Volume 2, No. 1, Jan-Feb 2011

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Active Learning Based Semantic Video Retrieval Using Single Graph

M.Ravinder*
Computer Science and Engineering
Sree Chaitanya College of Engineering
Karimnagar, India
ravinder_m.tech@yahoo.co.in

Dr.T.Venu Gopal
Computer Science and Engineering
JNTUCEH, Jagityal
Karimnagar, India
t_vgopal@rediffmail.com

***Abstract*—** A multimedia record has grown-up significantly over the last few years. Active learning and semi-supervised learning are significant machine learning techniques when labeled data is limited or costly to obtain. As an option of passively taking the training samples provided by the users, a model could be designed to actively seek the majority informative samples for training. We take up a graph based framework with semi-supervised learning method where each video shot is represented by a node in the graph and they are connected with edges weighted by their similarities. We apply active learning methods to select the most informative samples according to the graph structure and the current state of learning model.

***Keywords*-**video indexing ; retrieval ; learning algorithm; single graph

I. INTRODUCTION

Quantity of multimedia records has grown significantly over the past few years. With this growth is the ever-increasing want to effectively represent, arrange and retrieve this huge pool of multimedia stuffing, mainly for videos. Even though a lot of efforts have been dedicated to developing efficient video content retrieval systems, most current profitable video search systems, such as YouTube, still use standard text retrieval methods with the help of text tags for indexing and retrieval of videos [1]. A fundamental difference between video retrieval and text retrieval is that text representation is directly connected to human interpretations and there is no gap between the semantic meaning and representation of text. When a user search for the word"earth" in a collection of text documents, documents containing the word might be identified and returned to the user. But, when a user searches for"earth" in videos, it is not clear how to decide whether a video contains earth.

A video consists of visual features, text features and motion features. Visual features are extracted from key frames of a video shot the most common visual features that can be extracted include moments, color histogram, color coherence vector, color correlogram, edge histogram, and texture information [2].

Text features play a very important role in video retrieval, especially for news video retrieval [3]. Motion features are especially useful for queries about identify an action or a moving object, for example, identify fight scenes in a video, or look for shots with a train leaving the platform. There are statistical motion features and object-based motion features [4].

In this paper we implement active learning algorithm with single graph for semantic video retrieval. Section II discusses about active learning, section III talks about active learning with single graph, section IV focuses on experiments and results, section V ends up with conclusion.

II. ACTIVE LEARNING

During video retrieval, labeled video data are very limited for the reason that obtaining labels for video shots is an error-prone and costly task. Active learning approach can decrease users labeling effort by selecting only the most"informative" samples for the present learning models. Figure 2.1 shows the outline of an interactive video retrieval system with active learning.

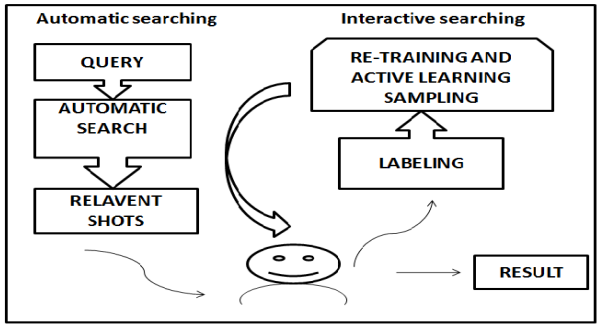


Figure 2.1 Interactive video searching using active learning

Most active learning methods focus on how to choose the most informative samples for a classification model and very few aims to select the most informative sample for ranking scenario [1].

A. Active Learning Strategies

The majority of the active learning algorithms fall into the group of active learning strategy with uncertainty, active learning strategy with error minimization, and hybrid active learning which the combination of both strategies.

B. Active learning algorithm

An active learning algorithm dynamically decides which are the most helpful data for the current model as well as asks the user to label those data as relevant or non-relevant. In order to choose the useful samples, an active learning strategy needs to sample training data according to the

current state of the model as well as the structure of the graph. The general active learning algorithm works in the following manner as shown in Fig 2.2:

(1) Initialize knowledge representation with initial excercise samples
(2) Choose samples from unlabled information by means of an active learning approch
(3) Request user to tag those samples and add them to the guidence dataset
(4) Revise model by means of the new guidance data
(5) Reiterate 2 to 4 for n iterations
(6) Yield final ranking list

Figure 2.2 Algorithm for active learning

III. Active Learning With Single Graph

Within graph based methods we first build a graph with nodes and edges. The nodes are the samples and the edges represent the similarity between those samples [5]. This graph captures the global structure of the data. Once the label of some data is recognized, it will be propagated next to the edges to other data points. [6] Proposed a method based on Gaussian random field and harmonic function. They formulated the learning problem as Gaussian random field over a relaxed continuous state space. They have carried out experiments on digit and text classification tasks. In [7] where active learning was combined with Gaussian random field and harmonic energy minimization.

