A DEEP LEARNING BASED APPROACH FOR PRECISE VIDEO TAGGING

Unmesh Papat, Prof. V.M. Lomte, Himanshu Deshmukh, Akash Sonwane, Rohit Shendage

Department of Computer Engineering, RMD Sinhgad School of Engineering,SPPU,India

ABSTRACT:

Video tagging is a complex problem that combines single-image feature extraction with arbitrarily long sequence understanding. By improving at the task of tagging videos with useful metadata labels, we necessarily improve our ability to understand the content and context of video data. A lot of existing methods fail to precisely tag videos because of their lack of ability to capture the video context. The context in a video represents the interactions of objects in a scene and their overall meaning. In this work, we propose a novel approach that integrates the video scene ontology with CNN based YOLO algorithm for improved video tagging.

Keywords: *YOLO algorithm, video tagging, feature extraction, post processing*

INTRODUCTION

Recent years have seen a rapid increase in the size of video collections due to the popularity of the Internet and of portable cameras, such as the smart-phone. These video collections have become the medium for many people to communicate and to find entertainment with the development of online services like YouTube and Vimeo. Finding match for a user submitted query is challenging on large multimedia data. To reduce the empirical search, video hosting websites often allow users to attach a description with the video. However, description or index terms can be ambiguous, irrelevant, insufficient or even empty. This creates a necessity for automatic video tagger. In this paper, we present an automatic video tagging tool, using CNN based YOLO for precise video tagging. It involves video segmentation that extracts distinct, representative frames from the input video by hierarchical combination of various image similarity metrics. In the next step, raw tags obtained from the segmented video frames are investigated to estimate semantic similarity information. Finally, we annotate the input video by combining raw tags with the inferred tags.

From a practical standpoint, there are currently no video classification benchmarks that match the scale and variety of existing image datasets because videos are significantly more difficult to collect, annotate and store. The goal for this project was to generate a classifier that most accurately labels a collection of

videos with up to appropriate tags that denote the genre and context of the video. The objective of our algorithm is to capture the overall theme of the video by extracting the key information from individual key frames and summarizing it into a few but precise tags. Automatic video tagging has wide spectrum of promising applications e.g., classification of videos into known categories, detecting prohibited categories (e.g. violence in kid's movies), and context based video search, and efficient archiving etc., motivating the interest of researchers worldwide. It can forever change the experience of video users and would allow better contextual searches.

Images and videos have become ubiquitous on the internet, which has encouraged the development of algorithms that can analyze their semantic content for various applications, including search and summarization. Recently, Convolutional Neural Networks (CNNs) have been demonstrated as an effective class of models for understanding image content, giving state-of-the-art results on image recognition, segmentation, detection and retrieval.

In this work, we propose a novel approach that integrates the video scene ontology with CNN (Convolutional Neural Network) based YOLO (You Only Look Once) for improved video tagging. Our method captures the content of a video by extracting the information from individual key frames. The key frames are then fed to a CNN based deep learning model to train its parameters. The trained parameters are used to generate the most frequent tags. Highly frequent tags are used to summarize the input video.

LITERATURE SURVEY

Today's digital contents are inherently multimedia: text, audio, image, video, and so on. Video, in particular, has become a new way of communication between Internet users with the proliferation of sensor-rich mobile devices. Accelerated by the tremendous increase in Internet bandwidth and storage space, video data has been generated, published, and spread explosively, becoming an indispensable part of today's big data. This has encouraged the development of advanced techniques for a broad range of video understanding applications including online advertising, video retrieval, video surveillance, etc. A fundamental issue that underlies the success of these technological advances is the understanding of video contents. Recent advances in deep learning in image and speech domains have motivated techniques to learn robust video feature representations to effectively exploit abundant multimodal clues in video data. Review of literature is given in table 1.

Table 1 Literature review

Sr. no.	Paper name	Author name, publication, year	System proposed	Techniques/ tools	Remark
1	A Deep Learning based Approach for Precise Video Tagging	Sadia Ilyas, Hafeez Ur Rehman, IEEE, 2019	The content of a video captured by extracting the information from individual key frames. The key frames are then fed to a CNN based deep learning model to train its parameters. The trained parameters are used to generate the most frequent tags. Highly frequent tags are used to summarize the input video.	CNN	This method managed to achieve an overall accuracy of 99.8% with an F1-score of 96.2%.
2	Content-based Video Emotion Tagging Augmented by Users' Multiple Physiological Responses	Wang, S., Chen, S., & Ji, Q., IEEE, 2017	An implicit hybrid video emotion tagging approach that integrates video content and users' multiple physiological responses, which are only required during training is proposed.	SVM	Proposed method with the help of physiological signals outperforms the baseline method, which uses video signals only. FI-score is 95%
3	Patwardhan, A. A., Das, S., Varshney, S., Desarkar, M. S., & Dogra, D. P.	ViTag: Automatic Video Tagging Using Segmentation and Conceptual Inference, IEEE, 2019	It involves video segmentation that extracts distinct, representative frames from the input video by hierarchical combination of various image similarity metrics. In the next step, raw tags obtained from the segmented video frames are investigated to estimate semantic similarity information. Finally, the input video is annotated by combining raw tags with the inferred tags.	natural language processing package (NLTK)	On a dataset of 103 videos belonging to 13 domains derived from various YouTube categories, system generates tags with 65.51% precision and 87% accuracy using reciprocal rank as a metric. The geometric mean of Reciprocal Rank estimated over the entire collection has been found to be 0.873.

