Concept Detector Refinement on Social Videos

Xueliang Liu
EURECOM Institute,France
xueliang.liu@eurecom.fr

Benoit Huet
EURECOM Institute,France
benoit.huet@eurecom.fr

ABSTRACT

The explosion of the social video sharing sites gives new challenges on video search and indexing technique. Because of the concept diversity in social videos, it is very hard to build a well annotated dataset that provides good coverage over the whole meaning of concepts. However, the prosperity of social video also make it easy to obtain a huge number of videos, which gives an opportunity to mine the semantic content from an infinite amount of video entities. In this paper, we focus on improving the performance concept detectors and propose a refinement framework based on semi-supervised learning technique. In our framework, the self-training algorithm is employed to expand the training dataset with automatically labeled data. The contribution of this paper is to demonstrate how to utilize the visual feature and text metadata to enhance the performance of concept classifier with a lot number of unlabeled videos. By experiment on a social video dataset with 21,000 entities, it is shown that after expanding the training set with automatic labeled shots, the concept detectors' performance is significantly improved.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Measure,Performance,Experimentation

Keywords

Social Video, Semantic Analysis, Semi-supervised Learning

1. INTRODUCTION

With the advances of digital capture equipment and multimedia storage, the recent explosion in video shared web technologies make it possible to upload the videos by the web users. Last five years has witnessed rapidly growing popularity of social video sites, such as YouTube[4], DailyMotion[1]. It results in the explosion of social video documents, with diverse concept and sparse text description.

The prosperity of social video gives new challenge on the traditional text-based search engine, though they have gain a remarkable success on text retrieval on the internet. Nowadays video search engine still adopt standard text retrieval technologies to index and search social videos according the accompanied metadata, such as tags, description, comments. This is a efficient way for video query but can not index video data semantically unless video data are well annotated by hand. Obviously it is becoming an emergency to develop video search techniques that can mine the semantic concept and no need manual labels.

On the other side, semantic video analysis techniques based on statistical models have make great progress in recent years. These research are currently focused on the analysis and mining the visual content of video by modeling the low level features extracted from video shots. These learning based techniques succeed in traditional video but face new challenges on dealing with social video. Due to the infinite amount of video with diverse concept, it is very hard to build a well labeled dataset with a good semantic concept coverage for training.

Both of the text-based technique and visual content modeling approach have their strengths and limitations on video indexing. It gives potentiality to fuse this two kinds of features together for better semantic analysis. In this paper, we try to integrate the text and visual feature of social video entities to improve the performance of concepts detection, and propose a semi-supervised learning based framework to obtain a group of concept detectors with better coverage by exploitation of unlabeled data.

The organization of the rest of this paper is as followed. In Section 2, we provide a brief review on the related work. In Section 3, we introduce our refinement framework based on visual feature and tags. In Section 4, we demonstrate the experiments and show the results. Finally, we conclude the paper in Section 5.

2. RELATED WORK

In the recent years, there already have been lots of studies on video concept analysis with thrust of video storage and machine learning techniques. To investigate the problem, the multimedia community has build many benchmark dataset, such as TrecVID[3], Caltech[10]. Semantic video analysis has traditionally involved these known datasets with

ACM MM '10 Firenze, Italy

fixed and limited sets of keywords and semantic concepts. Based on those well annotated datasets, many related framework are proposed with the advance of machine learning algorithm. A straightforward way of video analysis is to adapt the image search and indexing techniques directly, which utilize the visual feature obtained from shots keyframe to model the concept underlying in the video contents. For example, the authors of [9] have modeled image keywords using a multiple Bernoulli distribution for image annotation. To apply their method on videos they simply build their model for visual features within rectangular regions to the keyframes of a video and achieve better results. Other than adapted image indexing techniques, there are some works much more specifically designed for videos, which focus on mining on audio and spatio-temporal aspect of video. Motivated by the success of SIFT feature in image indexing, some similar work to represent shots with a spatio-temporal feature have been done[14, 13]. The authors of [14] proposed a local space-time features to capture local events . This technique have been shown effectiveness on people event recognition, such as people running or jumping. And in [13], the authors studied a shape-based feature for event recognition in crowded videos. To exploit the power of audio feature for video concept detection, Jiang et. al [12] investigated the joint audio and visual analysis at semantic concept detection, and propose a novel visual feature with background audio representation to improve concept detection.

