Capturing text semantics for concept detection in news video

Gang Wang, Tat-Seng Chua

Department of Computer Science
School of Computing, National University of Singapore
Computing 1, SINGAPORE 117590
{wanggang,chuats}@comp.nus.edu.sg

Abstract. The overwhelming amounts of multimedia contents have triggered the need for automatic semantic concept detection. However, as there are large variations in the visual feature space, text from automatic speech recognition (ASR) has been extensively used and found to be effective to complement visual features in the concept detection task. Generally, there are two common text analysis methods. One is text classification and the other is text retrieval. Both methods have their own strengths and weaknesses. In addition, fusion of text and visual analysis is still an open problem. In this paper, we present a novel multi-resolution, multi-source and multi-modal (M3) transductive learning framework. We fuse text and visual features via a multi-resolution model. This is because different modal features only work well in different temporal resolutions, which exhibit different types of semantics. We perform a multi-resolution analysis at the shot, multimedia discourse and story levels to capture the semantics in news video. While visual features play a dominant role at the shot level, text plays an increasingly important role as we move from the multimedia discourse towards the story levels. Our multi-source inference transductive model provides a solution to combine text classification and retrieval method together. We test our M3 transductive model on semantic concept detection on the TRECVID 2004 dataset. Preliminary results demonstrate that our approach is effective.

1. INTRODUCTION

The advancement in computer processor, storage and the growing availability of low-cost multimedia recording devices have led to an explosive amount of multimedia data. It is reported in [4] that there are 31 million hours of TV programs produced each year. The statistics from Internet study [13] shows that about 65% of Internet traffic is being taken up by transferring multimedia contents. Among them, about 73.79% is video related contents. In order to effectively use such large number of multimedia content, we need provide tools to facilitate the management and retrieval of multimedia contents. One of the most important tools is the automatic multimedia concept detectors, which index the multimedia data at the higher semantic

level. One such level is to index contents based on concepts frequently appear in queries such as the person-X or object X etc. However, it is very hard for visual object detectors to detect whether such concepts by visual information appear in a shot alone. This is because of the wide variations of visual objects in videos. The variations are caused by changes in appearance, shape, color and illumination conditions. Figure 1 shows examples of the concept "boat" in news video with different shapes and colors. On the other hand, we can obtain text from automatic speech recognition (ASR) in informational video such as news and documentary video. Thus, how to utilize text semantic to complement visual features to support concept detection is an important problem.

Fig. 1. Concept "boat/ship" with different shape and different colors

More formally, the concept detection task is defined as: given a set of predefined concept C: [C_1, C_2 ...Cn], develop a classifier to determine if concept C_i appears in shot S_k.

Most researchers first adopted either text classification under a supervised inductive learning framework [1] [22] [31] or text retrieval [3] [6] to capture text semantics. They then fused text and visual analysis by using heuristic rules, early/late fusion approaches or their combinations. In spite many efforts have been made, we are still far from achieving a good level of concept detection performance. Based on our analysis, we have identified two weaknesses in current systems that should be addressed to enhance the performance.

- How to make use of text analysis to overcome the problems of visual analysis and vice verse is an open problem.
- There are two types of knowledge: one is the concept text description and the other is the text and visual features in training data. Most current systems did not make use of both of knowledge simultaneously.

In this paper, we propose a M3 transductive framework to tackle the above two problems. The multi-resolution model is designed to let text and visual analysis support each other. In that model, we first analyze text and visual features at different resolutions. The analysis at any resolution will consider the evidence from the other resolution as context information. The transductive multi-source model integrates knowledge from the concept text description and training data. We also adopt a transductive inference model to analyze both text and visual features at the related

resolution. Because such an inference attempts to capture the distributions of training and test data by mapping test data to training data, we could know when we can make an inference via training data. For those test data that cannot be labeled by training data, our multi-source model brings web knowledge into the model to partially tackle the problem. We test our M3 transductive model on the concept detection task based on the TRECVID 2004 dataset. The test results demonstrate that our M3 transductive framework is superior to those systems based on text retrieval and classification. In addition, our system outperforms the state-of-the-arts systems based the single-resolution, single source supervised inductive inference framework.

