Exploring Knowledge of Sub-domain in a Multi-resolution Bootstrapping Framework for Concept Detection in News Video

Gang Wang
Department of Computer Science
National University of Singapore117590
wanggang@comp.nus.edu.sg

Tat-Seng Chua
Department of Computer Science
National University of Singapore 117590
chuats@comp.nus.edu.sg

Ming Zhao
Google Inc.
Mountain View, CA, USA, 94043
zhaoming@zhaoming.name

ABSTRACT
In this paper, we present a model based on a multi-resolution, multi-source and multi-modal (M3) bootstrapping framework that exploits knowledge of sub-domains for concept detection in news video. Because the characteristics and distributions of data in different sub-domains are different, we model and analyze the video in each sub-domain separately using a transductive framework. Along with this framework, we propose a "pseudo-Vapnik combined error bound" to tackle the problem of imbalanced distribution of training data in certain segments of sub-domains. For effective fusion of multi-modal features, we utilize multi-resolution inference and constraints to permit evidences from different modal features to support each other. Finally, we employ a bootstrapping technique to leverage unlabeled data to boost the overall system performance. We test our framework by detecting semantic concepts in the TRECVID 2004 dataset. Experimental results demonstrate that our approach is effective.

Categories and Subject Descriptors
H.3.1 **[Information Storage and Retrieval]** Content Analysis and Indexing methods; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding-video analysis.

General Terms
Algorithms, Experimentation.

Keywords
Domain Knowledge, Unlabeled Data, Text Semantics, Multi-resolution analysis, Transductive Learning, Bootstrapping.

1. INTRODUCTION
In recent years, the volume of digital video collections has increased exponentially, following the wide availability of low-cost multimedia recording and storage devices. There is increasing demand for effective solutions to manage such large-scale video databases. Because it is more convenient for novice users to express their information needs through concepts, the ability to index video contents at the semantic level (in terms of concepts) has attracted a lot of research interest.

In TRECVID evaluations, one of the important video genres is news video. News video usually includes several sub-domains such as sports, finance, live reporting etc. Techniques to classify shots into sub-domains are relatively mature [2]. Thus, many sub-domain analysis techniques have been used in the query-class dependent retrieval [6, 22]. These techniques first classify each user's query into one of the predefined categories and then employ query-class dependent weights to fuse the multimodal features for retrieval. This suggests that multiple sub-domain analysis should be effective for news video retrieval. However, few works have been done in capturing concept occurrence distributions from different sub-domains in the concept detection task [1, 5, 9, 18]. Because each sub-domain data set may have their own characteristics, training a model by mixing data of different sub-domains may lose important information and degrade the performance of the systems. On the other hand, if we segment the training data into several smaller data sets in different sub-domains and use them separately, in some domains, we may not have sufficient positive training data. Related works of this problem is the cross-domain adaption problem as investigated in [24]. However, given an existing classifier, they [24] required sufficient amount of labeled examples in the new dataset to learn the "delta function" between the original and the adapted classifier. This may not be practical as in many cases, there may not be much labeled data. Thus, given a training set in news video, we face the first problem on how we can make use of the knowledge of sub-domains to tackle the problems of sparseness and uneven distributions of training data.

Multimedia video processing refers to the idea of integrating information of different modalities [17], such as the text from automatic speech recognition (ASR) and visual contents, to analyze and process the contents of video. Yang et al [23] proposed a text-based retrieval method with visual constraints in Person X detection. However, it only works in person related concepts. Snoek et al [18] identify two general fusion approaches: namely early fusion and late fusion. The early fusion scheme integrates unimodal features before learning the concepts. Although it yields a truly multimedia feature representation, it has problems of high dimensionality and it is difficult to combine features into a common representation at a single resolution. The late fusion scheme first reduces the unimodal features to separate concept scores, and then integrates these scores to learn the concepts. Such an approach can focus on the individual strength

MM'08, October 26-31, 2008, Vancouver, British Columbia, Canada.

of each modality. However, this brings a potential loss of correlation in mixed feature space. Thus both approaches have their own strengths and weaknesses, and their relative performance across multiple test corpuses are mixed. Thus a second challenging problem is how to let evidence from multi-modal features support each other in a generic concept detection framework.

Naphade and Smith [15] surveyed the state-of-the-art systems and found that most researchers adopted a supervised learning approach to detect the concepts. Such a type of learning requires the estimation of an unknown function for all possible input values. This implies the availability of good quality training data, which includes most typical types of the data available in the test set. If such a condition is not satisfied, the performance of such systems may degrade significantly. Yan and Naphade [21] proposed a multi-view semi-supervised cross-feature learning method. They first used the labeled training set to learn one classifier in each view. They then boost the views from each other by augmenting the training set with unlabeled test data on which the other views can make high-confidence predictions. However, Tian et al [19] pointed out that unlabeled data helps only if labeled and unlabeled data are from the same distribution in a semi-supervised learning framework. Otherwise, the unlabeled data may degrade the performance when it is added. Other researchers adopted another way to use unlabeled data. For example, Qi et al [16] proposed a transductive learning method to infer unlabeled test data by finding related labeled training data via a clustering method. However, there are at least two open problems. One is how to segment the clusters until their contents are as pure and as large as possible. A pure cluster is defined as one where the labels of training samples are mostly positive or negative such that the entire cluster including the test samples can be labeled accordingly. The other problem is how to analyze the unknown clusters, which are impure clusters or clusters that include only test samples. Thus, a third problem is when and how to encode and explore the knowledge from unlabeled data properly in the inference framework.

In this paper, we propose a sub-domain based multi-resolution bootstrapping framework to tackle the above three problems. To tackle the first problem, we separate the whole corpus into eight sub-domain data sets and develop a sub-domain adaptive transductive learning algorithm. For the second problem, we leverage a multi-resolution inference structure and constraints to permit evidence from different modal features to support each other. To tackle the third problem, we make use of unlabeled data by combining two learning methods. We first employ transductive learning to capture the distributions of training and test data simultaneously so that we have the knowledge to know when we can make an inference via training data. We then combine it with a bootstrapping technique to further process the test results with low confidence from transductive learning. We test our framework on concept detection task based on the TRECVID 2004 dataset. The test results demonstrate that our framework is superior to reported systems.

The rest of the paper is organized as follows: Section 2 describes the design consideration. Section 3 presents an overview of our framework. Sections 4 discusses in detail on our multi-resolution, multi-source, multi-modal transductive learning. Our multi-resolution bootstrapping inference strategy is covered in Section 5. The experimental test-bed and evaluation results are presented in Section 6. Finally, Section 7 concludes the paper.

