SURVEY ON VIDEO ANNOTATION

M. SUMITHRA[1], V. MERCY RAJASELVI[2]

[1] Department of Computer Science and Engineering
Easwari Engineering College
Chennai,India
[2] Department of Computer Science and Engineering
Easwari Engineering College
Chennai,India

Abstract Video annotation is an important techique for easy video retrieval and also for keeping away from the intensive cost of labour for pure manual annotation. But several difficulties has been found out, such as insufficiency in training the data and curse of dimensionality. Various approaches and algorithms which have been proposed for video annotation has been surveyed in this paper.

Keywords: Video retrieval, Annotation, Key feature extraction, Object-level recognition, Scene-level recognition.

I. Introduction

Image annotation acts as a precursor to video annotation in various ways which provides an active field of research. Video features are directly extracted from image processing techniques. Many methods for image indexing are also easily applied to video. Advancement in digital video technology in past few years, leads to explosion of digital video data. Moreover, increased availability of Internet bandwidth defines a new means of video distribution, other than physical media. The major search engines have already started to provide specific services for indexing, searching and retrieving videos on the Internet.

Improving video accessibility is the true challenge. In fact, accessing the video data requires that video content is indexed appropriately but annotating it manually or tagging video is the best laborious and it is economically infeasible process. Therefore, one important subject of research has been combined with the study of novel techniques to extract information about video content automatically. For different data access modalities such as browsing, searching, comparison and categorization annotation process is the first step.

For the detection of video semantic concepts and the construction of semantic indices for videos annotation is the basic step. The approaches for video annotation are as follows.

a) Statistic-based.
b) Rule or knowledge-based.
c) Machine learning-based.

Video annotation process is also important for video management, such as video retrieval. Despite the continuous efforts in inventing new annotation algorithms, the annotation performance is usually not satisfied, 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 precision-recall of test videos.

In addition, a semantic video annotation tool should atleast support the following functionality:

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

Different types of modality issues is being considered while performing annotations that is Textual Modality, Visual Modality, Auditory Modality. Learning-based video annotation is a promising approach for facilitating video retrieval and it can avoid the intensive labor costs of pure manual annotation. But it encounters several difficulties, such as insufficiency of training data and the curse of dimensionality.

The challenges found in video annotation are:

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

The section 2 deals about the key feature extraction, section 3 deals about the object-level recognition, section 4 deals about the scene-level recognition, section 5 deals about the video annotation, section 6 is about the conclusion.

II. KEY FEATURE EXTRACTION

The first step in extraction of key frames is shot change detection. It mainly refers to the detection of transition between successive shots. Each shot consist of several frames and can be represented by one or more frames based on relative temporal differences. The commonly used methods in shot transition detection are pixel-based comparison and histogram-based method. The pixel-based methods are highly very sensitive to motion of objects. A color histogram method is adopted to segment the shots according to frame difference. The Histogram-based method is the most frequently used method to calculate frame difference. Since color histograms does not relate to the spatial information with the pixels of given color, and only records the amount of color information, images with similar color histograms can have dramatically different appearances.

Christoph Feichtenhofer et. al (2014), proposed the feature extraction step, static appearance has been captured by spatial orientations, image dynamics has been captured by spatio temporal oriented energies and chromatic information has been captured by color statistics. Subsequently, primitive features are being encoded into the mid-level representation that has been learned for task of dynamic scene recognition. Finally, the novel dynamic spacetime pyramid is being introduced. This dynamic pooling approach can handle both global as well as local motion by adapting to the temporal structure, as guided by the pooling energies. The resulting system provides that the online recognition of the dynamic scenes that thoroughly evaluated on two current benchmark datasets and yields best results to date on both datasets. In depth analysis tells about the benefits of explicitly modeling feature complementarily in combination with dynamic spacetime pyramid, indicates that this is the unified approach and it should be well-suited in many areas of video analysis.

M.Ravinder et. al (2016), proposed a novel algorithm for content-based video indexing and video retrieval using texture of key frames, edge and motion features are being presented. Using this algorithm and k-means clustering based method key frames from video is extracted using, followed by extraction of texture, edge, and motion features to represent a video with feature vector. This algorithm has been evaluated on database of three hundred and thirty five videos (collected from TRECVID 2005, Google, and BBC) of the four types. The performance of proposed method is compared with the volume local binary patterns (VLBP) method. The performance of proposed algorithm is more efficient compared to the VLBP method.