The rest of the paper is organized as follows: Section 2 discusses related work, while Section 3 describes the analysis on text semantic. Section 4 introduces our M3 transductive model. Sections 5 and 6 discuss visual analysis at the shot layer and text analysis at the multimedia discourse and story layers respectively. The experimental test-bed and evaluation results are presented in Section 7. Finally, Section 8 concludes the paper.

2. Related work

In order to implement a generic automatic news video semantic concept detection system with a good performance, we need to tackle at least two challenges. One is how to capture the text semantics, and the other is how to fuse the text and visual semantics to support concept detection. In this Section, the related work on these two topics are covered. We then introduce the background of transductive learning and multi-resolution analysis.

2.1 Text semantics

Text information is an important information source for informational video. There are two widely used methods to capture text semantics. One is text classification and the other is text retrieval. Text classification [11] works for concepts that are transcribed with a specific and limited vocabulary such as the concept “Weather” in CNN headline news. However, in general, the performance of text classification in the concept detection task is not good. This is partly because of the high dimensionality of text features and the limited training data. Text retrieval methods [3] [6] [39] regarded words from concept text descriptions or some predefined keywords as the query and employed the text retrieval with query expansion techniques to capture the semantics. Such methods are the only effective means when the training data is sparse. Based on the above discussion, we can find that both text analysis methods have their own strength and weakness. Given a concept with some training data, it is hard to know in advance which method is better.

In natural language processing [15], researchers captured text semantics not only from words at the sentence layer, but also topics at the story layer. Here the topic refers to the main focus of a story. In general, there are three types of methods to capture topic semantics. They are statistical-based [25], knowledge-based [10] and

hybrid [12]. Among these techniques, only word frequency counting can be used robustly across different domains; the other techniques rely on stereotypical text structure or the functional structures of specific domains. In video processing, some researchers [29] adopted knowledge-based approaches to identify topic in some specific domains, such as cooking instruction videos. However, as far as we know, no researchers adopted topic identification techniques to support concept detection in an open domain such as news video [33].

In addition, Rowe [27] used caption syntax to infer visual concept in the image. For example, he found that the primary subject noun phrase usually denotes the most significant information in the media datum or its "focus". However, we could not directly utilize such syntactic semantic technologies from image caption retrieval to concept-X detection in news video. This is because semantic parsers are designed for the grammatical written language and speech recognition text often contains too many errors that render the semantic parser ineffective.

In general, the analysis in news video based on text only is effective only if the desired concepts appear in both visual and text contents.

2.2 Fusion of multi-modal features

General speaking, three types of methods are proposed to fuse multi-modal features. They are the rule-based; the machine-learning based; and the mixture of the two approaches.

Some researchers such as [11] adopted the rule-based approaches. However, the drawbacks of such approaches are the lack of scalability and robustness.

To overcome the problems of rule-based approaches, many fusion algorithms adopt the supervised inductive learning methods such as [36]. In [31], the authors identified two general fusion approaches: namely early fusion and late fusion. The early fusion scheme integrates unimodal features before learning the concepts. The strength of this approach is that it yields a truly multimedia feature representation, since the features are integrated from the start. One of the weaknesses of this approach is that it is difficult to combine features into a common representation. The late fusion scheme first reduces the unimodal features to separately learned concept scores, and then integrated these scores to learn the concepts. The advantage of this approach is that it focuses on the individual strength of modalities. However, it has the high cost of learning effort and the potential loss of correlation in mixed feature space. In general, given a concept, it is difficult to decide which fusion method is better.

In addition, some image annotation algorithms, such as the translation model [8] and cross-media relevance model [14] and so on, adopted unsupervised approaches to describe images to a vocabulary of blobs as the basis for annotation. This causes the performance of the systems to be strongly influenced by the quality of visual clustering alone. It may result in images with different semantic concepts but similar appearance to be grouped together, while images with the same semantic contents may be separated into different clusters due to diverse appearance.