2. Design Consideration

In this Section, we introduce the motivation of our M3 framework. We focus our discussions on three topics: multiple sub-domain analysis, multi-resolution and multimodal fusion, and transductive learning with bootstrapping.

2.1 Multiple sub-domain analysis

In concept detection, many researchers have developed mid-level detectors to supplement the low-level features such as color and texture. Chang, et al [3] believe that abstracting low-level features to mid-level allows for inclusion of different modalities without resulting in an excessively high dimensionality. Several researchers have reported good performance on a number of mid-level detectors in news video. For example, Chaisorn [2] reported high accuracy of over 90% for shot genre detectors, such as anchorperson, live reporting, commercial and finance etc. Because of good performance for such mid-level detectors, researchers [1, 5] have made use of them to improve the concept detection results. In fact, such mid-level detectors can also be used to segment the corpus into multiple sub-domains. However, few research efforts have been carried out to use concept distributions in these sub-domains for the concept detection task.

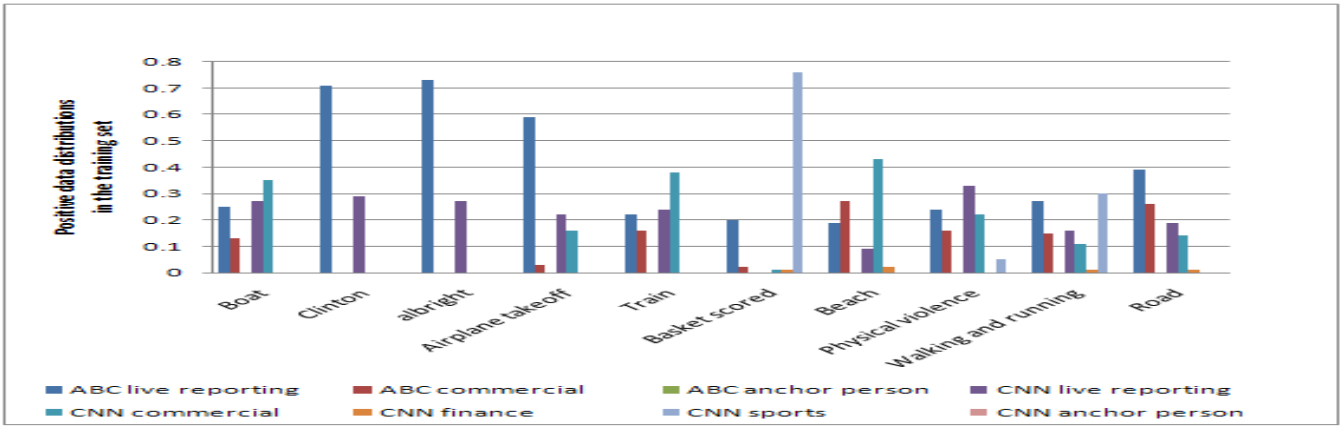

Figure 1: The distributions of positive data of the 10 concepts from TRECVID 2004 in the training set.

In our work, we segment the news video corpus into the categories of anchor person, sports, finance, commercial, with the rest placed under live-reporting. The reasons for the above choice are that the detectors for the first four categories are well defined [5], and the distributions of concepts in those 5 categories are distinctive. We used TRECVID 2004 as our test corpus, which has two series of news, ABC World News Tonight and CNN Headline News. Because the styles of these two sources of news are different, we segment them separately. This gives rise to eight sub-domains of: ABC live reporting, ABC commercial news, ABC anchorperson, CNN live reporting, CNN sports, CNN finance news, CNN anchorperson and CNN commercial news. From Figure 1, we can observe that the distributions of concepts in these sub-domains are very different. Thus, we should encode such distributions into the framework to improve the concept detection performance.

In addition, the characteristics of data from different domains may be different. From Figure 2, we find that shots sharing the same semantic concept for a product commercial usually have high similarity in both visual and text components. However, shots sharing the same semantic concept in live reporting may only share a few clue words in the ASR text, and tend to have large variations in the visual feature space. Thus, in order to capture the statistical pattern effectively, our analysis should base on distributions of multiple sub-domain data sets respectively.

well you'll need strength knowledge and experience on your sign that specific like helping you protect your family and plan to stick your retirement.

well you'll need strength knowledge and experience on your sign that specific like helping you protect your family and plan to stick your retirement.

(a) Two examples for concept "Boat/ship" in ABC commercial sub-domain.

She was the flagship of windjammer's fleet a 300 - foot reconditioned tall ship called the "fantome".

The ship had been held for five months in a Mexican port while authorities there tried to get the owners to pay their bills.

(b) Two examples for concept "Boat/ship" in ABC live reporting sub-domain

Figure 2: The characteristics of data from different domains may be different

2.2 Multi-resolution fusion

Currently, the state-of-the-arts systems fuse multi-modalities at a single resolution (mostly at the shot layer). As the shot boundary is designed to capture the changes of visual features, it is suited for visual analysis but fails to capture the text semantics well with breaks occurring often in the middle of a sentence. Yang et al [23] found that this is a common problem in news video analysis. In order to tackle the problem, we propose a multi-resolution fusion strategy. In our model, we define three resolution layers. They are the shot, multimedia discourse and story layer.

At the shot layer, we attempt to capture the semantics by finding similar images using color, texture and edge visual features. Such similar images are collected by an average-link clustering method. The choice of the clustering results is selected based on Vapnik Combined Bound [25] in a transductive inference framework. Figure 3 demonstrates the ability and limitation of visual feature analysis to cluster shots sharing the same concepts.

(a) Visual analysis is able to cluster two shots sharing the same concept "basket scored".

(b) Visual analysis fails to group two shots sharing the concept "boat".

Figure 3: Visual feature analysis at the shot layer.

Because of the limitation in discriminative power of visual analysis, it may cause many false alarms and misses. To overcome these limitations, we purify the clustering results and make further inference by text information. We perform text analysis at two layers. One is a multimedia discourse layer, which captures the synchronization between the text features at the sentence level and the visual features at the shot level. The MM discourse boundary occurs at the co-occurrence between the sentence and shot boundaries. In this work, we adopt the speaker change boundaries generated by the speech recognizer [8] as the pseudo sentence boundaries. This layer captures the semantics mainly by extracting a group of words from the enclosing ASR text. As it is insufficient to infer linguistic variations, we employ web-based knowledge (see Section 4.2) to bridge this gap in a transductive learning framework.

