manual annotation	Error prone,less Accurate annotation

In annotation we can go for multimodality

3. MULTIMODAL

Utilizing the available multimodality in video mediums, such as audio and sometimes enclosed text, has received relatively a good attention [2], In spite of that the multimodal features analysis usually increases certainty of video annotation, In this it was preferred to analyze input video's visual features only to keep focusing on wide domain. This was also to accommodate some domains where video clips lacks audio and enclosed text, or they are not so correlated with the visual features such as wild hunts and surveillance [3].

3.1 Multimodality

Different types of modality i.e Textual Modality , Visual Modality, Auditory Modality .Further also discussed Content based video indexing compromises of High-level indexing: Index on the basis of high level features e.g. action, time, and space. The main advantages of high-level indexing are that it can give more accurate semantic correct result. In high-level indexing, the high-level and low-level features are map to reduce the semantic gap. Low-level indexing: Index on the basis of low-level features e.g. colour, shape, and texture. Here no semantics is attached. Video can be retrieved by simple pattern matching and similarity measuring techniques. The main advantages of low-level indexing are that it is automatic and fast as compared to high-level indexing.

And Domain specific indexing: These technique uses high-level structure of video to constraints the low-level features extraction and processing. Also the Indexing Techniques of three types Segment-Based Video Indexing, Object-Based Video Indexing, Event-Based Video Indexing are discussed.

And focused on some issues that need to be considered.

a. Need for generalized multimodal video indexing techniques

b. Multimedia data (video) does not have a single unique semantic, so how do we highlight the semantic that will be further used for content based multimedia indexing.

c. The main challenge or complexity in video indexing and searching is that video data is multimodal. There is a need of a system that can decide that which modality is combined or used in order for maximum effectiveness and accurate searching.

d. Need of the framework for indexing that select the most appropriate mode for indexing or using the different modality combination[10].

4 .Semantic Video Annotation System

a prototype of a video annotation system, called Semantic Video Annotation System (SVAS) in which a three-level annotation architecture and a semantic video search language called Semantic Query Description Language for Video (SQDL-V) is used. SQDL-V engine based on SVAS is able to return more accurate search results in comparison to the formal video search method.[4].

4.1 Video semantic annotation using graph diffusion technique a novel and efficient approach for scalable to large data sets where only a couple of minutes improving large scale video semantic annotation using graph diffusion technique. The main concentration in this paper was on. Firstly, it allowed the online update of semantic context for addressing the problem of domain shift .Second , it was required to complete approach implemented over hundreds of concepts for thousands of video shots**[6].**

4.2Automatic video annotation method which determines the region of the foreground object and predicts its class.
The former was done using consensus foreground object template (CFOT) for moving object detection, and the later was achieved by the integration of heterogeneous features from different domains. In this work, the focus is on the challenging task of Web video annotation, in which most existing Web videos were captured under uncontrolled environments, with insufficient quality or limited tag information available[5].

The System has collected a complex, uncontrolled, and challenging Web video dataset from YouTube for the experiments carried out. The video data were captured by moving or shaky cameras and the moving object of interest were present in cluttered background. Significant scale and viewpoint variations of the objects were observed, and the resolution of a large portion of videos in dataset was low. The system considered six different moving object categories: Airplane, Ambulance, Car, Fire Engine, Helicopter, and Motorbike. Each object category had 25 to 30 video sequences, and each sequence has one moving foreground object present in it. Randomly select 10 from each class for training, and the remaining for testing

5 Different Learning Approaches To Video Annotation

i)Semi supervised

ii)Supervised

iii)Active learning

For supervised methods, the models of the semantic concepts are built over a labeled training set, and then the labels of new samples can be predicted by the learned models. Semi-supervised learning and active learning are two approaches to dealing with the difficulty of training data insufficiency in supervised methods. Semi-supervised learning methods exploit unlabeled data with certain assumptions, and they are expected to build more accurate models than those that can be achieved by supervised methods. Different from supervised and semi-supervised methods, active learning aims to organize a more effective training set. It works in an iterative way. In each round, the most informative unlabeled samples are selected for manual annotation, such that the obtained training set is more effective than that gathered randomly

5.1 Semisupervised Learning

In semi-supervised learning algorithms, self-training and co-training, can be enhanced by exploring the temporalconsistency of semantic concepts in video sequences. In the enhanced algorithms, instead of individual shots,time-constraint shot clusters are taken as the basic sample units, in which most mis-classifications can be corrected before they are applied for re-training, thus more accurate statistical models can be obtained.