A number of researchers [3] [39] attempted to combine machine learning based and rule-based fusion approaches together. However, the main problem of such a hybrid combination strategy is that it is hard to integrate both fusion schemes.

Furthermore, in news video processing, there is at least one common problem with the above three fusion methods. That is, the text keywords are not always aligned with related visual concept at the shot layer, such as the person X detection problem [37].

2.3 Transductive learning

Instead of obtaining a general hypothesis capable of classifying any "unseen" data under a supervised inductive learning framework, transductive learning [23] [26] [38] is concerned with directly classifying the given unlabeled data. The key to transductive learning is how to map specific (test) cases to specific (training) cases. Such a mapping could be obtained by a hierarchical clustering method [21]. However, there are at least two open problems. One is to segment the clusters until their contents are as pure and large as possible. A pure cluster is defined as the one where the labels of training samples are mostly positive or negative such that the entire cluster includes the test samples can be labeled accordingly. The other problem is to analyze the unknown clusters, which are impure clusters or clusters that include only test samples. Such unknown clusters can be analyzed using other sources of information to label the test samples contained in such clusters appropriately.

2.4 Multi-resolution analysis

The multi-resolution model is widely used in image processing, such as image pyramids [35]. Such an approach first analyses data at different resolutions to create a multi-resolution structure and then derives error metrics to help decide the best level of detail to use. Lin [20] and Li [17] used a multi-resolution model to detect shot and story boundaries for video and text documents respectively. They used information at the low resolution to locate the transition points and the high resolution to identify the exact boundaries by finding the maximal path. Similarly, Slaney et al [30] proposed a multi-resolution analysis method to detect discontinuities in video for story segmentation.

As far as we know, no multi-resolution models have been applied in the semantic concept detection task to fuse multi-modal features [33]. Most current approaches, especially those used in large scale TRECVID video concept detection and retrieval evaluations such as [6], employed a hybrid approach of using text to retrieve a subset of videos at the story layer before performing visual and text analysis at the shot level to re-rank the video shots. Such approaches are not multi-resolution fusion as the analysis at the story level is used as a filter, but not to reinforce the subsequent shot level analysis and vice versa. They may miss many relevant video shots that are not retrieved in the text–based story retrieval stage. An important characteristic of multi-resolution analysis is that the results of analysis at each resolution should support each other to overcome the respective weaknesses. Thus two key challenges of multi-

resolution video analysis are: (1) the definition of good units for fusion that leverage the strong points of text and visual features; and (2) the combination and integration of evidences from multi-resolution layers.

3. Analysis of text semantic

In this Section, we first discuss text analysis at different resolution layers. We then compare text retrieval and classification methods.

3.1 Text analysis at different resolution layers

In video analysis, one of the most widely used analysis unit is shot. A shot is an unbroken sequence of frames from one camera shot. 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 occur often in the middle of a sentence. Figure 2 illustrates the problem of analyzing text using shot units, where the sentence separated by three shot boundaries causes the mismatch between the text clue and the concept "Clinton". Yang et al [37] found that this is a common problem in news video analysis. To tackle this problem, Wilson and Divakaran [36] proposed to detect scene changes by using training data under a supervised learning framework.