Carnival cruise lines new smoke free ship began sailing out of Miami this week.

(a) Concept "Boat/Ship" and word vector "ship, Miami..." at the MM discourse layer

Captain Guyan march first headed west and then north to Belize. All 97 passengers disembarked safely.

(b) Concept "Boat/Ship" and word vector "Captain, passengers... " at the MM discourse layer

Figure 4: Text analysis at the MM discourse layer

Figure 4 shows the ability and limitation of text feature analysis at the MM discourse layer. From the Figure, we can infer that the shot in 4(a) has high degree of relevance to the concept "Boat/Ships" based on the word vector at the MM discourse layer; but it is hard to make such a decision in 4(b). In order to tackle this problem, we incorporate text analysis at the story layer into the framework. There are many story segmentation methods for news video as surveyed in [4]. In this paper, we perform a simple story segmentation using the heuristics based on anchorperson, some logos, cue phrases and commercial tags [5]. At the story layer, we attempt to capture the semantic concepts by exploring the relationship between the concept and topics of a story. Here the topic refers to the main focus of a story. We employ the method developed in [14] to extract topics, which mainly depends on a set of high frequency ASR words in a story. Based on our topic extraction system, we could find the topic vector for the shots in Figure 4(b) is {ships, storm, hurricane, fantome}. According to such a topic analysis, we can conclude that the enclosed shots may have some degree of relevance to the concept "boat/ship".

2.3 Transductive and bootstrapping learning

The common problem of the current learning approaches is that the inference is based on "static" data, which comes in the form of training data. We assume that we have the ability to make inferences via the knowledge from training data alone. However, news video often contains new reports, and thus the domain has the inherent characteristic that there are always some differences between the training data and test data. Based on our analysis, there are at least two types of variations between the training and test data. One is called "gradual transition". For example, given two news reports -- one is about "September 11 event", and the other is about "The progress of NATO invading Afghanistan". If one is the training data and the other test data, we may have difficulties to assign semantic label "violence" to the test data based on training data. However, if we have documents about "September 11 event and al-Qaeda forces" and "NATO invaded Afghanistan to remove al-Qaeda forces", we may transfer the semantic label "violence" from training to test data via these linked documents. The other variation is called "mutation". For example, the concept "Clinton" may occur in the event of a middle-east peace talk. It can also appear in the event of the Lewinsky scandal. Again, it may be difficult to transfer the semantic label "Clinton" from one event to the other event, such as the case shown in Figure 5.

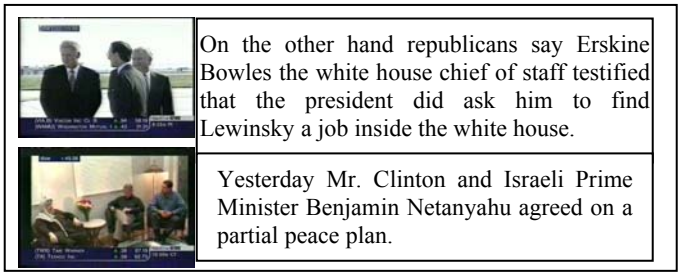

Figure 5: Different events include the same concept "Clinton" with very few non-stop words overlapping

In our framework, we propose the multi-source transductive learning under the bootstrapping framework. That is, we first employ transductive learning to capture the distributions of training and test data so that we have the knowledge to know when we can make an inference via training data. We then tackle the "gradual transition" problem by using a bootstrapping learning approach. It may add some linked documents to reduce the gap between training and test data. We tackle the "mutation" problem via our multi-source model, which captures the relationship between the words describing events and concepts such as "Clinton" via web statistics.

3. Overview of our M3 system

In this Section, we introduce our M3 framework. We will present the framework of our system, following by a presentation of the multi-resolution inference structure. We leave the detailed discussion on sub-domain transductive inference and multi-source, multi-resolution bootstrapping inference to Sections 4 and 5.

Figure 6 shows the bootstrapping architecture of our system. Given a corpus, we first employ the high performance mid-level detectors such as the anchorperson, commercial, finance and sports detectors [5] to segment the corpus into eight sub-domain data sets, as shown in Figure 1. We then perform the multi-resolution, multi-source and multi-modal (M3) transductive learning model as shown in Figure 7 to detect the concepts in each of the sub-domain data sets separately. After that, we select results with high confidence from all sub-domain data sets. If the number of positive test data is above the threshold, or when data propagation has converged, we will terminate the process. Otherwise, we employ a bootstrapping method to make further inferences. We will repeat the M3 transductive inference in those sub-domain data sets, which new test data are added into the training data.

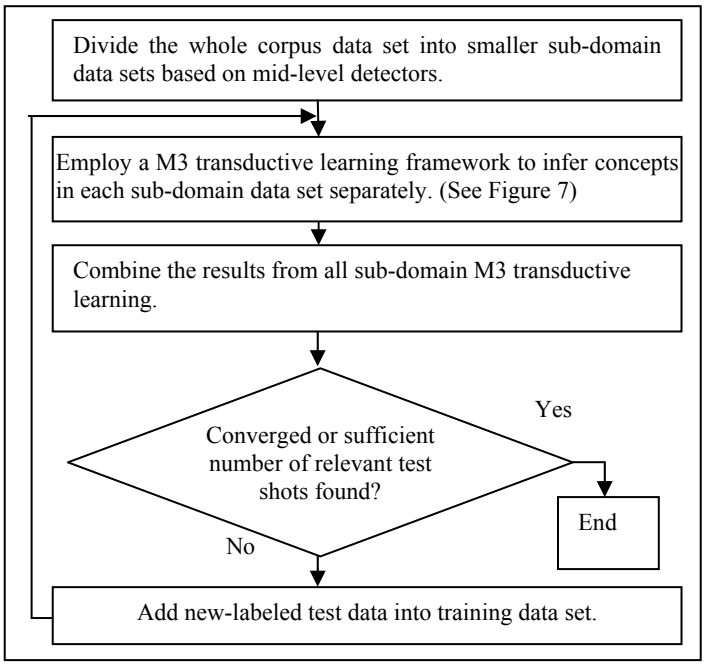

Figure 6: The bootstrapping architecture of our system

Figure 7 illustrates our M3 transductive learning model performed for each sub-domain separately. The transductive learning includes shot, MM discourse and story layer inference. At each layer, we carry out transductive inference to detect concepts by performing the average-link clustering. We analyze the resulting clusters based on the occurrence and frequency of positive training samples and multi-resolution constraints. We classify the test shots using the procedures outlined in Section 4.4 into three types: positive (P), unknown (U) and negative (N) sets. The so-called unknown category of test shots consists of those that cannot be labeled as positive or negative with sufficiently high confidence. Our inference begins with the highest resolution layer- the shot layer. Based on the information at the shot layer, we classify the test shots into (P1), (U1), and (N1) sets. The U1 set will be further processed at the lower resolution layer – the MM discourse layer by performing the web-based analysis and transductive learning. After the analysis at the MM discourse layer, we can divide U1 again into three sets – the positive (P2), the negative (N2), and the unknown (U2) sets. Finally, we further process the U2 clusters by labeling them using the topics extracted at the story layer. After the story layer analysis, the U2 data set is classified into (P3), (N3) and (U3) sets. The final ranking of the shots is based on: P1, P2, P3, U3, N3, N2, N1.

