Automatic Concept Detector Refinement for Large-Scale Video Semantic Annotation

Xueliang Liu and Benoit Huet
EURECOM Institute
Sophia-Antipolis, France
{xueliang.liu, Benoit.huet}@eurecom.fr

***Abstract*— With the explosion of sharing website, an infinite amount of multimedia items are made available for all on a day to day basis. Since search engine technologies rely essentially on textual information there is an urgent need to infer relevant semantic description through content based analysis on those multimedia documents. In this paper, we propose an approach which leverages the sheer volume of data available online to refine semantic concept detectors for videos annotation without requiring any additional human interaction. To address the problem in a realistic setting, we have collected a large video collection of about 42 thousand videos crawled from YouTube. A number of low-level features are extracted from those videos and are comprised within the corpus. Upon training on a small initial set of labeled video shots, the concept detectors are run on the large scale unlabeled corpus in order to identify and select new training samples. Thanks to this inexpensively obtained set of new training examples the concept detectors can be reinforced and enhanced based on a wider number of unlabeled samples and therefore better adapt to the corpus at hand. The experimental results reported here show that indeed the annotation accuracy improves when the training set is extended with automatically labeled samples.**

Keywords- Keywords— Online Video, Dataset, Annotation

I. INTRODUCTION

Semantic video analysis plays an important role in video indexing and retrieval. The annotation techniques, as one of the challenges in this field, allow us to categorize video data automatically. There has already been much prior work in this area [1-3]. However, in spite of the great progress made in the last decade in content-based multimedia analysis, state of the art approaches are still far to reach the level required to deal with the sheer volume of videos uploaded continuously by content sharing website users. Web 2.0 technologies have enabled many multimedia content sharing websites such as YouTube[1], Yahoo!Video[2], DailyMotion [3]. The growth of online videos creates new challenges for indexing and retrieving the embedded content. The research on online video needs to cope with the large amount of videos available and the unconstrained number of concepts present. The analysis of these shared videos shows a potential in improving the performance of traditional multimedia information analysis approaches and bridging the semantic gap between objective multimedia content analysis and subjective users' impressions.

In this paper, we address the problem of online video annotation, using content-based information originating from visual characteristics to extend the training set without extra human annotation efforts. Firstly, we collect an online video dataset for our research, and then we develop an application to annotate automatically unlabeled videos with a set of semantic concepts. The rest of this paper is organized as follows: Section II reviews the related works. In section III, we describe how the large scale dataset was collected and extended with content based features. In section IV, we present the online video annotation system along with the automatic training data enhancement mechanism proposed. The results are then reported and discussed. Finally, section V gives a summary and the perspective of this work.

II. RELATED WORK

We introduce the related work from two perspectives: Annotation algorithms on one side and the user generated and shared multimedia content and platforms on the other side.

Video annotation research in the multimedia academic community has devoted most of its attention to mapping low-level features to high-level semantic concepts with learning algorithms. Those learning-based approaches try to discover semantic pattern in low-level visual feature space, which may subsequently be employed to realize content-based video search. Many traditional machine learning algorithms such as the SVM **Error! Reference source not found.**, K-NN [1], semi-supervised learning [3], and active-learning [4] are employed to improve the precision and recall performance. Those approaches are a promising direction to enable content-based video search. However, due to the complexity of both video dataset and semantic corpus, existing techniques for automatic video annotation are suffering from the difficulties of dealing with large-scale video dataset and large-scale concept set, in terms of both annotation accuracy and computational cost.

With the popularity of digital cameras and the prosperity of video sharing portals, it becomes easier to access those digital video from the internet. Unlike traditional video collections, web video collections have unlimited concept vocabulary and rich metadata such as filename, tags and brief description, which are potentially useful to index videos with the assistance of those text. To address the problem of semantic web video analysis, Chua et al. have created a web

[1] http://www.youtube.com
[2] http://video.yahoo.com
[3] http://www.dailymotion.com/fr

image dataset for image concept analysis [5]. Similarly, Juan Cao et al. have released a web video dataset [6]. Beside these web video datasets built very recently, some studies have been presented on the concept detection or tags analysis in online image and videos. In [7], A. Ulges et al. built a group of concept detectors based on the online videos. And in [8], Lei Wu et al. proposed and implemented an automatic tagging recommendation system for large-scale web image retrieval.

In this paper, we build a web-scale video dataset and adopt a feedback model for automatically annotation. With this model, training set is extended with automatically labeled samples, and the concept detectors are reinforced and enhanced.

III. OUR DATASET

For the purpose of our research, a very large and realistic collection of videos along with the metadata available (categories/tags, description, and so on) is needed. Here we detail the way we proceeded to create our publicly available dataset.