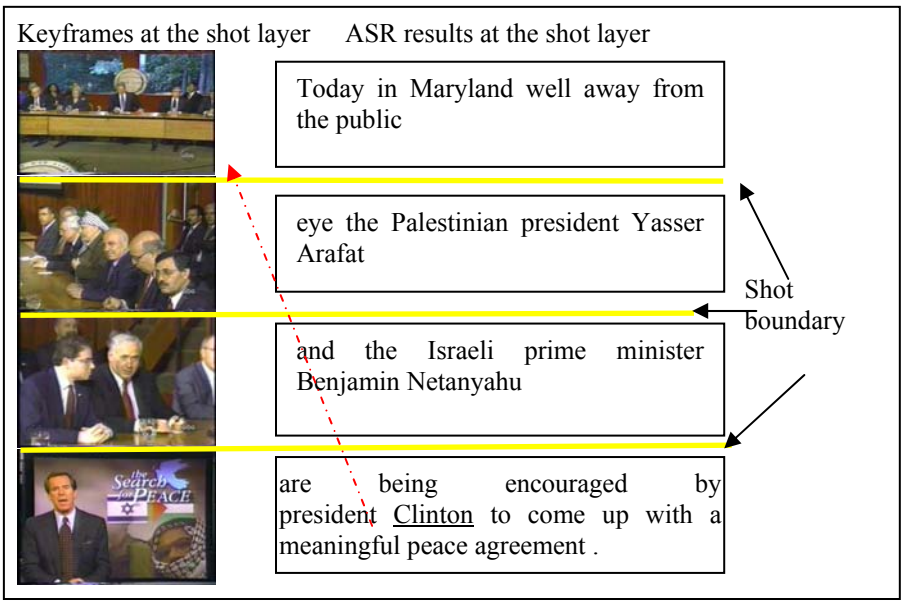

Fig. 2. The sentence separated by three shot boundaries causes the mismatch between the text clue and the concept "Clinton".

Because of collecting training data is time consuming, we propose a new unit, namely multimedia discourse to tackle the problem. The so-called multimedia (MM) discourse aims to capture the synchronization between the visual features at the shot level and the text feature at the sentence level. The MM discourse boundary occurs at the co-occurrence between the sentence and shot boundary. In this work, we adopt the speaker change boundaries generated by the speech recognizer [9] as the pseudo sentence boundaries. At the MM discourse layer, we capture the semantics mainly by extracting a group of keywords from the enclosed ASR text. We did not extract topic at the MM discourse layer, because there are often insufficient contents in such a unit to extract topics.

In general, there are three types of relations between keywords-based text semantics and shot-based visual semantics.

a) Type 1: We could infer the visual concept based on the text clues. Figure 3 shows an example where we found text clue word "Clinton" and visual content showing "Clinton" simultaneously.

Yesterday Mr. Clinton and Israeli Prime Minister Benjamin Netanyahu agreed on a partial peace plan. .....

Fig. 3. Text clue word "Clinton" co-occurred with visual concept.

b) Type 2: We could find the related text clue words, but the visual concept is not present. Figure 4 shows an example in which the keyword "Clinton" appears in the ASR transcripts, but we could not find semantic concept "Clinton" occurring in the shot.

President Clinton said today the U. S. must try to limit the international financial crisis before it gets worse in Latin America. .....

Fig. 4. An example of the text clues appears, but the concept did not occur.

c) Type 3: The visual concept is present but the related text clue words are absent. Figure 5 shows an example in which the concept occurs in the shot, but it is difficult to capture the text clues.

His defenders though are sending out a dual - pronged message.

Fig. 5. An example of the visual concept occurred, but we could not capture the text clues.

The above analysis highlights one challenge. That is how to find the words from ASR transcripts to describe the image content. In our framework, we first cluster visual similar images together. We then utilize the frequently occurring words as the label of visual image clusters.

We extract the text labels for an image cluster result (vcr_i) by using the following Equation:

$$P(W_k, vcr_i) = \frac{NumofShotsInTheClusterIncludes(W_k)}{NumofShotsInTheCluster} \quad (1)$$

If $P(W_k, vcr_i) > \beta$, we regard such a keyword as a text label for the cluster. For each of the visual cluster, we collect a group of words and build a text vector $TV(w_1, w_2, ..., w_n)$.