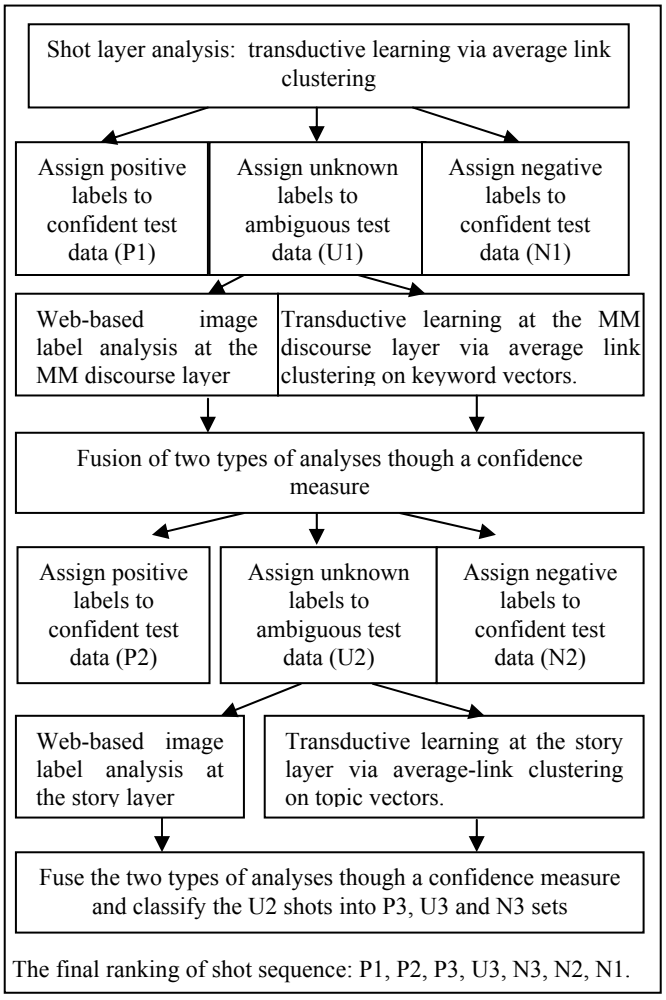

Figure 7: Our M3 transductive learning inference structure

4. Transductive learning at the multi-resolution layer

In this Section, we first introduce multi-modal features in our framework. We then discuss the constraint based clustering. After that, we present our transductive learning model. Finally, we discuss how our algorithm tackles the problems of sub-domain adaptation.

4.1 Visual features at the shot layer

At the shot layer, we use common low-level visual features as used in most other works to analyze the key frame images for each shot. The visual features used include Edge Histogram Layout (EHL), Color Correlogram (CC), Color Moments (CM), Co-occurrence Texture (CT) and Wavelet Texture Grid (WTG). For each shot, we extract the above visual features and generate a feature vector $f(f_1, f_2, f_3, ..., f_t)$. As discussed in the previous Section, clustering is performed as part of our transductive learning. In a clustering process, one of the most important aspects is the definition of similarity measure. At the shot layer, we adopt the cosine similarity between feature vectors as the similarity between shot i and shot j:

$$\cos sim(i, j) = \frac{\sum_{k=1}^{t}(f_{ki} \bullet f_{kj})}{\sqrt{\sum_{k=1}^{t} f_{ki}^{\ 2} \bullet \sum_{k=1}^{t} f_{kj}^{\ 2}}} \qquad (1)$$

4.2 Text features at the MM discourse and story layer

Because the characteristics between MM discourse and story layer are different, we capture two types of text semantics respectively. Usually, one MM discourse includes one or very few sentences, which come from automatic speech recognition (ASR) results. Due to ASR errors and limited number of words, there is insufficient text information to capture text semantics. Thus, we just extract keywords at the MM discourse layer. On the other hand, a story includes many sentences and covers a complete end. This provides relatively rich linguistic information. We employ a topic extraction algorithm to capture text focus of the story. Compared to keywords, using topic words to infer visual concepts is more effective. This is because visual information is indirectly represented in the focus of the topics.

Both the keyword vector at the MM discourse layer and topic vector at the story layer are used to represent the text content of individual entity at both layers. We denote such text vector as $T(w_1, w_2, ..., w_n)$. On the other hand, we need to model the text content of multiple entities at the cluster level. Here we regard the frequently occurring terms as the labels of the clusters, where the clusters are derived from higher resolution analysis. That is, when we form keyword vectors at the MM discourse layer, the cluster results come from the shot layer analysis. Similarly, when we form topic vectors at the story layer, the clusters came from the MM discourse layer. Given a cluster vcr_i, we can extract such text labels to represent vcr_i using the following term ranking formula:

$$P(W_k, vcr_i) = \frac{NumofShotsInTheClusterIncludes(W_k)}{NumofShotsInTheCluster} \qquad (2)$$

If $P(W_k, vcr_i) > \beta$, we regard the term W_k as a text label for such a cluster. For each cluster, we model it as a text vector $TC(w_1, w_2, ..., w_n)$.