III. OBJECT-LEVEL RECOGNITION

The type of the actions performed in a video can be easily recognized and detected by the humans. However, the automatic recognition of the human action is a contest in computer vision with growing applications for the automated surveillance, the content-based video retrieval, the video summarization, the elderly home monitoring for assisted living, and the human–computer interaction. The confusion lies in people who are performing the same action in noticeably different ways, leading to the errors of omission. Also, the individuals who are performing the different actions visually appear to be similar, which leads to the errors of commission. In addition, the illumination and the view/scale changes create further challenges to automatically interpret the scene.

Fatemeh Tabib Mahmoudi et. al (2015), proposed the object recognition strategy mainly in two steps: single view and multi views processes. In the single view process, for defining region's properties in each of segmented regions, the object-based image analysis (OBIA) is being performed independently on individual views. In second stage, the classified objects of all views are being combined together through a decision-level fusion based on scene contextual information for refining the classification results. Sensory information, analyzing the visibility maps, the height, and the structural characteristics of the multi views classified objects define the scene contextual information. For Evaluating the capability of proposed context aware object recognition methodology, two datasets are being performed: 1) multi angular Worldview-2 satellite images Rio de Janeiro in Brazil and 2) multi views digital modular camera (DMC) aerial images over complex urban area in Germany. The obtained results represents that using contextual information together with a decision-level fusion of multi views, the object recognition is difficult and an ambiguity are also decreased and the overall accuracy and the kappa are gradually improved for both of theWorldView-2 and the DMC datasets.

Amir H. Shabani et. al(2013), proposed the concept of learning multiple dictionaries of action primitives at the different resolutions and consequently, the multiple scale-specific representations for given video sample. Using a decoupled fusion of the multiple representations, the improvement in human classification accuracy of the realistic benchmark databases by about 5%, compared with the state-of-the art methods.

Jose M. Chaquet et.al(2013), proposed to wrap up the lack of complete description of most important public datasets for the video-based human activity and the action recognition and to guide the researchers in election of the most suitable dataset for the benchmarking their algorithms.

IV. SCENE RECOGNITION

Dynamic scenes are characterized by the collection of dynamic patterns and their spatial layout, as captured in the short video clips. For instance, a beach scene might be characterized by drifting the overhead clouds, mid-scene water waves and the foreground of static sandy texture. Other examples includes the forest fires, the avalanches and the traffic scenes. These scenes may be captured by the either stationary or moving cameras. Thus, while scene motion is characteristic, and it is not the exclusive of camera induced motion. Indeed, dynamic scene classification in the presence of the camera motion has been proven to be more challenging than when this confounding attribute is absent.

Aaron O. Thomas et.al (2016), proposed to explored influence of the video annotation conditions upon the meta comprehension accuracy and learning performance with the group of 81 undergraduate students of various majors. Findings suggest that the video annotation systems designed for the simultaneous note taking may have a deleterious effect upon the meta cognitive monitoring in general and meta comprehension in particular. Text-based strategies used to improve the meta comprehension accuracy such as delay in the production of the keyword to summarize the essence of an instructional topic that do not appear to the impact meta cognitive performance in the context of video annotation. Interestingly, participants in the control condition performed as well in the both learning performance and meta comprehension accuracy as their counterparts. These findings have been implications for the design of the video annotation systems and the learner best practices in use of video annotation, particularly in online and blended learning formats.

Chien-Li Chou et.al (2016), proposed the near-scenes, which contain similar concepts, the topics, or the semantic meanings, are designed for the better video content understanding and annotation. A novel framework of hierarchical video-to-near-scene (HV2NS) annotation not only preserve but also to purify the semantic meanings of the near-scenes. To detect near-scenes, a pattern-based prefix tree is the first constructed to fast retrieve near-duplicate videos. Then, the videos containing similar near-duplicate segments and the similar keywords are being clustered with the consideration of multi-modal features including visual and textual features. To enhance the precision of the near-scene detection, a pattern-to-intensity-mark (PIM) method is proposed to perform the precise frame-level near-duplicate segment alignment. For each the near-scene, a video-to-concept distribution model is being designed to analyze the representativeness of the keywords and discriminations of the clusters by the proposed potential term frequency and inverse document frequency (potential TFIDF) and entropy. Tags are ranked according to the video-to-concept distribution scores, and the tags with the highest scores are being propagated to the near-scenes detected. Extensive experiments are demonstrate that the proposed PIM outperforms state-of-the-art approaches compared in terms of the quality segments (QS) and the quality frames (QF) for near-scene detection. Furthermore, the proposed framework of the hierarchical video-to-near-scene annotation is achieved the high quality of the near-scene annotation in terms of mean average precision (MAP).