4	Technical Analysis of Multi -Text Video Standardization Based on Tag System	Yuan Zhang; Shufeng Li, IEEE, 2020	The text analysis method is used to extract the video keywords to assist the personalized recommendation of the video.	http adaptive streaming (HAS)	
5	Tag-based Video Retrieval with Social Tag Relevance Learning	Takeda, H., Yoshida, S., & Muneyasu, M., IEEE, 2019	A formula for calculating the tag relevance score is formulated considering the tag occurrence frequency imbalance	Tag neighbor voting algorithm	This approach is effective and efficient
6	Incorporating Geo- Tagged Mobile Videos into Context-Aware Augmented Reality Applications	To, H., Park, H., Kim, S. H., & Shahabi, C., IEEE, 2016	This paper investigated three general approaches for incorporating video content into AR applications; pre- defined content, on-demand content, and suggested content by hotspots.	Augmented reality, tracking algorithm, hotspot algorithm	The proposed algorithms are fast and able to find interesting video segments. Also, the hotspot algorithm efficiently found all hotspots in the dataset.
7	Knowledge-Augmented Multimodal Deep Regression Bayesian Networks for Emotion Video Tagging	Shangfei Wang, Longfei Hao, and Qiang Ji, IEEE, 2020	A MMDRBN is proposed to capture the relationship between audio and visual modalities for emotion video tagging. The structure of the MMDRBN is modifying to incorporate domain knowledge. The main audio and visual elements are also summarized used by filmmakers to convey emotions and formulate them as semantical meaningful middle-level representation, i.e., attributes.	multimodal deep regression Bayesian network	the LIRIS-ACCEDE database demonstrate that this model successfully captured the intrinsic connections between audio and visual modalities, and integrate the middle-level representation learning from video data and semantical attributes summarized from film grammar.

8	Personalized video emotion tagging through a topic model.	Wu, S., Wang, S., & Gao, Z., IEEE, 2017	During training, the proposed topic model exploits the latent space to model the relationships among personal characteristics, video content and video tagging behaviors. After learning, this model generates meaningful latent topics, which help personalized video emotion tagging.	Modified LDA, topic model	Accuracy, FI score and kappa is 72.2%, 0.643 and 0.418 respectively
9	Crowd sourced time-sync video tagging using semantic association graph	Yang, W., Ruan, N., Gao, W., Wang, K., Ran, W., & Jia, W., IEEE, 2017	SW-IDF first generates corresponding semantic association graph (SAG) using semantic similarities and timestamps of the time-sync comments. Then it clusters the comments into sub-graphs of different topics and assigns weight to each comment based on SAG. This can clearly differentiate the meaningful comments with the noises. In this way, the noises can be identified, and effectively eliminated	Semantic Weight-Inverse Document Frequency (SW-IDF)	This system achieves 0.3045 precision and 0.6530 recalls in high-density comments; 0.3800 precision and 0.4460 recalls in low-density comments. F-score-0.4153
10	Tagging and Solution-Based Video Recommendations in Learning Video Environments	Lehmann, A., IEEE, 2019	An approach is presented to remove obstacles in a learning video environment	Supervised learning	This system allows socializing such a learning environment by using problem tagging to recommend learning videos regarding a specific problem, recommended by other learners, to speed up and facilitate the learning process for each learner

PROPOSED SYSTEM

Video captioning is a new problem that has received increasing attention from both computer vision and natural language processing communities. Given an input video, the goal is to automatically generate a complete and natural sentence, which could have a great potential impact, for instance, on robotic vision or on helping visually impaired people. Nevertheless, this task is very challenging, as a description generation model should capture not only the objects, scenes, and activities presented in the video, but also be capable of expressing how these objects/scenes/activities relate to each other in a natural sentence. In this section, we discussed proposed model for captioning/tagging using deep neural network. Block diagram of proposed system is given in fig 3.1