During graph-based learning, the uncertainty of a sample in the graph is very much associated to the global and local structure of a graph. Consider the case when the graph is not connected and has several connected components. This is a practical situation in ϵ-similarity graph where only nodes with similarities $> \epsilon$ are connected with an edge. Nodes within a connected component are alike to each other while nodes belonging to different connected components are less alike. An insightful idea is to select at least one sample from each connected component. Since if a connected component has no labeled node in it, no matter how we propagate scores in network, those nodes in this isolated components cannot receive any information. For learning initialization, we firstly recognize all the connected components in a graph and then begin the graph-based learning by sampling single node from each component. A reasonable method to sample within every component is to choose number of samples proportional to the size of the connected component. Uniform sampling on the graph possibly will be a good approximation to the sampling approach described above. We consider active learning following initialization. Since the scores are propagated down the edges and the score of every node is obtained from the scores of its neighbors, nodes that are the furthest from labeled nodes are the most uncertain. Nodes that have a lot of links with unlabeled nodes are more informative for graph-based learning model. We define two degrees for a node, $\text{degree}_L(a)$ and $\text{degree}_U(a)$

$$\text{degree}_L(a) = |\{x_b \,|W_{ab} > 0, x_b \in L\}| \quad (1)$$

$$\text{degree}_U(a) = |\{x_b \,|W_{ab} > 0, x_b \in U\}| \quad (2)$$

An uncertainty based active learning selection principle that selects nodes with a small degree with labeled nodes but large degree with unlabeled nodes

$$\text{ActiveLearning}_{\text{uncertainty}}(a) = \frac{\text{degree}_U(a)}{\text{degree}_L(a)} \quad (3)$$

Then the active learning strategy greedily selects the top *n* nodes with the highest value of $\text{ActiveLearning}_{\text{uncertainty}}(a)$.

A. Construction of Graph

Graph construction is a very essential step in graph-based method and we should include prior domain knowledge about the data. There are a variety of methods for constructing a graph. We can construct a completely connected graph with an edge between every pair of nodes. This graph will take up O(n2) memory, which is not practical for huge video dataset. It has been shown empirically that a complete graph performs worse than sparse graphs [9].

For TREC video dataset, because of the scale of dataset, we construct k-nearest neighbor graph so that we can manage the sparseness of the graph simply. Within k-nearest neighbor graph a node is only connected to its k nearest neighbors. One particular feature of video data is the temporal relation between the shots, i.e. if a certain scene appears in single shot, it is highly possible that related scenes appear in close to shots within the same video. In order to include the temporal relation of video shots, we reinforce the graph so that all shots inside the same video are connected, i.e. the sub-graph on each video is a completely connected graph.

B. Features of Video Data

We bring in a few of the extensively used visual feature descriptors in content based video retrieval. Features can be extracted from video key frames. We also make use of text feature which is extracted from the audio track of a video as well as high level concept that serve to bridge the semantic gap between the low level visual features and high level semantic meanings.

[a] Movement of color :

Movements of color have been verified to be an effective way in representing color distributions of a videoframe. The first three moments of each color component are defined as:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} P_i \quad (1)$$

$$\sigma = \left(\frac{1}{N}\sum_{i=1}^{N}(P_i - \mu)^2\right)^{1/2} \quad (2)$$

$$s = \left(\frac{1}{N}\sum_{i=1}^{N}(P_i - \mu)^3\right)^{1/3} \quad (3)$$

where P_i is the value of the color component at pixel i and N is the entire number of pixels. Movement of color is a very compact global representation of a video frame. In order to improve its discrimination power, we divide each video frame into five × five grids and we compute the first three color moments for every one of the three components.

[b] Histogram for Edge Information:

We first quantize the edge information of a video frame in to a number of bins and the number of pixels drop into each bin is calculated. Every video frame is divided into five regions, with four regions uniformly partitioning the video frame and one region in the middle that over laps with the other four regions. After that edge information can be obtained as a result of applying various edge detection filters along different directions. For every pixel, the direction with the largest magnitude is set as the direction of the pixel.

