Editorial

ICSES Transactions on Data Science, Engineering and Technology (ITDSET)

Journal Homepage: www.i-cses.com/itdset

Automatic Semantic Video Annotation

A. A. Kalaivani [1,*] and B. RajaSuguna [2]

[1] Assistant Professor (SG), Department of CSE, Chennai, India.
[2] Assistant Professor, Department of CSE, Chennai, India

* Corresponding Author: kalaivania.sse@saveetha.com

THE rapidly increasing quantity of publicly available videos has driven research into developing automatic tools for indexing, rating, searching and retrieval. Textual semantic representations, such as tagging, labeling and annotation, are used to represent appropriate semantics for search and retrieval. The semantics should be inspired by the human cognitive way of perceiving to describe videos. The difference between the low-level visual contents and the corresponding human perception is referred to as the *'semantic gap'*. Tackling this gap is harder in the case of unconstrained videos due to lack of semantics knowledge.

Video Analysis

Video based applications - such as video surveillance, road traffic control, sports events detection requires a strong human intervention when a semantic understanding of contents is needed to detect objects, actions or events within a video stream. Manual analysis of video sequences is a very time consuming task and it often leads to inaccurate results due to the "video blindness". In the video surveillance domain, for example, it has been stimulated that an operator can miss up to 95% of scene activities after only 22 minutes of analysis.

In the last years, great efforts by the computer vision research community leads to the development of robust and reliable algorithms for video analysis tasks at different levels:

- Low-level video analysis methods address the ability to find the image regions corresponding to objects of interest (detection) and then track them across different frames while maintaining the correct identities (tracking) .
- Mid-level video analysis methods face the problem of recognizing simple or "atomic" events or activities
- High-level video analysis methods concentrate on the detection of "complex" events or activities

While low-level processing aims at generating feature descriptions to summarize characteristics of data in a quantitative way, high-level processing is more related to the interpretation and reasoning with visual data: it takes features descriptors as input and generates abstract, qualitative descriptions about contents, addressing the so called semantic gap problem . Through the years, high-level methods have been implemented using various forms of artificial intelligence, from symbolic knowledge representation, based on hierarchical structures using semantic networks or more traditional object oriented data models , up to rule-based systems. Many works have been focused on improving the ability to search and retrieve specific contents from large repositories.

Video Annotation Techniques

Video annotation also known as "Video semantic annotation" can be performed in three of the following techniques: Traditional manual annotation, rule-based annotation and machine learning technique. Manual annotation is laborious, time consuming, ambiguous, too subjective and error prone process. Furthermore, this technique has low efficiency and speed [2]-[3]. Rule-based annotation technique classifies the annotation using expert knowledge. Since, commonly the undeveloped hierarchical annotation forms are used then it neither cover all the semantic content of video nor the versatile requirement of video annotations [1]. The third group, machine learning technique, can act as a supervised classification task to use in automatic video annotation to cope with the weaknesses of the other techniques [4].

The key idea of Automatic Video Annotation (AVA) is to construct the model(s) through automatic learning of semantic concept from many videos (even shots or keyframes), then to utilize these concept model(s) for predicting appropriate annotation/label for any new video. Later, these annotated videos can be retrieved by textual queries. However, the performance of this

framework highly depends on: First, the video content representation; second, feature extraction; third, feature selection which means to choose more visual features for training classifiers result in: (1) Having various visual characteristic of video, (2) Improve the classifiers capability to recognize different video concepts and (3) Enhance classification accuracy. Fourth, employing an effective classification algorithm and proper dataset since any inappropriate use of dataset for constructing model during training stage will lead to deterioration in the performance of AVA due to the lack of adequ-ate concepts in their annotation vocabularies [6].