It has become possible to attempt multimedia search and indexing on a large size of dataset recently, with the prospect of social share website such as YouTube [4], Flickr [2], where a large number of videos are available benefited from the contribution of web users. However, working on real world web datasets provides new challenge on the traditional video concept detection techniques. Clearly it is very time-consuming and tedious work to build a well labeled dataset for the training purpose. To address the problem of semantic web video analysis, some large size of dataset with the multimedia data crawler from shared portals have been built [8, 6]. Beside those web video datasets built very recently, a number of research works in the image domain have shown acceptable results by investigation on semi-supervised learning techniques[16, 11]. A similar work as we propose in this paper, the authors of [15] developed an active learning based concept classifier refinement system on a large scale of image dataset, the concept classifiers can be reinforced by updating the positive and negative training samples iteratively, but it still need users to review the output of the classifier. In this paper, we report an automatic framework based on semi-supervised learning for concept refinement and expect a similar result on social videos.

3. OUR PROPOSED SCHEME

Visual features are popularly used in previous research [9, 14, 13]. They are believed as the representative feature to video content. In social video shared site, the videos entities are uploaded always along with some text metadata. Though these metadata are labeled manually by the video user arbitrarily, and are not so accurate because of noise among them, these text also give a rough indication to the real concept behind the video shots. so they are of importance for video content representation. The proper combination of textual and visual feature can boost the performance of video analysis technique greatly. In this paper, we take a web video entity as composition of a group of shots along with a tag set, and focus on refine the concept detectors by exploration into unlabeled pools based on semi-supervised self-training approach.

3.1 Refinement with Self-training Strategy

At first let us introduce the general semantic video concept analysis problem briefly. It is the process by which a computer system automatically assigns caption or keywords to a digital video shot, which can be often regarded as a classification problem. Suppose we have a well annotated dataset $X = \{x_i\}$ along with its label $Y = \{y_i\}$. Our goal is to find out a group of classifiers

$$\mathcal{F} = \{f_i | f_i : X \rightarrow Y\} \tag{1}$$

Their parameters $\{\lambda_i\}$ can be obtained from

$$\{\lambda_i\} = \arg\max_{\lambda_i} P(Y|X, \{\lambda_i\}) \tag{2}$$

For the video annotation problems, the dataset X can be the shot, which is represented by the visual feature vectors, and the label Y is the concept to be annotated. Associated with each concept, there is a model f_i to compute the probability that a shot belongs to this concept with. In order to obtain the predict model, lots of data should be labeled well for the training process. However, it is unaffordable to annotate a large-scale video corpus with a good coverage over the whole meaning of concepts because of the work intensity and time consumption. On another side, it becomes easier to obtain a huge number of unlabeled data with the booming of video share website. This make it possible to capture more underlying meaning of the concepts. There are also related some work done with a semi-supervised learning framework to mine semantic meaning among data pool without labels [17]. Semi-supervised learning is a group of algorithm that make use of the labeled and unlabeled data. The one we used in this work is called self-training. Besides the labeled dataset X_1^l, supported we have other unlabeled dataset U_{l+1}^N and $l << N$. In self-training, the classifiers are firstly trained from the small amount of labeled data X as shown in Equation 1, then used to predict label for the unlabeled data, and the most confident unlabeled data are added to the training set.

$$X^* = \{x | \max_i (P(x|f_i)) > \theta, x \in U\} \tag{3}$$

And

$$X^{'} = X + X^* \tag{4}$$

The classifiers are then re-trained on the extended training set $X^{'}$ with the same method as Equation 2.

$$\{\lambda_i^{'}\} = \arg\max_{\lambda_i^{'}} P(Y, Y^*|X^{'}, \{\lambda_i^{'}\}) \tag{5}$$

Where $Y^{'}$ is the label of X^* predicted by model $\mathcal{F}$, and $\lambda_i^{'}$ is the parameters of updated model $\mathcal{F}^{'}$.