Although TC is the text label vector for the clusters, we employ the same text similarity measure. Thus, in the discussion on text similarity, we use one notation T to represent text vectors. As different text vectors may express the same concept, we propose a new web-based concept similarity measure that uses the information redundancy of the Web to assign high similarity scores to those relevant text vectors with few or even non-overlapping words such as the cases in Figure 8, in which both are about the concept "*Clinton*". Of course, such a method will still assign high similarity scores to two text vectors when there is high word overlap between them. The definition of such a similarity measure is:

$$Sim_{unit}(T1, T2) = 1 - |P_{web}(C_x \mid T1) - P_{web}(C_x \mid T2)| \quad (3)$$

where T1, T2 are two different text vector instances, C_x is the word from the concept text descriptions. We estimate $P_{web}(C_x \mid T)$ in Equation (3) as follows:

$$P_{web}(C_x \mid T) = \frac{\#(C_x, T)}{\#(T)} \quad (4)$$

where we compute # (C_x,T) by using the text description of concept X together with text vector T as the query to the Google search engine, and count the estimated number of hits that include the query terms. # (T) is computed in a similar manner.

To improve the effectiveness of Web searches, we need to select a few dominant terms in the text vector as query. Here we employ a text weighting scheme based on tf.rf developed in [13]. Such a method measures the importance of a term based on its frequency (tf) and relevant frequency (rf). Here the relevant frequency is obtained by computing the ratio of the term's occurrences in the positive and negative training data. In our application, we found that some important terms may occur only in test data; while the relevance frequency rf in the tf.rf approach does not consider terms occurred only in test set. In order to tackle this problem, we leverage the web statistics to obtain other relevance information. The new weighting Equation is:

$$Weight(W_i) = tf * [\alpha_w * \frac{\#(W_i, C_x)_{training}}{\#(W_i)_{training}} + (1 - \alpha_w) * \frac{\#(W_i, C_x)_{web}}{\#(W_i)_{web}}] \quad (5)$$

where $\#(W_i, C_x)_{training}$ and $\#(W_i)_{training}$ are obtained by counting the co-occurrence between term W_i and C_x, and the occurrence of term W_i in training data respectively; while $\#(W_i, C_x)_{web}$ and $\#(W_i)_{web}$ are computed in a similar manner as the variables in Equation (4). α_w is designed to balance the training data and web statistics, and is estimated as:

$$\alpha_w = \begin{cases} Log_{(\lambda+1)}(1+tf) & \text{tf}<\lambda \\ Log_{(\lambda+1)}(1+\lambda) & \text{Otherwise} \end{cases} \quad (6)$$

where tf is the term frequency and λ is a predefined threshold. That is if the term has sufficiently high frequency in training data, then the value of rf is computed based on the statistics in training data. Otherwise, we will incorporate web statistics for smoothing. The resulting scheme considers all the words in the whole corpus instead of just the words in the training data.

4.3 The constraints-based clustering

The key to transductive learning is how to map specific (test) cases to corresponding (training) cases. Such a mapping could be obtained by an average-link clustering. The ideal clustering results for transductive learning are that the clusters are pure and large. If the clusters are pure, we could achieve high precision. If the size of the clusters is large, we could obtain high recall. However, it is often hard to achieve both advantages at the same time. Thus, our strategy includes three steps.

First, we attempt to obtain small and pure cluster results. In order to make the cluster results as pure as possible, our shot layer clustering process is based not only on visual features, but also constraints from different resolutions. Such text constraints from lower resolutions provide the cannot-link constrains that avoid the clustering of semantically dissimilar shots together. Figure 8 illustrates the cannot-link constraints from MM discourse and story layer to purify the shot clustering results. If we were to measure the similarity between these two shots by global visual features alone, they may have some degrees of similarity as shown in Figure 8. However, when we consider its contextual information at the MM discourse and story layers, we would know that one is related to the concept "Clinton" while the other is irrelevant. That is, the two shots are not similar to the concept "Clinton" in the semantic view. The text constraints come from the measure of homogeneity of text semantics. The text-based Cannot-Link constraint is defined as follows. For a shot layer clustering, given two shots S(i) and S(j) with high visual similarity, if $Sim_{MD}[S(i), S(j)] < \delta_1$ and $Sim_{ST}[S(i), S(j)] < \delta_2$ then shots i and j cannot be clustered, where $Sim_{MD}()$ and $Sim_{ST}()$ are text similarity at the MM discourse and story layer respectively. Thus, the clusters we obtained in this step are relatively pure and small.

The word vector at the MM discourse layer is {Clinton, Chile, summit....}

The topic vector at the story is {Clinton}

The word vector at the MM discourse layer is {Nancy Reagan, lumpectomy ...}

The topic vector at the story layer is {Nancy Reagan}

Figure 8: An example of cannot-link text constraints purified visual shot clustering results

Second, we further cluster the results from the first step by using the must-link and cannot-link constraints, along with the text features at the MM discourse layer. The must-link constraints, derived from visual shot clustering, ensure that two highly "visually" similar shots that were gathered in the shot layer analysis remain clustered at the MM discourse layer. It helps to establish the linkage between visual features and ASR terms. That is, given two shots S(i) and S(j), and vcr is a cluster among visual-

shot based clustering results; then the must-link constraint at the MM discourse cluster is defined as follows: $\exists S(i), S(j), k$, if $S(i)$ and $S(j) \in vcr_k$, the shot i and j must be linked together at the MM discourse layer analysis. We also introduce a cannot-link constraint at the MM discourse layer from the lower story layer too. That is, given two shots S(i) and S(j) with high text similarity at the MM discourse layer, if $Sim_{ST}[S(i), S(j)] < \delta_2$ then shots i and j cannot be clustered together, where $Sim_{ST}()$ are text similarity at the story layer.

Third, we further cluster the shots based on the results at the MM discourse layer by using the must-link constraints and text features at the story layer. The must-link constraints at the story layer clustering is defined as: suppose mmcr is a cluster from MM discourse layer clustering results, $\exists S(i), S(j), k$, if $S(i)$ and $S(j) \in mmcr_k$, then shot i and j must be linked together at the story layer.

Finally, we rank the results based on the shot, MM discourse and story layers. Because the cluster results at the shot layer are the purest, we have the highest confidence for the results and we assign them the highest ranking. With the sizes of the cluster results becoming larger and larger at the MM discourse and story layer, our confidence of the clusters become lower. Thus, we assign lower rankings to these results. Based on the above strategy, we could approximately obtain purer and larger cluster results to facilitate our inference.

4.4 Transductive inference

The transductive inference is used to analyze both the visual and text features in our framework. It involves three stages.

In stage 1, a series of clustering are applied as different inference hypotheses using the constraint-based average-link clustering, which is discussed in Section 4.3.