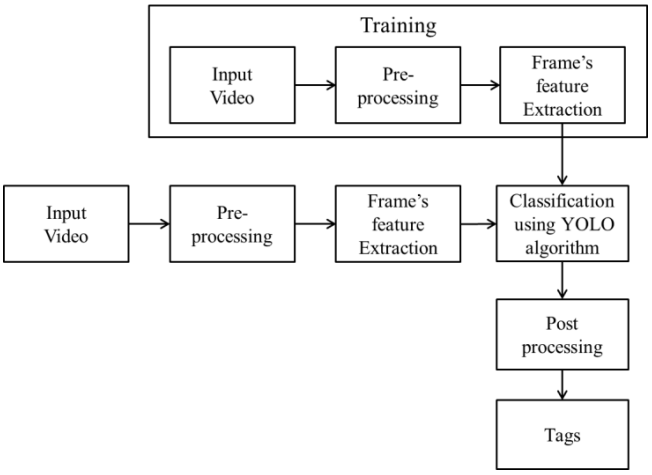

Fig 1 block diagram of proposed system

Input video is an offline video of arbitrary length from database (for training) and browsed video (testing video). Pre-processing is a common name for operations with images at the lowest level of abstraction both input and output are intensity images. The aim of pre- processing is an improvement of the image data that suppresses unwanted distortions or enhances some image features important for further processing. At first input videos are to processed to extract frames. The frame dimensions differ, so to bring them into uniform dimensions we resize them to 324 x 240 which is the minimum size frame in the dataset. Before discussing the extraction of feature points it is necessary to have a measure to compare parts of images.

To avoid redundancy we extracted key frames. Key frames are the video frames that are not too similar. The extraction and matching of features of each key frame is based on these measures. Besides the simple point feature a more advanced type of feature is also presented. Feature extraction technique is used to extract the features by keeping as much information as possible from large set of data of image. Dataset is given to train CNN. Classification is performed using CNN.

The output of the previous step i.e., a pool of classes undergoes some post processing. This phase decides the resulting tags of the input video. Individual key frames of the video may belong to more than one target tags of the video ontology. The final tags are decided by pruning the labels that are less dominant in terms of their occurrence. Repetition is removed and the count of occurrence of the predicted classes is stored. It is then normalized by taking the average. The classes having higher occurrence count are selected as the final tags for the input video.

CONCLUSIONS

Video tagging is the process of associating useful information with the content of the video. Improved indexing and automated analysis of millions of videos could be accomplished by getting videos tagged automatically. We proposed a novel method for accurate video tagging using YOLO algorithm by utilizing the video based action ontology. The proposed technique has a number of interesting uses, for example, genre classification of movies to: adventure, romance, sci-fi, action, etc.

REFERENCES

[1] Sadia Ilyas, Hafeez Ur Rehman, "A Deep Learning based Approach for Precise Video Tagging", 2019 15th International Conference on Emerging Technologies (ICET) 10.1109/ICET48972.2019.8994567

[2] Wang, S., Chen, S., & Ji, Q. (2017). Content-based Video Emotion Tagging Augmented by Users' Multiple Physiological Responses. IEEE Transactions on Affective Computing, 1–1. doi:10.1109/taffc.2017.2702749

[3] Patwardhan, A. A., Das, S., Varshney, S., Desarkar, M. S., & Dogra, D. P. (2019). ViTag: Automatic Video Tagging Using Segmentation and Conceptual Inference. 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM). doi:10.1109/bigmm.2019.00-12

[4] Yuan Zhang; Shufeng Li, "Technical Analysis of Multi - Text Video Standardization Based on Tag System", 2020 International Conference on Culture-oriented Science & Technology (ICCST) IEEE, **DOI:** 10.1109/ICCST50977.2020.00015

[5] Takeda, H., Yoshida, S., & Muneyasu, M. (2019). Tag- based Video Retrieval with Social Tag Relevance Learning. 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE). doi:10.1109/gcce46687.2019.9015338

[6] To, H., Park, H., Kim, S. H., & Shahabi, C. (2016). Incorporating Geo-Tagged Mobile Videos into Context-Aware Augmented Reality Applications. 2016 IEEE Second International Conference on Multimedia Big Data (BigMM). doi:10.1109/bigmm.2016.64

[7] Shangfei Wang, Senior Member, IEEE, Longfei Hao, and Qiang Ji, "Knowledge-Augmented Multimodal Deep Regression Bayesian Networks for Emotion Video Tagging", 1520-9210 (c) 2019 IEEE Transactions on Multimedia (Volume: 22, Issue: 4, April 2020), **DOI:** 10.1109/TMM.2019.2934824

[8] Wu, S., Wang, S., & Gao, Z. (2017). Personalized video emotion tagging through a topic model. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi:10.1109/icassp.2017.7952680

[9] Yang, W., Ruan, N., Gao, W., Wang, K., Ran, W., & Jia, W. (2017). Crowdsourced time-sync video tagging using semantic association graph. 2017 IEEE International Conference on Multimedia and Expo (ICME). doi:10.1109/icme.2017.8019364

[10] Lehmann, A. (2019). Problem Tagging and Solution- Based Video Recommendations in Learning Video Environments. 2019 IEEE Global Engineering Education Conference (EDUCON). doi:10.1109/educon.2019.8725254