A. Video download

A well designed dataset is of importance for research addressing the online video analysis problem. The size of the dataset should be as large as we can possibly handle and contain very diverse items. However, we should also notice that it is not possible to collect a really web-scale video dataset because of two reasons: Firstly, as there is an infinite amount of videos on the internet, it is not realistic to attempt to capture them all with common lab equipment, and secondly, online videos are updated dynamically and are occasionally removed sometimes. Let us take YouTube, the renowned video sharing site, as an example. It is reported that this portal offers more than hundreds of millions video entities. About 65 thousands videos are uploaded every day and lots of videos are removed when they are out of date. With this in mind, we decided to focus on the collection of the most representative videos for the construction of our dataset.

To construct a representative web video dataset, we download the videos and their respective metadata from YouTube with the assistance of its public API[4]. Firstly, we retrieve the most popular videos information in a whole month (from May 1st 2009 to June 1st 2009) and download 635 videos. From those videos' meta-data, we collect 1875 meaningful tags after removing lapses words. Secondly, we use those tags as the new seeds to retrieve videos and download about 42,000 YouTube videos. For each of the downloaded entry, the processing described in following subsections is performed to extract features for each video.

B. Low level visual features

Videos are segmented into shots and for each shot three representative keyframes are extracted based on the method proposed in [9]. In this algorithm, the shot boundary is detected based on the color-histogram. Then we employ the approaches to construct the NUS-WIDE dataset [5] to extract the low level visual features for our dataset. The low level feature used in this dataset are 64-D Color histogram, 144-D Color auto-correlogram, 73-D Edge direction histogram ,225-D Block–wise Color moment, and 128-D SIFT feature.

C. The training/evaluation subset

Having detailed the set of features used to represent the shot content, we shall now describe the part of the corpus used for initial training and algorithm evaluation. Though it is more accurate to annotate videos manually for a standard training dataset, it is also a tedious and labor-intensive process. Here, we use the method proposed in [7] to collect the videos with a special concept. The basic idea of this method is to use complex Boolean words to query from the video shared website, which has shown competitive performance. We first query videos from YouTube with a keywords and refine the query words after checking the result manually, and then loop this process until all of the returned videos are high relevant with the query concept. At last we download all of these queried videos. It should be noticed that the web video tags are not labeled on shot level, so we just query the videos whose length is less than 3minutes in practice, in order to guarantee most of the video shots are highly relevant with the concept. In practice, we have queried YouTube with the 39 concepts used in TrecVid 2006 and downloaded about 8000 videos. Some Boolean query examples used for collecting the concept-based dataset are list in TABLE I. . Furthermore, we remove the concept whose queried results are too less and avoid the imbalance issues for different concept. We also use the KNN algorithm to remove the noisy shots, resulting in a pruned video dataset composed of 5,712 videos containing 25 semantic concepts. We should argue that the query words used here cannot cover the intentional and extensional meaning of those concepts, but it is indeed an effective way to find the representative video shots.

TABLE I. BOOLEAN QUERY WORDS EXAMPLES

Concept	Queries	YouTube Category
Weather	Weather –forecast –song	Events
Outdoor	Mountain & river	Travel&events
Building	Building &city&view	--
Mountain	Mountain&tavel	Travel&events
Sky	Cloud –music –computing	science
Maps	map geography	education

Out of those 5712 videos, 50% are kept for evaluating the algorithm's performance while the remainder serves for training the initial detectors.

IV. AN ANNOTATION APPLICATION

Automatic semantic concept detection plays an important role in content-based video search. Here, we propose an annotation application on our dataset. The purpose of our research is to investigate the possibility to use the online

[4] http://www.youtube.com/dev

videos and their associated user generated content, as an alternative video source for semantic analysis.

A. The feedback annotation model

Figure 1 illustrates the feedback model we propose in this paper to annotate as many shots as possible from the unlabeled pools with the concept detectors. In this model, a semi-supervised self-learning strategy [12] is employed. Firstly, we initiate the training of the concept detectors with the previously labeled subset as described in section III.C. We run those newly trained detectors on the unlabeled video pools collected from YouTube. The videos shots that are highly similar with a concept, in other words, when the probability estimation output on the concept detector is above a given threshold, will be added to the training set with the aim of improving the correct detection rate for the corresponding concept. The concept detectors are then re-trained on the automatically extended training set. Both the original concept detector and the retrained ones are then evaluated on the testing dataset which has been held out for performance evaluation only.