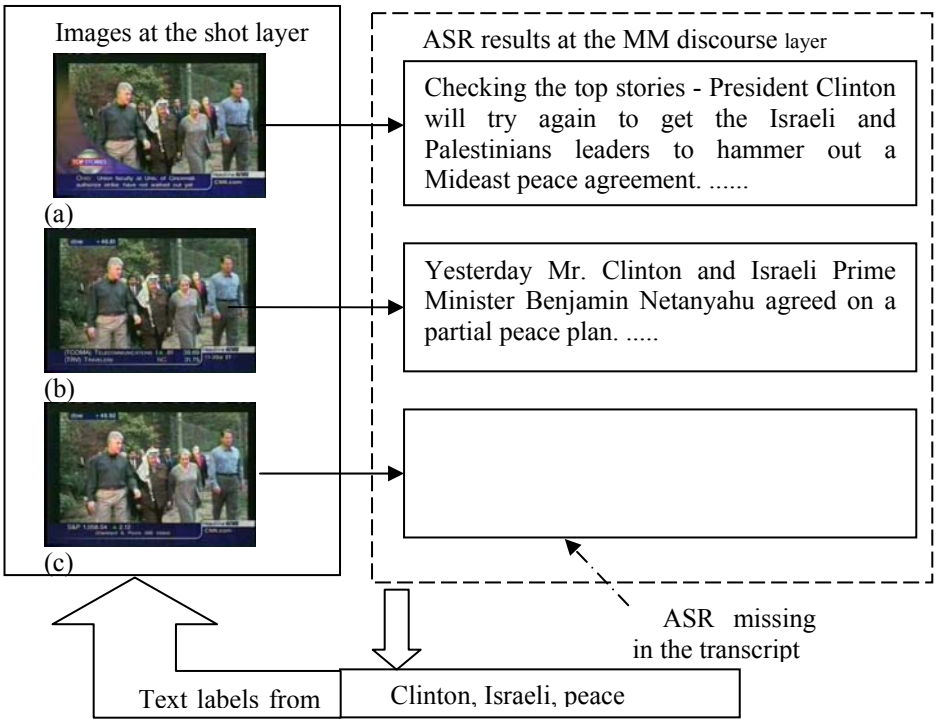

Fig. 6. An example of labeling a visual cluster by text information

Figure 6 gives an example of image labeling by keywords at the MM discourse layer. The visual cluster result vcr_i is labeled by a keyword vector TV_i= {Clinton, Israeli, peace}.

However, in some cases, we could not obtain any text labels, because the ASR words in a cluster exhibits large diversity. Figure 7 shows an example of such a case. Because of such a characteristic, we could partially tackle the problem in Figure 4. Based on the above analysis, we can see that we have partially tackled the problem of

inference visual concept from the available text clues by using Equation (1).

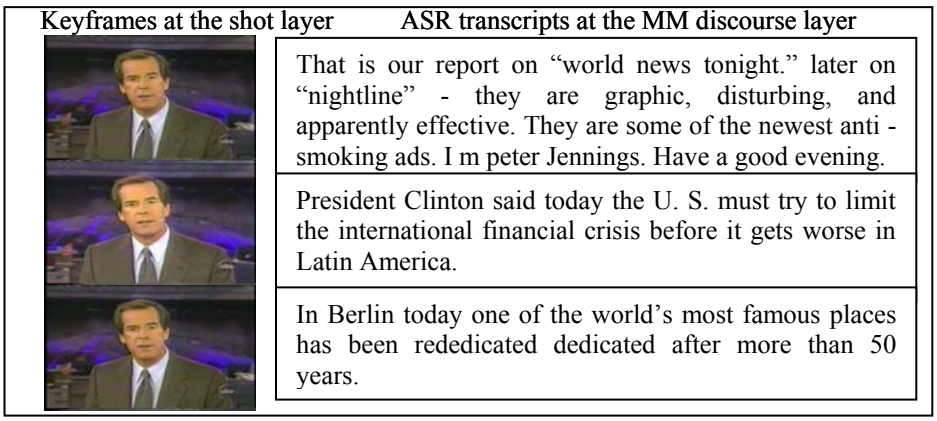

Fig. 7. An example of zero text labels could be extracted for the image cluster.