[c] Wavelet of Texture information:

Texture is about continual pattern in a video frame. There are two broad types of texture representation methods: structural and statistical [10]. Structural methods are more effective with very regular patterns while statistical methods, including Tamura feature, multi-resolution filtering techniques, etc, use statistical information of the density of a video frame to characterize texture. Each video frame is first transformed into gray scale and divided into three×three grids. Then we perform three levels of Haar Wavelet transform to each of the grid.

[d] High level Features:

For bridging the semantic gap between low level visual features, such as color, edge and texture, and high level semantics, one approach is to make use of a set of intermediate semantic concepts identified as high level features that can be used to describe repeated visual and audio content entities in video collections. High level features concepts include road, water, buildings, etc. Once defining the set of high level concepts, video can be annotated first to indicate the existence or nonexistence of those concepts in the video.

C. Proposed System:

Our system is composed of three main components : graph construction unit, graph based learning unit and active learning unit. After features have been extracted from video shots in the data set, the graph construction unit will build one graph for each feature. In single graph based learning, the graph-based learning unit will perform Gaussian random fields and harmonic functions learning [6] on the graph with the current training samples. And active learning unit selects training samples from all unlabeled data for the user to label according to different active learning strategies.

IV. EXPERIMENTS AND RESULTS

We utilize the dataset from TRECVID Retrieval Evaluation 2007 [8] for experiments. This is one of the biggest annotated video data set that has been extensively used in evaluating video retrieval systems performance. The data set includes 100 hours of a wide variety of video, together with educational, cultural, youth oriented programming, news periodical, chronological footage etc. Table I consists of the key statistics of the Trecvideo 2007 data set.

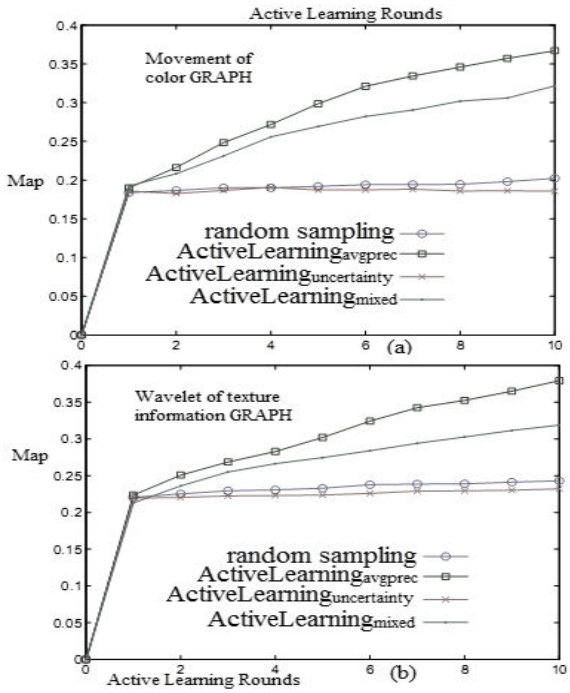
TABLE I. KEY STATISTICS OF TREC VIDEO 2007 DATA SET

Data span(hours)	100
Quantity of shots	18,142
Standard shot length	20sec
Standard number of shots per video	166
Quantity of video programs	109
Number of distinctive program titles	47

Figure 4.1 result of query, find shots with a big crowd of people

A few sample shots for the query "Find shots with a big crowd of people" is as shown in Figure 4.1. For TREC video dataset, for the reason that of the scale of dataset, we construct k-nearest neighbor graph so that we can manage the sparseness of the graph without difficulty. In k-nearest neighbor graph a node is only connected to its k-nearest neighbors. Within our experiments, we put k =30. One particular feature of video data is the temporal relation between the shots, In order to integrate the temporal relation of video shots, we reinforce the graph so that all shots inside the same video are connected.

We look at the effectiveness of different active learning strategies on single graph based learning. We contrast random sampling, uncertainty based active learning ($ActiveLearning_{uncertainty}$), and average precision based active learning ($ActiveLearning_{avgprec}$), which selects the top *n* samples based on the ranked list and we also use a mixed active learning strategy ($ActiveLearning_{mixed}$) that combines $ActiveLearning_{uncertainty}$ and $ActiveLearning_{avgprec}$ it selects top *n/2* samples from the ranked list and *n/2* uncertain samples. Figure 4.1 shows the results on TREC video data set.