5.1.1 Self-Training

For self-training, firstly a classifier is trained from a small amount of labeled samples, which is then used to classify unlabeled samples. Typically the classified samples with high confidence levels are added to the training set.

5.1.2 Co-Training

For co-training, it is assumed that the features can be split into two sets that are conditionally independent given the class, and each feature set is sufficient for training a “good” classifier. Initially two separate classifiers are trained based on these two feature sets with a set of labeled samples respectively. Each
classifier then classifies unlabeled samples, and adds those with high confidence levels to the training set, which is applied to “teach” the other classifier. Afterwards two classifiers are re-trained from the new training set based on the corresponding feature sets, and the process repeats.

6. Text Extraction Using Clustering Algorithm:

One of the issue in video annotation is to extract text from frames. Using the measured similarities between frames, shot boundaries can be detected. Shot boundary detection approaches can be classified into threshold-based and statistical learning-based. Approaches to extract key frames are classified into six categories : sequential comparison-based, global comparison-based, reference frame-based, clustering based, curve simplification-based, and object/event based [7] . Preference can be given to clustering based, These algorithms cluster frames and then choose frames closest to the cluster centers as the key frames. Selection of key frames can be done using the complete link method of hierarchical agglomerative clustering in the color feature space. Another method to extract key frames using the fuzzy K-means clustering in the color feature subspace. The merits of the clustering-based algorithms are that they can use generic clustering algorithms, and the global characteristics of a video can be reflected in the extracted key frames. The limitations of these algorithms are as follows: First, they are dependent on the clustering results, but successful acquisition of semantic meaningful clusters is very difficult, especially for large data, and second the sequential nature of the video cannot be naturally utilized: Usually, clumsy tricks are used to ensure that adjacent frames are likely to be assigned to the same cluster. For above text extraction in video following properties of text can be considered:

1) Dense intensity variety (or gradient);

2) Contrast between text and its background;

3) Structural information;

4) Texture property

Key frame used for representing the main content of a video shot. Key Frame Extraction is the key technology for video retrieval,video query, video index, video browse and video abstraction. The algorithm for key frame extraction will affect the establishment and retrieval efficiency of video retrieval system.Key Frame Extraction-based video retrieval generally includes such steps as follows.Firstly, a video is divided into different shots, and keyframes are extracted from these shots. Then the low-levelvisual features such as color, texture and shape are extracted from the key frames. These features are being used to build index and will be kept in database. After that,users can search videos from database by different search mechanisms.Current key frame extraction techniques mainly include:shot-based method, content analysis-based method, motion analysis-based method and clustering-based method.

(1)In shot-based method, video is divided in to different shots, the first or the last frame in a shot is regarded as the key frame. Although this method is simple, it is only appropriate for static video.

(2)In motion analysis-based method, key frames were extracted on the basis of object motion or camera motion in video.This work is supported by Natural Science Foundation for Young Scientists of Shanxi Province (2010021016-1).

the acceleration of moving objects in video. It firstly detected the motion type of camera and then extracted key frame based on the order of motion types[4].This disadvantage of this method is due to its very large calculation.

(3)In clustering-based method, similar frames are clustered to the same category. Clustering can be used in scene or in single shot. When it is used in scene, it can distinguish different shots and then key frames are extracted from the shots. When it is used in shots,sub-shots are generated This method has been proved to be effective. Nevertheless, it breaks the temporal sequence of the key frames in original video.

(4)Content analysis -based key frames selection depends

on the changes of video content. The key problem in content analysis-based method is whether it can well capture the underlying key frames when there are lots of changes in contents

7 USES OF VIDEO ANNOTATIONS

1. Broadcasters generally annotate material that will be used later for either immediate "highlights" purposes, or for archiving

2. "Production Logging" in which producers will mark up an event live, to note shots to be edited into highlights packages and "Posterity Logging" in which librarians make detailed annotation of video tape for long term reuse, where depth and historical context is also noted.

3.Faster retrival process.

7.1 OBJECTIVE

1. The enhanced annotations resulting can be used directly in improving existing text-based search engines.

2. Automated video annotation must explicitly address the issue of scalability, both in terms of the quantity of video and the expansiveness of the annotation vocabulary.