However, Equation (1) could not solve the problem in Figure 5. In order to tackle the problem, we add text analysis at the story layer into the framework. There are many story segmentation methods for news video as surveyed in [5]. In this paper, we perform a simple story segmentation using the heuristics based on anchorperson, some logos, cue phrases and commercial tags [6]. At the story layer, we attempt to capture the semantic concepts by exploring the relationship between the concept and the topics of a story. We employ a simple method developed in [19] to extract topics, which mainly depends on a set of high frequency ASR words in a story. We then build the linkage between topics and visual contents as similar to that at the MM discourse layer. The difference between the story layer labeling and the MM discourse layer labeling is that the target of text analysis at the story layer is topics. For the example shown in Figure 5, we are able to extract the topic labels of {Clinton, President}, as shown in Figure 8. Based on such topic labels, we can then conclude that the enclosed shots may have some degree of relevance to concept "Clinton". This could partially tackle the problem in Figure 5.

Fig. 8. The story layer context information.

3.2 Text classification and retrieval

Text classification usually refers to a supervised inductive learning algorithm using text features. However, such type of learning requires the estimation of unknown function for all possible input values. This implies the availability of good quality training data, which covers most typical types of data available in the test set. If such a condition is not satisfied, then the performance of such systems may drop significantly. One solution to obtain a good quality training data is to label as many training data as possible. However, preparing training data is a very time consuming task. Thus, in many cases, we need to face the sparse training data problem [24].

Text retrieval may be effective, when training data is not sufficient and test content includes some query terms. For example, for test data 1 in Figure 9(b), text retrieval could capture the concept "boat/ship", because the query word "ship" appeared in the ASR transcript. On the other hand, the text classification method may fail, because of the large gap between training and test data. For test data 2 in Figure 9(c) text classification can work well, but text retrieval will fail. This is because the ASR transcripts do not include any keyword related to the queries "boat" or "ship" and text retrieval fails to use the knowledge from training data. Hence, text classification and retrieval have their own strengths and we need to combine them to take advantage of their strengths in concept detection.

Furthermore, if we employ text analysis without support from visual feature analysis, we could not overcome the problems we discussed in the previous section. Thus, in our design, no matter whether we employ the classification-based or retrieval-based methods at the MM discourse and story layers, we must first build the linkage between the visual contents and terms from the ASR transcripts by using Equation (1).

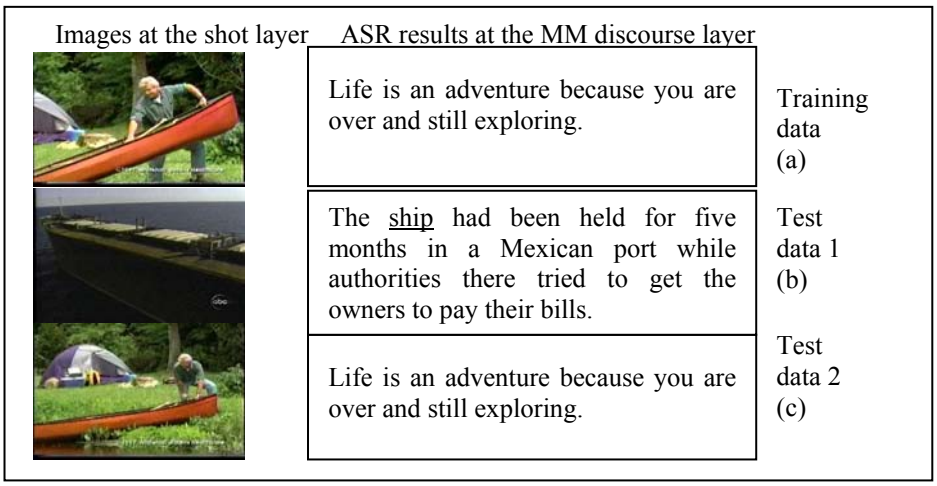

Fig. 9. An example of detecting concept "boat/ship" using two text analysis methods.