In stage 2, a hypothesis is selected based on Vapnik Combined Bound [26] to determine the confidence of the series of clusters. That is, given a hypothesis $h \in H$ and unlabeled test set X_u , the predicted risk $R_h(X_u)$ of unlabeled samples is:

$$R_h(X_u) \le R_h(X_m) + \sqrt{(\frac{m+u}{u})(\frac{\tau + \log(C-1) + \ln\frac{1}{\delta}}{m})} \quad (7)$$

where m is the number of labeled samples in the training data; u is the number of unlabeled samples in test data; δ is the confidence; C is the maximal partitions in the corpus; τ is the number of clusters in current hypotheses (cluster); and $R_h(X_m)$ is the total number of positive and negative training data in the same clusters.

Such clustering typically results in three types of clusters:

Type1: The cluster contains data from both training and test sets. Only in this type of clusters, we could use labeled training data to predict the relevance of the unlabeled test data.

Type 2: The cluster contains only data from the training set. This shows that such training data is not useful in predicting the relevance of the unlabeled test set.

Type 3: The cluster contains data from the test set only. We do not know whether such a cluster is relevant to concept X or not. We call such clusters ambiguous/unknown clusters.

In stage 3, we label the test sample in the selected hypotheses by using the training data in the same cluster. That is, given a test shot S_{unit} appearing in a cluster containing both training and test data under a certain resolution, we compute the probability of S_{unit} appearing in cluster C_x, or $P(C_x \mid S_{unit})$, as:

$$P(C_x \mid S_{unit}) = \frac{NumofTrainingShotsWith(C_x)\text{int}heCluster}{NumOfTrainingShotsIntheCluster} \quad (8)$$

However, some clusters may include very few training data, which may violate the "Law of large numbers" in probability inference. Thus, we have to add a variable: confidence index (CI), to partially tackle this problem. We estimate CI in a similar way as α_w using Equation (6), and we use TD to replace the item tf. TD represents the number of training data in a cluster.

In addition, we include the probability of concept C_x in sub-domain D_i, or $P(C_x \mid D_i)$, into the final score function for S as:

$$Score(S) = CI * P(C_x \mid S) * \log_2[1 + P(C_x \mid D_i] \quad (9)$$

where CI is the confidence index for the cluster that includes the test shot S. Because in some sub-domains there is even no positive training data, we employ the smoothing method [12] to estimate the probability of concept C_x in sub-domain D_i as:

$$P(C_x \mid D_i) \approx \frac{ShotWith(C_x)in(D_i)+1}{ShotsIn(D_i)IntheTraining+1} \quad (10)$$

where D_i is a sub-domain data set.

Also because some clusters include only test data, that is type 3 clusters, we could not compute Equation (8). Thus we adopt a multi-resolution analysis strategy and a web-based text smoothing approach to tackle this problem. At the MM discourse layer, we bring web statistics into the framework when the training data is not enough. This is because the current web is a huge data depository and we can make use of the term co-occurrence relationship to explore the semantics. That is, given a test shot S, we can find a MM discourse layer cluster $mmcr_j$, which includes the test shot S. The text label vector for the cluster is TC, which is obtained by Equation (2). The semantic concept inference is defined as follows:

$$Score(C_x \mid S) = [CI * P_{corpus}(C_x \mid S) + (1-CI) * P_{web}(C_x \mid TC)] * \log_2[1 + P(C_x \mid D_i)] \quad (11)$$

$$P_{Corpus}(C_x \mid S) = \frac{NumofTrain\ ingShotsWi\ th(C_x)In(mmcr_j)}{NumofTrain\ ingShotsIn(mmcr_j)} \quad (12)$$

$$P_{web}(C_x \mid TC) = \frac{\#(C_x, TC)_{web}}{\#(TC)_{web}} \quad (13)$$

We obtain $\#(C_x, TC_i)_{web}$, $\#(TC_i)_{web}$ in a similar manner as in Equation (4) and CI is the confidence index.

At the story layer, the inference is similar as that at the MM discourse layer.

After each layer analysis, a shot classification component is used to divide the test shots into positive (P), unknown (U) and negative (N) sets. We can classify the test shots S at a certain resolution layer as follows:

a) If $Score(C_x \mid S_{layer}) > \alpha_{layer}$, we label it as positive data and put it into the P shot set.

b) If $Score(C_x \mid S_{layer}) < \delta_{layer}$, we label it as negative data and put it into the N shot set.

c) Otherwise, we assign an unknown label to it and put it into U set for the lower resolution layer inference.

where $\alpha_{layer}, \delta_{layer}$ are pre-defined thresholds with $\alpha_{layer} > \delta_{layer}$.

4.5 An adaptive transductive algorithm

As shown in Figure 1, there is the problem of imbalanced distribution of training data in certain segments of the sub-domain. However, the existing cross-domain adaption algorithms could not tackle the problem. This is because they adopted supervised learning. One of the most important assumptions in supervised learning is that the training samples have the same distribution as that of future test samples. Thus, if there is a problem of mismatch or imbalanced distribution of training data (say very few or even no positive training data) in some sub-domain data sets, it is hard for these algorithms to adapt their classifiers.

Here, we develop a pseudo-Vapnik combined error bound transductive learning approach to partially tackle this problem. Our inference follows the label of training data if and only if there is enough training data with the same label in the same cluster as the target test data. However, as we have discussed in the previous Section, the function of Vapnik combined error bound is to select a cluster hypothesis. If there are very few or even no positive data, it is hard to compute the term $R_h(X_m)$ in Equation (7) accurately. To tackle this problem, we develop a pseudo-Vapnik combined error bound adaption algorithm. Given that there is insufficient training data for the current clusters in current sub-domain dataset, we leverage on training data in other sub-domains to estimate the Vapnik combined error bound. We obtain similarity values from those sub-domain data sets that have enough positive and negative training data. We then use the average of these similarity values as the pseudo-Vapnik combined error bound for the sub-domain with imbalance training data.

The detail of the adaptive cross sub-domain transductive learning algorithm in our M3 framework is outlined as follows:

Input: A full sample set $X=\{X_1, X_2 \ldots X_{m+u}\}$;

A training set with semantic labels $\{(X_1, Y_1) \ldots (X_m, Y_m)\}$ algorithm.

Step 1: Compute the similarity between each sample pairs (X_i, X_j) and build a similarity matrix.

Step 2: If there is a constraint between each sample pairs (X_i, X_j), then we set $Sim(X_i, X_j)=0$ for a Cannot-Link constraint; or $Sim(X_i, X_j)=1$ for a Must-Link constraint.