For social video, the useful information we can explore are visual features and textual metadata, we consider them for our refinement as follows.

3.2 Visual Feature Based Refinement

In most of previous annotation research, visual features are used to represent the content of video shot. The intuitional approach of refinement is to utilize the visual feature directly. As shown in Figure 1, we first initiate the training of the concept detectors with the previously labeled subset. Then those newly trained detectors are run on the unlabeled video collection to predict labels. The videos shots that are with a high similarity to a concept, in other words, when the probability estimation of the concept detector is above a given threshold, will be added to the training set. The concept detectors are then re-trained on the automatically extended training set. Both the original concept detector and the re-trained ones are then evaluated on the testing dataset which has been held out for performance evaluation only.

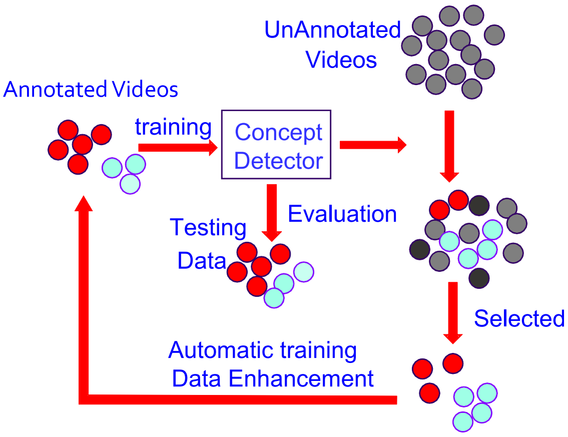

Figure 1: Visual Feature Refinement

3.3 Tags-Based Visual Supervision Refinement

Compared with traditional videos, social videos are commonly accompanied with metadata such as tags, description, script, etc..., which are uploaded by the users themselves. Though the textual information is often erroneous and sparse, and not accurate enough to provide the required knowledge for effective content-based retrieval, the analysis of the auxiliary text shows a potential in improving the performance of traditional multimedia information analysis approaches.

Before we utilize the text metadata for semantic analysis, there are some problems that need to be dealt with. In video annotation, the concepts should be labeled on each shot. However, the web videos tags are given to the whole video entity, so we can not use the tags as a kind of weak label directly. Additionally, the synonymy and polysemy problems make it more complicated. Synonymy is used to describe the fact that there are many ways to refer to the same object and polysemy means that most words have more than one distinct meaning. Considering the synonymy and polysemy, it is hard to mine the meaning from so brief and sparse text description. For example, the video shots tagged with "boat" and "ship" should be annotated with the same concept, and if a video shot is tagged as "Apple", the user's intention may be "a kind of fruit", "A kind of electrical device", or "a famous company". To solve the synonymy, we expand the concept with keywords that have similar meaning in semantic level. And for polysemy, it should be noticed that for each word, the different meaning has a different visual appearance, while there will be a special distribution in feature space.

With this in mind, we propose our tag-based visual supervision (TBVS) refinement framework as shown in Figure 2. We query with keywords for each concept from our dataset, and initialize the annotation of all the *Shots* with such concept for each returned video entity. Then we use the same strategy as Section 3.2. A group of trained visual concept detectors are run on those shots, and sort the result by visual similarity. Those whose probability are above a given threshold are reserved to the training set.

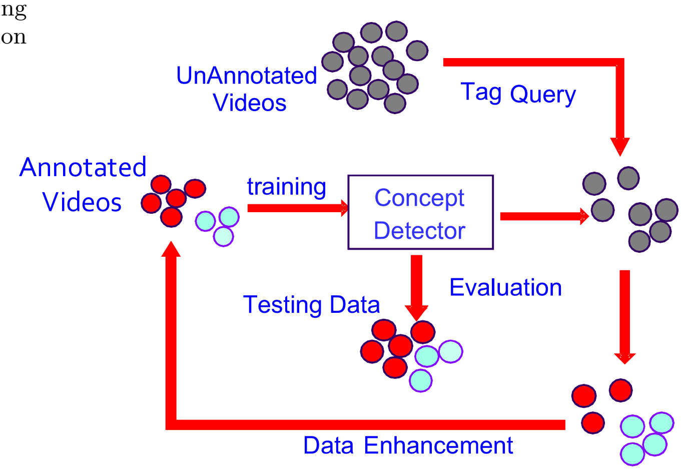

Figure 2: Tags-Based Visual Supervised Refinement

Though it seems that similar strategies are used in the two refinement methods, they are even too different. For visual feature refinement, only visual similarity are taken into consideration. However for the tag-based refinement, we filter the video shots from a word query result, it is more possible to obtain relevant shot under a semantic meaning.