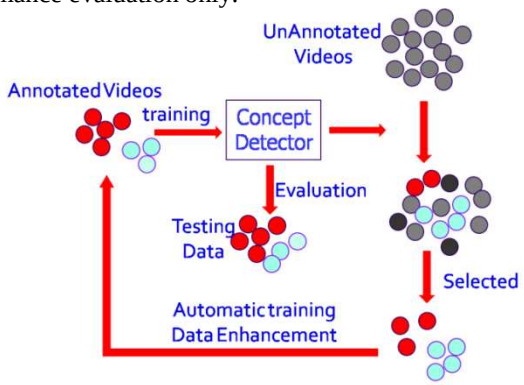

Figure 1. The feedback annotation model

B. Concept detectors training

We use the 2806 training videos, annotated with the 25 concepts to initially train the concept detectors. The features used in the experiment reported here are the global color moment (CM) and local SIFT feature as described in section III.B, which have been shown efficient and effective in generic concept detection [10]. Based on the two selected features, one-vs.-all classifiers are trained for the task of detecting each individual concept. The training algorithm we used here is nonlinear Support Vector Machine (SVM) implemented by the latest LIBSVM [11] with a Radial Basis Function (RBF) kernel. We use the cross-validation method to determinate the parameter setting of the SVM models. When the concept detectors are obtained, we apply them on the un-annotated pool to annotate those unlabeled shots automatically.

C. Automatic selection of new training samples

We want to validate the idea that enhancing the training set with automatically labeled data is able to improve the performance of concept detectors in terms of accuracy. In order to save the computational time, half of our unlabeled data is used as the unlabeled data. There are about 21,000 videos and 424,000 shots totally. The CM feature on those shots is used in our experiments. However, only a third subset of SIFT feature is used because of the limitation of memory storage. We run the trained concept detectors on those video shots and select automatically new training samples based on the result. If the output probability of a video shot for a SVM classifier is higher than a given threshold, the video shot is reserved and added to the training corpus under the corresponding concepts. It is obvious that the threshold plays a crucial role in this model. On one hand, it should be noticed that classifier performances can be degraded if the automatically labeled samples are too erroneous. If the threshold is too small, more shots will be kept to the labeled set, and the number of shots labeled incorrectly likely increases, which will contaminate the training set with potentially noisy data and directly bring down the performance of concept detector in the subsequent training process. On the other hand, a high value threshold will lead to fewer shots reserved and the concept detectors' performance will not improved much. Figure 2 shows the relationship between the number of positively retrieved shots and the selection threshold.

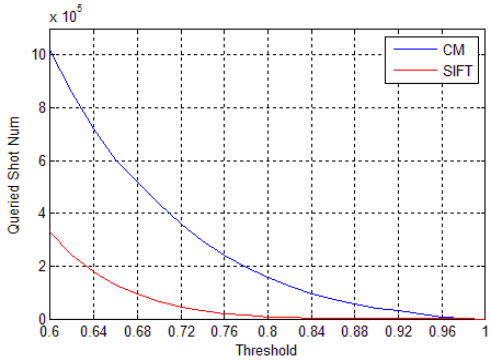

Figure 2. Number of retrieved shots with respect to the selection threshold

D. Evaluation

For the purpose of evaluating the annotation accuracy improvements, we train another groups of concept detectors with the union of labeled data and reserved data, and run this groups of concept detector on the same testing data. Here, we report the results concerning annotation accuracy after a unique database enhancement / concept detector re-training. Under normal circumstances, we foresee such feedback process to run continuously in order to adapt constantly the concept models.

In this experiment, the average precision (AP) and mean average precision (MAP) are used as performance measures. AP is a standard performance measure for image and video semantic concept detection. And MAP is the arithmetic mean of average precision values across all of the concepts.

E. Results

A well chosen probability threshold for selection novel training set exemplars plays an important role in the

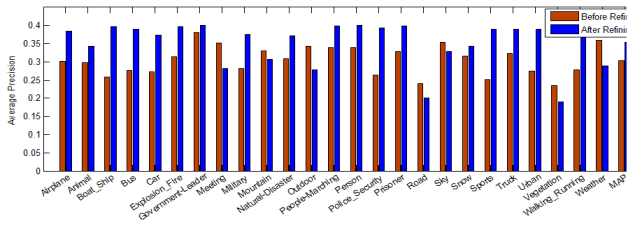

Figure 3 The AP of concept detectors with CM feature

experiments. In order to compare the performance of our model with CM and SIFT feature, we select the top 15% unlabeled data according the probability output under concept detectors for both groups of the data. In details, for CM feature, the threshold of 0.92 leads in 29,510 shots annotated with 75K labels, corresponding to about 3 labels for each shot on average. And for SIFT feature, 9,623 shots are annotated with about 24K labels under the threshold 0.80.