3.Research in video search and mining techniques is progressing rapidly yet most works are limited by small vocabularies and dataset sizes we can develop a prototype system to enhance web scale video search with automated Video annotation .Testing the model on a portion of YouTube can demonstrates the scalability and efficacy of our approach that will be used.

7.2 GAPS

It is well-known that analyzing and reasoning about video data are not easy due to

(1) the difficulty of approaching and simulating human being's perception by computers, and

(2) the lack of semantically meaningful annotations and technologies in understanding complex audio/visual data This is often referred to as the "semantic gap" in the multimedia retrieval community which limits the retrieval effectiveness.

8. CONCLUSION

In this paper, we have recalled some problems related with different techniques. retrieval. The state of the art of existing approaches in each major issue has been described with the focus on the following tasks: video structure analysis including shot boundary detection, key frame extraction and scene segmentation, extraction of features of static key frames, objects and motions, video data mining, video classification and annotation, video search including interface, similarity measure and relevance feedback, and video summarization and browsing.In this paper uses,gaps and objectives of video annotation is been given.

9. REFERENCES

[1] Amjad Altadmri and Amr Ahmed "Automatic Semantic Video Annotation in Wide Domain Videos Based on Similarity and Commonsense Knowledgebase's ,"in IEEE International Conference on Signal and Image Processing Applications ,2009.

[2] M. L. Shyu, Z. Xie, M. Chen, and S. C. Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," IEEE Transactions on Multimedia, vol. 10, no. 2, pp. 252–259, 2008.

[3] N. Haering, R. J. Qian, and M. I. Sezan, "A semantic event-detection approach and its application to detecting hunts in wildlife video," IEEE Transactions on Circuits and Systems for Video Technology, vol. 10, no. 6, pp. 857–868, 2000

[4] ChengZhi Xu, HuiChuan Wu1, Bo Xiao1, Philip C-Y Sheu1, Shu-Ching Chen," A Hierarchical Video Annotation System".

[5] . Shih-Wei Suna, Yu-Chiang Frank Wanga, Yao-Ling Hung "AUTOMATIC ANNOTATION OF WEB VIDEOS" a Institute of Information Science and Research Center for Information Technology InnovationAcademia Sinica, Taipei, Taiwan

[6] Yu-Gang Jiang,Jun Wang, Shih-Fu Changand Chong-Wah Ngo ,"Domain Adaptive Semantic Diffusion for Large ScaleContext-Based Video Annotation",by ieee ICCV sept29-oct 2,2009

[7] Vincenzo Lombardo & Rossana Damiano "An intelligent tool for narrative-based video annotation and editing", in 2010 International Conference on Complex, Intelligent and Software Intensive Systems

[8] Meng Wang, Xian-Sheng Hua Jinhui Tang and Richang Hong "Beyond Distance Measurement: Constructing Neighborhood Similarity for Video Annotation" IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 11, NO. 3, APRIL 2009 465.

[9] Jenny Yuen, Bryan Russell& Ce Liu1 Antonio Torralba, " LabelMe video: Building a Video Database with Human Annotations " in International Journal of Research and Reviews in Information Sciences

[10] Nida Aslam, Irfanullah, Kok-Keong Loo, Roohullah,"Growing Trend from Uni-to-Multimodal Video Indexing ", in nternational Journal of Digital Content Technology and its Applications Volume 3, Number 2, June 2009

[11] Khasfariyati Razikin, Dion Hoe-Lian Goh, Ee-Peng Lim, Aixin Sun, Yin-Leng Theng, Thi Nhu QuynhKim, Kalyani Chatterjea, Chew-Hung Chang, "Managing Media Rich Geo-spatial Annotations for a map-based Mobile Application using Clustering"in 2010 IEEE

[12] Jinhui Tang, Xian-Sheng Hua, Guo-Jun Qi , Zhiwei Gu , Xiuqing Wu ,"Beyond Accuracy: Typically Ranking For Video Annotation " ICME 2007 IEEE

[13] Emily Moxley, Tao Mei, and B. S. Manjunath Video Annotation Through Search and Graph Reinforcement Mining,", ieee transactions on multimedia, vol. 12, no. 3, april 2010

[14] Matthew Butler, Tim Zapart, and Raymond Li," Video Annotation – Improving Assessment of Transient Educational Events", in Proceedings of the 2006 Informing Science and IT Education Joint Conference Salford, UK – June 25-28