In both refinement approaches, obviously the selection metric chosen is crucial. We attempt to reduce the impact of this issue by incorrectly labeled examples that are added to the training set, as incorrect samples that added in the training set will lead to a wrong result. This issue will be study further in Section 4.3.

4. EXPERIMENTS

To validate our proposed approach on social videos, we use a social video dataset and conduct a group of experiments. The dataset contains about 21,000 videos entities which are crawled from video share website Youtube[4]. Those videos are segments into 240,000 shots and a text corpus with 300 keywords are built from the video tags and titles. We study our approach on such a data in term of performance improvement.

4.1 Dataset

A well designed dataset is very important for our concept detector refinement problems. Here we use a subset of our dataset created before, which contains 42,000 videos and their associated metadata from YouTube [4]. Half of the whole data are used in the experiments, which means about 21,000 videos. All of the videos are segmented into shots and a keyframe is extracted for each shot. We also obtain multiple types of low level visual feature for the keyframe(64-D color histogram, 225-D color moment, 250-D Bag-Of-Words). For simplification, only the 225-D color moment feature, which has have been shown efficient and effective in generic concept detection [5], is used in this experiment.

Besides the visual feature, we also build a keywords dictionary from the text metadata along the videos. From video title and tags we obtain 562K textual words. We sort them by frequency after removing the stop-word and words stemming. We also remove some meaningless words such “video”, “music” manually and reserve the top 300 words as our keywords corpus.

In this experiments, we manually choose five visual concept: *Airplane*, *Animal*, *Boat_Ship*, *Person*, *Snow*, which have recognizable appearance and good distribution in our dataset. We expand those concept semantically with the keywords in our corpus for synonymy as shown in Table 1.

Table 1: Concept Expanding

Concept	Keywords
Airplane	Airplane, Flight
Animal	Animal, Dog, Tiger,Lion
Boat_Ship	Boat, Ship
Person	Person, People, Girl, Boy
Snow	Snow

4.2 Learning Process and Evaluation

The self-learning is a practical wrapper approach, there still need a baseline machine learning algorithm. Here *support vector machine* (SVM) is employed for the learning process. SVM is a effective method to solve binary-class or multi-class classification problems. A classification problem is considered on a given a set of labeled training data $(\vec{x_i}, y_i)$ where samples $\vec{x_i} \in R^d$ and binary labels are given as $y_i \in \{1,-1\}$ for binary-class problems and $y_i \in Z$ for multi-class problems. In the case of multimedia information retrieval, we can consider R^d the d-dimensional space of low-level visual features so that each image or video has a unique feature vector descriptor. The labels y_i are used to indicate which concept examples are relevant with. The solution of SVM is to construct a hyperplane or set of hyperplane in a high feature dimensional space, which can be used for classification, as well as regression or other tasks. The hyperplanes have the largest distance to the nearest training data points of any class and make the classification error of the classifier to be lower. The implementation used here is the latest LIBSVM [7] with a *Radial Basis Function* (RBF) kernel. We use the cross-validation methods to determinate the parameters in the SVM models.

In this experiment, the *average precision* (AP) and *mean average precision* (MAP) are used as criteria to measure the performance. AP is a standard performance measure for image and video semantic concept search and indexing. It is almost the same as the area under the un-interpolated precision-recall curve. And MAP is the arithmetic mean of average precision values across all of the concepts.