V. VIDEO ANNOTATION

Video annotation refers assigning videos a set of labels which describes their content in both syntactic and semantic levels. The labels facilitate various kinds of video manipulations, such as retrieval and summarization. However, giving the proper label to videos is a challenging task due to video data scale and the complexity of video contents.

To perform video annotation the different steps has to be performed such as key feature extraction, object-level recognition, scene-level recognition. The following is the survey of different video annotation process that is performed.

Hongsen Liao et.al (2015), proposed a visualization based batch mode sampling method to handle problem. An iso-contour based scatter plot is used to provide intuitive clues for representativeness and informativeness of samples and assist users in sample selection. A semi-supervised metric learning method is incorporated to generate an effective scatter plot reflecting the high-level semantic similarity for visual sample selection. Moreover, both quantitative and qualitative evaluations are provided to show that visualization based method can be effectively enhance sample selection in active learning.

Zengkai Wang et.al(2016), have proposed a soccer video annotation approach based on semantic matching in coarse time constraints, where video events and external text information (match reports) are being synchronized using semantic correspondence in temporal sequence. Unlike the state-of-art soccer video analysis methods that assume time of an event's occurrence is given precisely by external text information, this work describes the problem of annotating soccer videos using match reports with coarse-grained time information.

The contributions of proposed approach are: 1) We propose more generalized approach that synchronizes video events with text descriptions using high-level semantics with coarse time constraints, rather than assuming that timestamp is given exactly in text description; 2) The detection of event boundaries is being improved by attack–defense transition analysis (ADTA); 3) A robust and fast center circle detection algorithm is being proposed for classification of soccer field zones and ADTA; 4) Unlike conventional audio-based whistle detection, we propose a novel Hough transform (HT) based algorithm for perspective of image processing. This allows the game start time that is to be detected, and further helps synchronization of video and text events. Experimental results conducted on large number of soccer videos, validate the effectiveness of the proposed approach.

Wenjing Tong et.al (2015), proposed novel video shot boundary detection of framework based on interpretable TAGs learned by Convolutional Neural Networks (CNNs). Firstly, we adopt candidate segment selection to predict the positions of shot boundaries and discard most non-boundary frames. The preprocessing method can help to improve both accuracy and speed of the SBD algorithm. Then, cut transition and gradual transition detections which are based on interpretable TAGs are conducted to identify the shot boundaries in candidate segments. Afterwards, we synthesize the features of frames in shot and get semantic labels for the shot. Experiments on TRECVID 2001 test data show that proposed scheme can achieve a better performance compared with state-of-the-art schemes. Besides, the semantic labels obtained by framework can be used to depict the content of shot.

Aftab Khan et.al(2014), have proposed four variants of novel hierarchical hidden Markov models strategy for rule induction in context of automated sports video annotation including a multilevel Chinese takeaway process (MLCTP) based on Chinese restaurant process and novel Cartesian product label-based hierarchical bottom-up clustering (CLHBC) method that employs prior information contained within label structures. Our results show significant improvement by comparison against flat Markov model: optimal performance is obtained using hybrid method, which combines MLCTP generated hierarchical topological structures with CLHBC generated event labels and also show that methods proposed are generalizable to other rule-based environments including human driving behavior and human actions.

VI. CONCLUSION:

In this survey, we discussed about the Video annotation, Characteristics, Key feature extraction, Object-level recognition, Scene-level recognition and also done a detailed survey on various object and scene-level recognition in video annotation. The issues in this are also detected such as 1.Only small set of objects are being used. 2.The analysis done in the object and scene-level recognition are less accurate. 3.Object and scene-level recognition are done only in specific manner.