As far as we know, no efforts have been made to combine text-based classification and retrieval methods together to detect concept X. Our combined model first employs a transductive learning classification-based approach to label those test data that can be confidently labeled from training data by using either visual or text features. It then estimates the occurrence of concept for the remaining ambiguous test samples by using a multi-resolution analysis that incorporates web-based knowledge in a retrieval framework.

4. An introduction to our M3 TRANSDUCTIVE MODEL

In our M3 transductive framework, we analyze text and visual features at the shot, MM discourse and story respectively. While visual features play a dominant role at the shot level, text plays an increasingly important role as we move towards the multimedia discourse and story levels. In our design, we model the semantic concept detection problem as a conditional probability problem. That is, given a concept C_x, we want to rank a given test shot S, according to $P(S \mid C_x)$. Let us represent the visual part of a shot by S_v and the text part by S_t. This can be expanded as below.

$$P(S \mid C_x) = P(S_t, S_v \mid C_x) = \frac{P(C_x \mid S_t, S_v) P(S_t, S_v)}{P(C_x)} \quad (2)$$

In Equation (2), the denominator can be ignored for ranking the shots given any concept C_x. In addition, we assume that all shots are equally likely. This simplifies Equation (2) to:

$$P(S \mid C_x) \propto P(C_x \mid S_t, S_v) \quad (3)$$

As discussed in Section 3, we do not analyze text semantics at the shot layer, and capture text semantics only at the MM discourse and story layers, which we denote them as MD_t and ST_t respectively. In order to compute Equation (3), we make our inference via multi-resolution analysis, which is shown in Figure 10.

At the shot layer, we infer the labels of test shots by clustering shots via a transductive learning framework. The confidence of our inference depends on the amount of training data and its purity in any cluster. We divide test data into three categories (P1, U1, N1). The shots in P1 and N1 clusters can be labeled as positive and negative test shots respectively by using the training data in the same cluster with high confidence. The shots in U1 set are the shots that cannot be labeled as positive or negative with high confidence. Two situations may give rise to such unknown shots. One is that the cluster does not include any training data; and the other is when the number of training data is small or the purity of the cluster is low.

In order to label the U1 shots, we annotate such visual clusters by the keyword vector at the MM discourse layer. Two types of methods will be applied to make

further inference. One is to supplement the text analysis using web knowledge by capturing the relationship between the keyword vector at the MM discourse layer and words from web statistics. The other is to further cluster shots by a transductive learning method based on the web-enhanced keyword vectors. After the analysis at the MM discourse layer, we can divide U1 into three sets, which are a positive (P2), negative set (N2) and unknown set (U2).

Finally, we further disambiguate the U2 clusters by using the topics extracted at the story layer. We perform a similar text inference as the MM discourse layer and rank the U2 shots based on the story layer inference. We save the ranking result in the story result set (SR). The final ranking of the shots is as follows: P1, P2, SR, N2, N1.