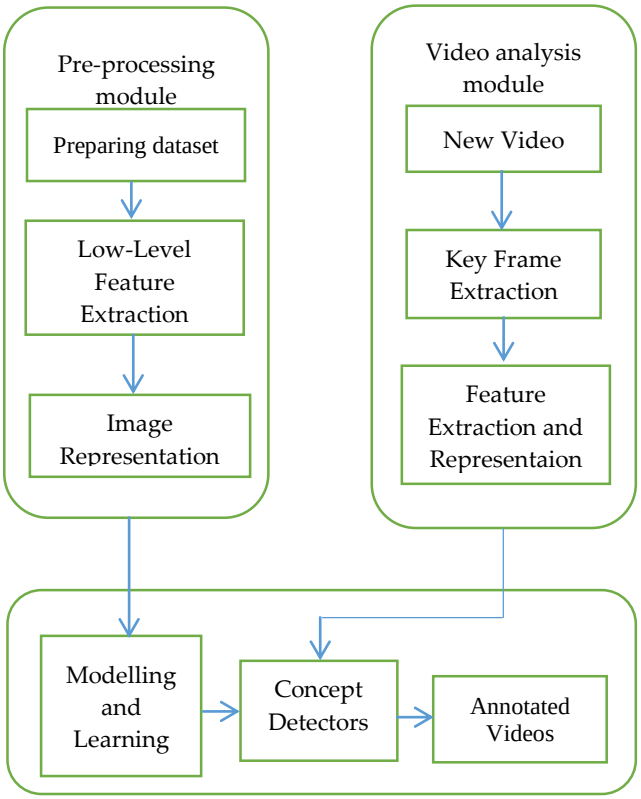

Figure 1. Framework for Video Semantic Annotation

Video Annotation Tools

Video annotation usually takes the shot as the basic unit; the video content feature on the other hand is extracted from the keyframes.Thus,theimagefeature is the major feature of the video. Traditional global feature such as color and texture generally encounter the complex difficulty expressing semantic information. But recently the booming technologies based on local features including scale-invariant feature transform (SIFT), rotation-invariant feature transform (RIFT), bag-ofwords (BoW), and bag-of-features (BoF) represent enormous potential to convey semanteme.

Video annotation tools make a rather poor utilisatiotin of Semantic Web technologies and formal meaning, XML being the most common choice for the capturing and representation of the produced annotations. The use of MPEG-7 based descriptions, may constitute a solution towards standardised video descriptions, yet raises serious issues with respect to the automatic processing of annotations, especially the descriptive ones, at a semantic level. The localisation of temporal segments is performed mostly manually, indicating the issues involved in automatically identifying the time interval corresponding to the semantic notion addressed by the annotation; only Advene, SVAT and VideoAnnex perform automatic shot detection. Furthermore, VideoAnnex, VIA and SVAT[5] are the only ones annotation of spatial regions on frames of the video, as well. Anvil has recently presented a new annotation mechanisms called spatiotemporal coding aiming to support point and region annotation, yet currently only points are supported.

Two main issues in semantic annotation of unconstrained videos emerge are : firstly, the extraction of a compact representation composed of spatio-temporal features, suitable for efficient matching of objects, actions and scenes; and secondly, the representation and use of semantic relationships between objects, actions and scenes to validate annotation, compensating for the limitations of the raw visual information, including variable appearance, occlusions and ambiguity.

A challenging issue in video annotation concerns the representation of structural and by consequence temporal information in an erective manner so as to avoid overwhelming volumes of metadata.

Reagards,
A. A. Kalaivani
August 2018

REFERENCES

[1] Dorado, A., J. Calic and E. Izquierdo, "A rule-based video annotation system", *IEEE Trans. Circ. Syst. Video Technol*., vol. 14, pp. 622-633, 2004.

[2] Settles, B., M. Craven and L. Friedland, "Active learning with real annotation costs", in Proc. of *the NIPS Workshop on Cost-Sensitive Learning*, pp. 1-10, 2008. URL: http://burrsettles.com/pub/settles.nips08ws.pdf

[3] Tang, J., X.S. Hua, G.J. Qi, Z. Gu and X. Wu, "Beyond accuracy: Typicality ranking for video annotation", in Proc. of *the IEEE International Conference on Multimedia and Expo*, Beijing, July 2-5, 2007, pp. 647-650.

[4] Tao, D., D. Xu and X. Li, *Semantic Mining Technologies for Multimedia Databases*. 1st Edn., Information Science Reference, New York, 2009. ISBN-13: 978-1605661889

[5] S. Dasiopoulou, E. Giannakidou, G. Litos, P. Malasioti, and Y. Kompatsiaris, "A Survey of Semantic Image and Video Annotation Tools," Lecture Notes in Computer Science, pp. 196–239, 2011.

[6] Luca Greco, Pierluigi Ritrovato, Mario Vento, "On the use of semantic technologies for video analysis", *Semantic Web*, pp. 1-21, 2017.