Step3: Place each sample in X as its own cluster, creating the list of clusters C: $C=c_1, c_2 \ldots c_{l+u}$

While (there exists a pair of mergeable clusters) do

(a) Select a pair of clusters c_i and c_j according to the minimal average group distance

(b) Merge c_i to c_j and remove c_i

(c) Save each partition as a hypothesis to the disk.

Endwhile

Step 4: For each hypothesis, we compare it with pseudo-Vapnik combined bound and select the hypothesis that satisfies our pseudo-Vapnik combined bound constraints as our final clustering result.

Step 5: Label the test samples for those clusters that include both training and test data.

Figure 9: A constraint based transductive learning algorithm

5. The bootstrapping framework

We employ the bootstrapping technique to further process the unknown test results from the M3 transductive learning. We attempt to add some linked documents from test data to reduce the gap between the training and test data to tackle the "graduation transition" problem as discussed in Section 2.3. Up to now, many bootstrapping algorithms are available. Most of them assume that the newly added unlabeled data belongs to the same distribution as the labeled data. However, this is not always true. In order to reduce the errors from newly added unlabeled data, we propose a new bootstrapping algorithm, shown in Figure 10. The basic idea is that we use test data with high inference confidence to rerank the data in unknown clusters from our initial M3 transductive model. The main differences between our approach and other bootstrapping work [7] are (a) in order to reduce the risk of adding unlabeled data with wrong annotation labels, we set the confidence of the newly added test data to a relatively low value as compared to the labeled training data; and (b), the bootstrapping method only processes the data in the unknown clusters from our M3 transductive learning, rather than the whole test set.

The detail of our bootstrapping algorithm is outlined as follows:

Notation: P(i)(j) is a positive shot set, where i shows the layer of resolution for inference with i=1 denoting the inference at shot layer; i=2 for MM discourse layer; and i=3 for story layer. Index j records the number of iterations in the bootstrapping module. N(i)(j) and U(i)(j) are defined in a similar manner for the negative and unknown shot sets respectively.

Step1: j=0; initialize K=T, where T is a constraint (say T=50). We perform an initial M3 transductive inference. We obtain an initial shot ranking sequence RS{P(1)(0), P(2)(0), P(3)(0), U(3)(0), N(3)(0), N(2)(0), N(1)(0)};

Step2: If U(3)(j) is empty or the number of the positive sequence in RS[P(1)(0),P(2)(0), P(3)(0), ...P(i)(j)] is above the user's requirement or data propagation has converged, we stop the program.

Else, goto step 3.

Step3: We obtain the top K shots from the shot ranking sequence RS as newly added positive labeled data and the bottom K shots from RS as newly added negative labeled data.

Step 4: We redo the M3 transductive inference. We divide U(3)(j) into two sets. One set is a labeled set: which includes three positive P(1)(j+1), P(2)(j+1), P(3)(j+1) and three negative N(1)(j+1), N(2)(j+1), N(3)(j+1). The other set is still the unknown set U(3)(j+1).

Step 5: Add the new inference results into the shot ranking sequence: RS{P(1)(0), P(2)(0), P(3)(0) ... P(1)(j+1), P(2)(j+1), P(3)(j+1), U(3)(j+1), N(3)(j+1), N(2)(j+1), N(1)(j+1)...N(3)(0), N(2)(0), N(1)(0)};

Step 6: Update j and K for next iteration, j=j+1; K=K+T; Goto step 2

Figure 10: Our bootstrapping algorithm

6. Experiment

We use the training and test sets of the TRECVID 2004 corpus to infer the visual concepts. The corpus includes 137 hours of news video from CNN Headline News and ABC World News Tonight; 67 hours of news video are used for training and 70 hours for testing. We measure the effectiveness of our model using all the 10 semantic concepts defined for the TRECVID 2004 semantic concept task. The concepts are listed in Figure 1.

The performance of the system is measured using mean average precision (MAP) based on the top 2000 retrieved shots for all the ten concepts. This is the same to the evaluation used in TRECVID 2004.

6.1 Test1: Multi-resolution analysis

We perform three experiments on concept detection based on: (a) shot layer visual analysis without text, (b) shot layer + MM discourse layer analysis, and (c) full M3 model with story layer analysis. We list the results in Figure 11.

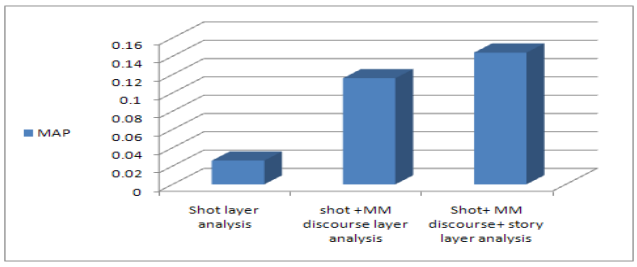

Figure 11: The results based on different combination of multi-resolution analysis

From the Figure, we observe that using only the shot layer visual analysis without text, we could achieve only very low MAP of 0.026. By incorporating text semantics at MM discourse layer, we could improve the result substantially to an MAP of 0.116. The best result is achieved when we perform a multi-resolution analysis at the shot, MM discourse and story level, with a MAP of 0.144. The results suggest that different modal features only work well in different temporal resolutions and different resolutions exhibit different types of semantics.

6.2 Test 2: comparison with the reported systems on TRECVID 2004 Dataset

In order to compare our results with other related systems on the same corpus, we tabulate in Figure 12 the results of all reported systems that have completed all the ten concepts in TRECVID 2004. The systems are ranked based on their MAP performance from left (the best) to right (the worst).

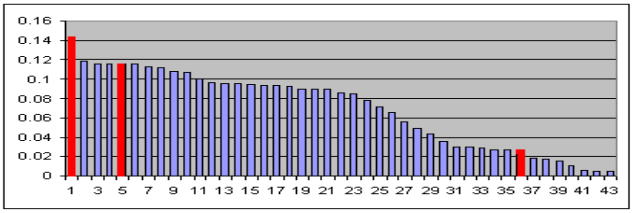

Figure 12: The comparison with other systems in TRECVID

Figure 12 shows that our three combinations of systems are ranked as 1st, 5th, and 36th. Compare to the best reported system ranked 2nd in Figure 12, our M3 transductive framework achieved more than 21% improvement in MAP. We achieve better performance, mainly because:

a) We encode the distributions of concepts in sub-domains into the framework. This benefits the detection and discrimination of concepts that have high occurrence in certain sub-domains. In addition, we can also obtain better statistical patterns, because we use the training data from different sub-domain separately so as to better capture the characteristics of different sub-domains.

b) We employ the multi-resolution fusion strategy to combine text and visual features.