4.3 Parameters Setting

The key issue in self-learning is how to find the proper metric to decide which examples to add to the training set. In our refinement process, we use a threshold to decide the amount of new adding shots for simplification. It is obvious that the threshold plays a crucial role in this model. On one hand,a high value threshold will lead to fewer shots reserved and the training set are still far to reach a good coverage in the feature space, which will lead to the new trained concept detectors' performance will not improved much. On the other hand, it should be noticed that classifier performances can be degraded if there are many incorrectly labeled sample in the new training set. If the threshold is too small, more shots will be inserted to the training set, and the number of shots labeled incorrectly will increase, This will contaminate the training set with potentially noisy data and directly bring down the performance of concept detector in the subsequent training process. Figure 3 shows the percent of reserved shots in both refinements strategy. From this Figure, we can see that with the increasing of threshold, the percent of expanding shots decreases. And for the two process, we use different threshold because of two reasons. First, we know that there are some kind of underlying truth in tags labeled by users, so we can forecast that in the tag-based refinement it is no need to use a threshold with the same value as visual feature refinement. Secondly, because of the sparsity of keywords in text metadata, there is no large amount of shots return queried by keywords. So a high probability threshold value will block too many shots into the refinement process. In our experiment, we find the optimal threshold of 0.92 for visual feature based refinement and 0.72 for tag based refinement can achieve a better result.

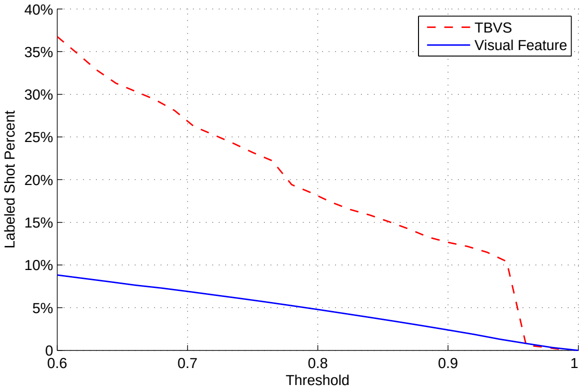

Figure 3: retrieved shots with respect to selection threshold

4.4 Results

To validate our methods, a group of experiments are done in our dataset: a) training with annotated shots; b) training with on all shots from tag query result; c) refinement based on visual feature; d) refinement based on tags query and visual supervision. All of the detectors are tested on the same data that are labeled well.

Figure 4 gives the detector performance measure result on the experiments. From the figure, we can see that training on the data queried by tags gain the worst performance in all of the concept as well as the mean measure, as we expect, because of the noise among user tags. Compared with the performance of classifiers trained on labeled data, both the visual feature refinement and tag-based visual supervised refinement achieve better results. In visual feature refinement, the detection accuracy is improved when new shots are added automatically through self-learning scheme for most of concept. Significant AP gains are achieved for "Boat-Ship" by 43.1%, "Person" by 28.5%, "Airplane" by 19.0%. The overall MAP is improved by 21.7% after a single iteration.

Figure 4 also show the remarkable improvement on tag-based visual supervised refinement. Similar with visual feature refinement, this group of concept detectors is also enhanced by coming of the new shots. With an overall MAP improved by 23.5%, concept detectors also gained significant advance, such as "Boat-Ship" by 53.5%, "Airplane" by 27.7%, and "Person" by 17.7% respectively.

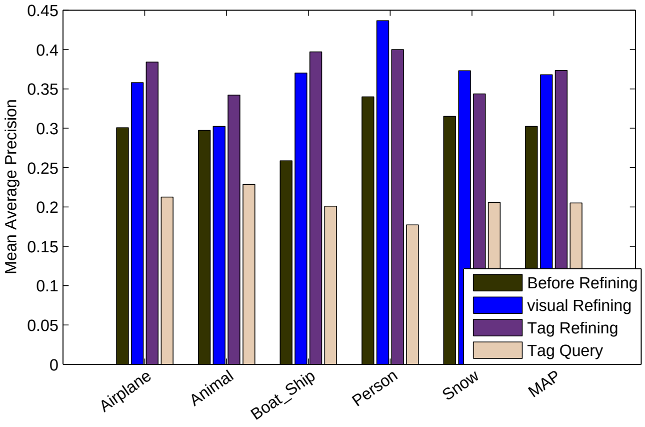

Figure 4: The AP of concept detectors

From the result, we also can observe that the concept coverage in this dataset can be reinforced by the automatic annotated data, and concept detector's performance gain a remarkable improvement with the two refinements process. Other things we obtain from the results is that though the video tags are sparse and erroneous for a single video, there indeed are semantic truth in a groups of shots. As shown in Figure 5, the training data expanding with less shots and low metric leads to almost the same refinement results.