Figure 3 gives the detector performance based on the CM feature before refinement and after refinement. From the figure, we can see that for most of concept, the detection accuracy is improved when new shots are added automatically through our feedback mechanism. Significant AP gains are achieved for "Sports" by 55.00%, " Boat_Ship" by 53.55%, "Police_Security" by 49.40%. The overall MAP is improved by 16.68% after a single iteration. The result of SIFT feature refinement shows on Figure 4. Similar with CM feature, most of the concept detectors' performance is also improved. With an overall MAP improved by 10.56%, some concept detectors also gained significant advance, such as "Sky" by 61.3%, "Outdoor" by 46.2%, and "Prisoner" by 35.8 %. Those result further support that concept detectors can be improved with the unlabeled data. To our best knowledge, it is the first time to employ a semi-supervised learning strategy on such a huge web-scale video dataset.

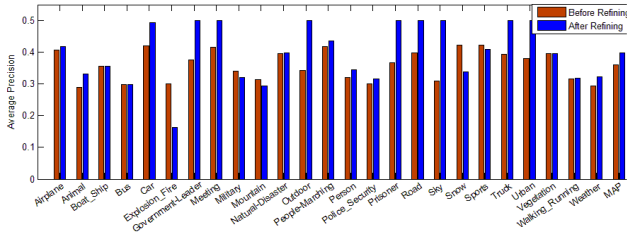

Figure 4 The AP of concept detectors with CM feature

However we also should notice the performances degradation on some classifiers. For some concept, such as "Meeting", "Weather", "Vegetation" in Figure 3 the detectors performances are regressing when new shots are added to the training set (25.48%, 23.96%, 23.59% respectively). The same phenomenon can be observed in Figure 4, such as "Explosion_Fire" by 45.6%, "Snow" by 20.2% and "Mountain" by 6.8 %. This indicates that the original model was not sufficiently robust and too many error samples are added in the training dataset. This issue could easily be taken care of by removing those training added exemplars before updating the selection probability threshold and eventually selecting new shot candidates. This is part of the work that we are currently pursuing to improve our approach.

V. CONCLUSION

In this paper, we have introduced a large scale video corpus, providing about 42,000 videos along with both low-level visual features, including an annotated subset of about 6,000 videos for semantic video analysis. We have developed an annotation application based on this dataset, with the aim of investigating the possibility of generating new training samples for improving and refining concept detectors accuracy without extra human cost. The evaluation of our proposed approach shows that when the training set is extended using automatically annotated video shots, the concept detection accuracy (average precision) can be improved by as much as 16% on average.

We are currently extending the approach reported here in a number of ways. Firstly, by employing the textual information associated with the video to further refine the concept models. Secondly, by employing all the low-level features extracted from the video shots for concept detection. We are also investigating the effect of performing the automatic training set enhancements and its corresponding concept detector re-training multiple times on video shot annotation accuracy.

REFERENCES

[1] C.Cortes and V.Vapnik. Support vector networks. Machine Learning, 20:273--297, 1995.

[2] Min-Ling Zhang, Zhi-Hua Zhou. A k-Nearest Neighbor Based Algorithm for Multi-label Classification. IEEE International Conference on Granular Computing, (2005), pp. 718-721 Vol. 2.

[3] Zheng-Jun Zha, Tao Mei et al. Graph-Based Semi-Supervised Learning with Multi-Label. IEEE Conference on ICME, pp. 1321-1324, Hannover, Germany, June 23-26, 2008.

[4] Xian-Sheng Hua and Guo-Jun Qi. Online multi-label active annotation: towards large-scale content-based video search. Proceeding of the 16th ACM international conference on Multimedia , Canada.2008.

[5] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. ACM International Conference on Image and Video Retrieval. Greece. Jul. 8-10, 2009.

[6] J. Cao, Y.D. Zhang, Y.C. Song, Z.N. Chen, X. Zhang, and J.T. Li. MCG-WEBV: A Benchmark Dataset for Web Video Analysis. Technical Report, Institute of Computing Technology, May. 2009.

[7] A. Ulges et al. Learning automatic concept detectors from online video. Computer vision and Image understanding. 2009.

[8] Lei Wu, Linjun Yang, Xian-Sheng Hua, Nenghai Yu. Learning to Tag. In Proceedings of the 18th international conference on World Wide Web.2009.

[9] Tang Wang; Tao Mei et al, "Video Collage: A Novel Presentation of Video Sequence," In Proceedings of ICME, Beijing, 2007.

[10] Amir, et al. IBM Research TRECVID-2004 Video Retrieval System. In NIST TRECVID Workshop, Gaithersburg, MD, 2004.

[11] C.C. Chang and C.J. Lin. LIBSVM: a Library for Support Vector Machines. 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[12] J. Zhu. Semi-supervised learning literature survey. Computer Sciences Technical Report TR 1530, University of Wisconsin-Madison. 2005.