6.3 Test 3: the bootstrapping approach

After performing the sub-domain based M3 transductive learning, we continue to run several iterations of the bootstrapping algorithm. The result is shown in Table 1.

Table 1: The results from our bootstrapping approach

	Our M3 approach	M3+bootstrapping approach
MAP	0.144	0.145

Table 1 shows that there is further improvement of 1% when we employ the bootstrapping approach. The improvement is statistically significant as judged by using paired t-test [10] ($p<0.05$). This shows that the bootstrapping method is feasible.

7. Conclusion

Although research on semantic concept detection has been carried out for several years, the analysis based on multiple sub-domain concept distribution, multi-resolution fusion and bootstrapping technique is relatively recent. This paper outlines a sub-domain based M3 bootstrapping learning framework. In this framework, we exploit the concept occurrence distribution in the sub-domain to boost the performance of the system and propose an adaptive cross sub-domain algorithm to tackle the imbalance in the concept distribution. In the multi-resolution model, we integrate visual and text features by using a multi-resolution inference structure and constraints such that the evidences for inference come from the integration of multiple modalities. Finally, we employ bootstrapping to process the unknown data from our initial M3 framework. The experimental results on TRECVID 2004 data set demonstrate that our approach is able to achieve over 22% improvement in MAP over the best-reported system.

This work is only the beginning. Although the performance of our system is better than the reported systems, it is still far from a satisfactory level of performance for general use. Further research can be carried out as follows:

- In our framework, we regard the statistical dependency as causality. However, this is not always true. We plan to integrate statistical corpus knowledge, together with the knowledge of human annotations and manually built encyclopedia such as Wikipedia to further improve the performance.
- We plan to encode concept relationships in concept detection.
- We plan to improve our bootstrapping techniques.

REFERENCES

[1] A. Amir et al, "IBM research TRECVID 2005 video retrieval system", Proceedings of TRECVID 2005, Gaithersburg, MD, November 2005 available at: http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/

[2] L. Chaisorn, "A Hierarchical Multi-Modal approach to story segmentation in news video", PhD thesis in National University of Singapore, 2004

[3] S.F. Chang, R. Manmatha, and T.S. Chua, "Combining text and audio-visual features in video indexing", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp1005-1008, 2005

[4] T.S. Chua, S.F. Chang, L. Chaisorn, and W. H. Hsu, "Story Boundary Detection in Large Broadcast News Video Archives-Techniques, Experience and Trends", Proceedings of the 12th ACM International Conference on Multimedia pp. 656-659, 2004

[5] T.S. Chua et al, "TRECVID 2004 Search and Feature Extraction Task by NUS PRIS" Proceedings of (VIDEO) TRECVID 2004, Gaithersburg, MD, November 2004, available at : http://www-nlpir.nist.gov/projects/tvpubs/

[6] T.S. Chua et al, "TRECVID 2005 by NUS PRIS", Proceeding of TRECVID 2005, Gaithersburg, MD, November 2005, available at http://www-nlpir.nist.gov/projects/tvpubs/

[7] H.M. Feng, R. Shi, T.S. Chua, "A bootstrapping framework for annotating and retrieving WWW images." In Proceeding of the 12th ACM Multimedia conference International conference pp. 960-967, 2004

[8] J.L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System." Speech Communication, 37(1-2) pp89-108, 2002.

[9] A. Hauptmann et al, "Multi-Lingual Broadcast News Retrieval" Proceedings of TRECVID 2006 available at: http://www-nlpir.nist.gov/projects/tvpubs/

[10] D.A. Hull, "Using statistical testing in the evaluation of retrieval experiments". Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pp329-338, 1993

[11] A.K. Jain, M.N. Murty , and P.J. Flynn , "Data Clustering: A Review", ACM Computing Surveys, Vol 31, No. 3, pp. 264-323,1999

[12] D. Jurafsky and J. H. Martin, "Speech and language processing", published by Prentice-Hall Inc 2000.

[13] M. Lan, C. L. Tan and H. B. Low "Proposing a new term weighting scheme for text categorization", Proceeding of the 21st National Conference on Artificial Intelligence, AAAI-2006

[14] C.Y. Lin, "Robust Automated Topic Identification" Ph.D. Thesis, University of Southern California 1997

[15] M. R. Naphade and J. R. Smith , "On the detection of semantic concepts at TRECVID", Proceedings of the 12th ACM Multimedia pp.660-667, 2004

[16] G.J. Qi, X.S. Hua, Y. Song, J.H. Tang, H.J. Zhang, "Transductive Inference with Hierarchical Clustering for Video Annotation" International Conference on Multimedia and Expo, pp.643 – 646, 2007

[17] L. A. Rowe and R. Jain, "ACM SIGMM Retreat Report on Future Directions in Multimedia Research", ACM Transactions on Multimedia Computing, Communications, and Applications, Vol 1, issues 1 pp3-13, 2005

[18] C. G.M. Snoek, et al, "The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia", Proceedings of the 14th ACM Multimedia International conference, pp.421 – 430, 2006.

[19] Q. Tian, J. Yu, Q. Xue, and N. Sebe, "A New Analysis of the Value of Unlabeled Data in Semi-Supervised Learning for Image Retrieval", Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2004), Vol.2, pp.1019-1022, 2004.

[20] V.N. Vapnik, "Statistical learning theory", Wiley Interscience New York. pp120-200, 1998,

[21] R. Yan and M. R. Naphade "Semi-supervised Cross Feature Learning for Semantic Concept Detection in Video" Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR), vol. 1,pp.657- 663, 2005.

[22] R. Yan, J. Yang and A. G. Hauptmann, "Learning Query-Class Dependent Weights for Automatic Video Retrieval", In Proceeding of the 12th ACM Multimedia conference International conference, pp. 548 – 555, 2004

[23] J. Yang, A. Hauptmann, M. Y. Chen, "Finding Person X: Correlating Names with Visual Appearances" , International Conference on Image and Video Retrieval (CIVR'04), Dublin City University, Ireland, July 21-23, 2004

[24] J. Yang, R. Yan and Hauptmann, "Cross-Domain Video Concept Detection Using Adaptive SVMs", In Proceeding of the 15th annual ACM international conference on Multimedia, pp. 188-197 , 2007

[25] R. E. Yaniv, and L. Gerzon, "Effective Transductive Learning via PAC-Bayesian Model Selection." Technical Report CS-2004-05, IIT, 2004.