5. CONCLUSIONS

As the amount of social videos content continues to enlarge, there is an immediate need for automatic tools for semantic video search and indexing. However it is very hard to provide a well labeled datasets for learning purpose because of the infinite and diversity of the concepts. In this paper we focus on improving the performance of concept detection with semi-supervised learning. A refinement framework that utilize the visual feature and text meta data.

In the future, we will consider how to build concept detectors with a weak effect of well labeled set. We know that in the refinement framework proposed in the paper, a pre-trained model is necessary to filter the unrelevant noise. but it is not easy to obtain in real world. So how to obtain a group of concepts automatically and analyze them semantically is our future work.

6. REFERENCES

[1] *DailyMotion*. http://www.dailymotion.fr.

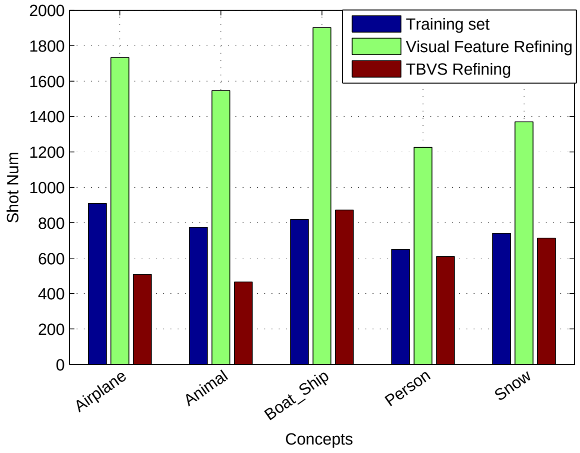

Figure 5: Shot Number for training and refinement

[2] *Flickr*. http://www.flickr.com/.

[3] *TrecVID*. http://trecvid.nist.gov/.

[4] *YouTube*. http://www.youtube.com.

[5] A. J. O. Argill, M. Berg, and et al. IBM research TRECVID-2004 video retrieval system. In *NIST TRECVID 2004 Workshop*, Gaithersburg, MD, 2004.

[6] J. Cao, Y. Zhang, Y. Song, Z. Chen, X. Zhang, and J. Li. MCG-WEBV: A benchmark dataset for web video analysis. Technical report, May. 2009.

[7] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[8] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. NUS-WIDE: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval*, Santorini, Greece.

[9] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1002–1009, 2004.

[10] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical report, California Institute of Technology, 2007.

[11] R. Hong, G. Li, L. Nie, J. Tang, and T.-S. Chua. Explore large scale data for multimedia QA. In *ACM conference on Image and Video Retrieval*, Xi'an, China, 2010.

[12] W. Jiang, C. Cotton, S.-F. Chang, D. Ellis, and A. C. Loui. Short-term audio-visual atoms for generic video concept classification. In *Proceeding of ACM international conference on Multimedia (ACM MM)*, October 2009.

[13] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *International Conference on Computer Vision*.

[14] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *International proceeding of Pattern Recognition*, 2004.

[15] M. yu Chen, M. Christel, E. Hauptmann, and

H. Wactlar. Putting active learning into multimedia applications: Dynamic definition and refinement of concept classifiers. In *Proceedings of ACM Multimedia*, pages 902–911. ACM Press, 2005.

[16] Z.-J. Zha, T. Mei, J. Wang, Z. Wang, and X.-S. Hua. Graph-based semi-supervised learning with multi-label. *ACM Trans. Program. Lang. Syst.*, 20(5):97–103, 2009.

[17] X. Zhu. Semi-supervised learning literature survey. Technical report, CMU, 2006.