REFERENCES:

[1]. Aaron O. Thomas*, Pavlo D. Antonenko, Robert Davis (2016), “Understanding meta comprehension accuracy within video annotation Systems”, Computers in Human Behavior 58 269e277.

[2]. Aftab Khan, David Windridge, and Josef Kittler (2014),” Multilevel Chinese Takeaway Process and Label-Based Processes for Rule Induction in the Context of Automated Sports Video Annotation”, IEEE transactions on cybernetics, vol. 44, no. 10, october.

[3]. Amir H. Shabani , John S. Zelek , David A. Clausi(2013),”multiple scale specific representations for human action recognition”, Pattern Recognition Letters 34 (2013) 1771–1779

[4]. Chien-Li Chou, Student Member, IEEE, Hua-Tsung Chen, Member, IEEE, and Suh-Yin Lee, Senior Member (2016),” Multi-Modal Video-to-Near-Scene Annotation”, IEEE 1520-9210 (c).

[5]. Christoph Feichtenhofer, Axel Pinz Richard, P. Wildes (2016), “Dynamic Scene Recognition with Complementary Spatiotemporal Features”, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[6]. Fatemeh Tabib Mahmoudi, Fellow, IEEE, Farhad Samadzadegan, and Peter Reinartz, Jr., Member (2015),”Object Recognition Based on the Context Aware Decision-Level Fusion in Multiviews Imagery”, IEEE journal of selected topics in applied earth observations and remote sensing, vol. 8, no. 1, january.

[7]. Hao Wang, Dit-Yan Yeung Senior Member (2015),“Towards Bayesian Deep Learning: A Framework and Some Existing Methods”, IEEE 1041-4347 (c) 2016 IEEE.8. journal of latex class files, vol. 14, no. 8, august,

[8]. Hongsen Liao, Li Chen, Yibo Song, Hao Ming (2013),”Visualization Based Active Learning for Video Annotation”, IEEE .

[9]. Iván González-Díaz , Tomás Martínez-Cortés, Ascensión Gallardo-Antolín, Fernando Díaz-de-María (2015),” Temporal segmentation and keyframe selection methods for user-generated video search-based annotation”, Expert Systems with Applications 42 488–502.

[10]. Jose M. Chaquet , Enrique J. Carmona , Antonio Fernández-Caballero,” A survey of video datasets for human action and activity recognition “,Computer Vision and Image Understanding 117 (2013) 633–659.

[11]. Kee-SungLee , Ahmad Nurzid Rosli , Ivan Ariesthea Supandi , Geun-SikJo (2014) ,” Dynamic sampling-based interpolation algorithm for representation of clickable moving object in collaborative video annotation”, Neurocomputing146(2014)291–300.

[12]. M.Ravinder and T.Venugopal(2016),”Content-Based Video Indexing and Retrieval using Key frames Texture, Edge and Motion Features”, International Journal of Current Engineering and Technology Vol.6, No.2 April

[13]. Marina Ivasic-Kos , MiranPobar , SlobodanRibaric (2016),“Two-tier image annotation model based on a multi-label classifier and fuzzy-knowledge representation scheme”, , PatternRecognition52 287–305.

[14]. Simon Jones, Ling Shao (2013), “Content-based retrieval of human actions from realistic video databases”, Information Sciences 236 56–65.

[15]. Thomas B. Moeslund , Adrian Hilton, Volker Kru¨ger,” A survey of advances in vision-based human motion capture and analysis”, Computer Vision and Image Understanding 104 (2006) 90–126.

[16]. Wenjing Tong1, Li Song1,2, Xiaokang Yang1,2, Hui Qu1, Rong Xie1,2 (2008), ”CNN-Based Shot Boundary Detection and Video Annotation”.

[17]. Yirui Wu, Palaiahnakote Shivakumara, Tong Lu, Member, IEEE, Chew Lim Tan, Michael Blumenstein, Senior Member, IEEE, and Govindaraj Hemantha Kumar (2016), “Contour Restoration of Text Components for Recognition in Video/Scene Images”, IEEE transactions on image processing, vol. 25, no. 12, december.

[18]. Zengkai Wang, Junqing Yu, Member, IEEE, and Yunfeng He (2015),” Soccer video event annotation by synchronization of attack defense clips and match reports with coarse- grained Time information.”, Neurocomputing.