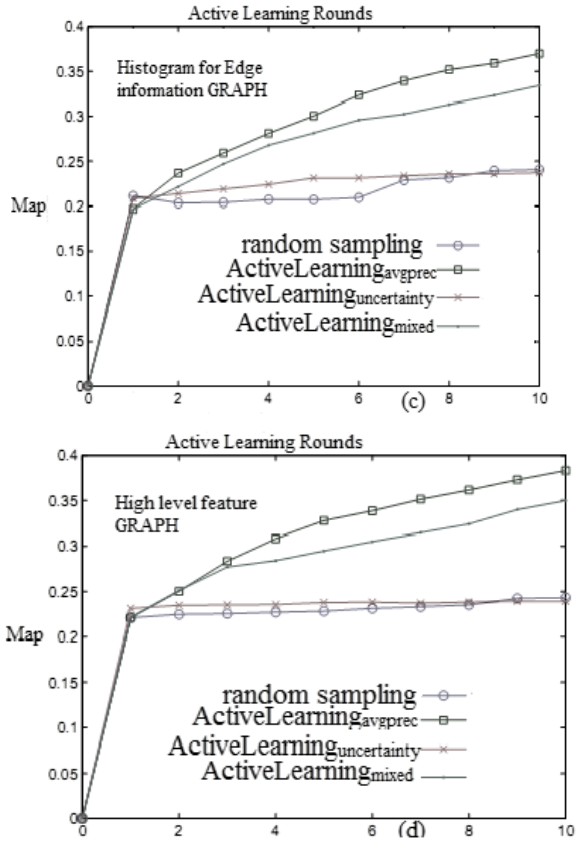


Figure 4.2 Active Learning - Single Graph TREC video

We look at the number of relevant training samples in every activelearninground. From the graphs shown in Figure 4.2 We examine that average precision based sampling selects more relevant training samples than random sampling. We examine that there is a very strong correlation between the number of related samples and the Average Precision performance.

V. CONCLUTION

In this paper we examined the performance of different features and active learning strategies on single graph based learning. We found that the effectiveness of the features depend on data. Video frames that contain distinctive characteristics are easy to retrieve. We have also demonstrated that Average Precision based active learning strategy performs well on real retrieval problems where relevant samples are unusual.

VI. REFERENCES

[1] T.S.Huang, C.K.Dagli, S.Rajaram, E.Y.Chang, M.I.Mandel, G.E.Poliner, and D.P.W.Ellis. "Active learning for interactive multimedia retrieval," In Proceedings of IEEE, 2008.

[2] Fuhui Long, Hongjiang Zhang, and David Dagan Feng. "Fundamentals of content based image retrieval,".

[3] Paul Over, Tzveta Ianeva, Wessel Kraaijz, and Alan F. Smeaton. Trecvid 2006 an overview. In MIR '07: "Proceedings of the 9th ACM International Workshop on Multimedia Information Retrieval," NewYork, NY, USA, 2006. ACMPress.

[4] Chih-Wen Su, H.-Y.M.Liao, Hsiao-Rong Tyan, Chia-Wen Lin, Duan-Yu Chen, and Kuo- Chin Fan. "Motion flow-based video retrieva," IEEE Transactionson Multimedia, 2007.

[5] Xiaojin Zhu. "semi-supervised learning literature survey" 2007.

[6] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. "Semi-supervised learning using gaussian fields and harmonic functions," In In ICML, pages 912–919, 2003.

[7] Xiaojing Zhu, John Lafferty, and Zoubin Ghahramni. "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions," Pro-ceedings of the ICML-2003 Workshop on the continuum from labeled to unlabeled data,2003.

[8] Alan F. Smeaton, Paul Over, and Wessel Kraaij. "Valuation campaigns and trecvid," Proceeding of International workshop on multimedia information retrieval, 2007.

[9] Xiaojin Zhu. " Semi-Supervised Learning with Graphs," PhD thesis, Carnegie Mellon University, 2005.

[10] Fuhui Long, Hongjiang Zhang, and David Dagan Feng. "Fundamentals of content based image retrieval," 2003.

[11] Sheng Tang, Yong-Dong Zhang, Jin-TaoLi, Ming Li, Na Cai, Xu Zhang, Kun Tao, Li Tan, Shao- Xi Xu, and Yuan- Yuan Ran. "Trecvid 2007 high-level feature extraction by mcg- ict- cas," 2007.