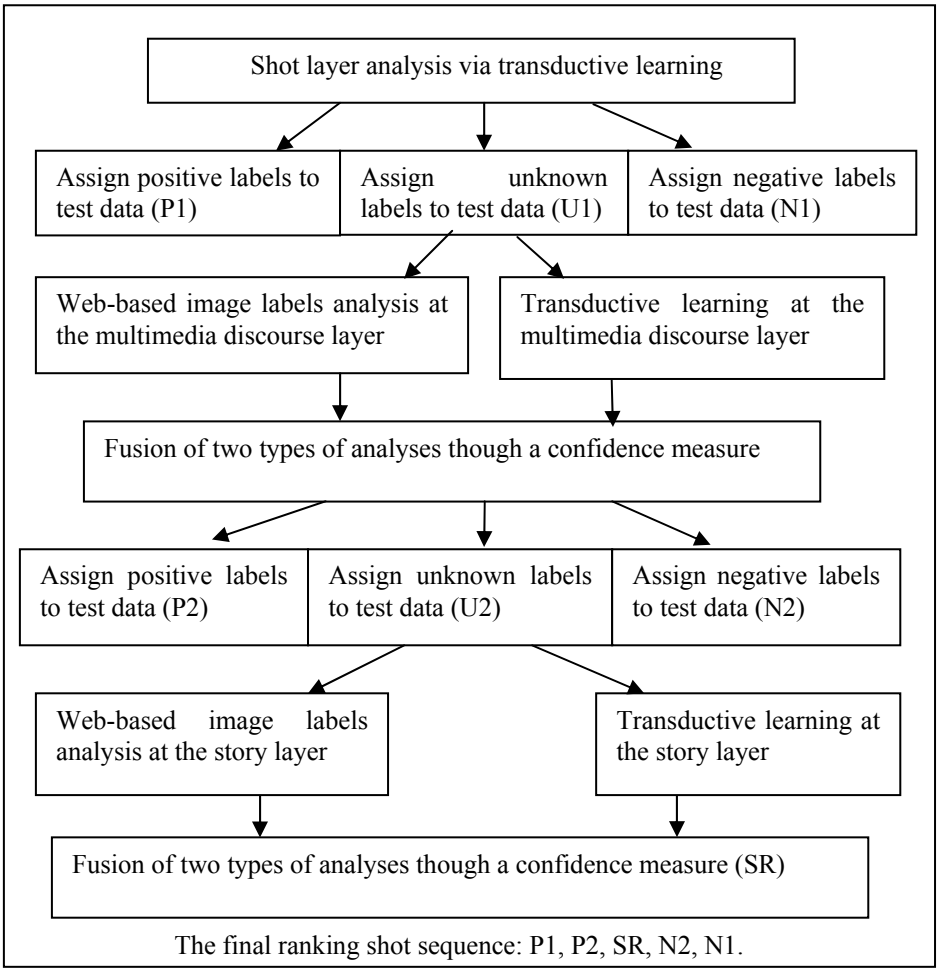

Fig. 10. The architecture of the system

Overall, through the multi-resolution analysis, Equation (3) can be expressed as:

$$P(C_x \mid S_t, S_v) = \begin{cases} P(C_x \mid S_v) & S \in P1 \text{ or } S \in N1 \\ P(C_x \mid S_{MD_t}) & S \in P2 \text{ or } S \in N2 \\ P(C_x \mid S_{ST_t}) & \text{Otherwise} \end{cases} \quad (4)$$

The term $P(C_x \mid S_{unit})$ can be computed from the training data in the cluster using the following procedure where S_{unit} in the multi-resolution framework can be S_v, S_{MD_t}, S_{ST_t}.

$$P(C_x \mid S_{unit}) = \frac{NumofTrainingShotsWith(C_x)\text{int}heCluster}{NumOfTrainingShotsIntheCluster} \quad (5)$$

Generally, there are two important assumptions in the probability theory. One is the "Law of large numbers", and the other is that training data needs to cover all the cases in test data; otherwise, we need some form of a smoothing method to estimate the probability of "unseen" cases.

Because our inference is based on the clustering results and some clusters may include very few training data, which may violate the "Law of large numbers" in the probability inference, we have to add a variable: confidence index (CI) to partially tackle the problem. Suppose T is the number of training data in a cluster and α is a predefined threshold, CI for that cluster can be computed as follows:

$$CI = \begin{cases} Log_{(\alpha+1)}(1+T) & T<\alpha \\ Log_{(\alpha+1)}(1+\alpha) & \text{Otherwise} \end{cases} \quad (6)$$

This is because the law of large numbers is a theorem in probability that describes the long-term stability of a random variable. That is, given a sample of independent and identically distributed random variables with a large number (α) of observations, the average of these observations may approach and stay close to the population mean. Thus, we assign a high confidence to the result. On the other hand, if the number of observations is small, the average of such observations may be far from the population mean. We assign a low confidence to the result.

The final decision score is modified by the confidence score as follows:

$$Score(S) = CI * P(C_x \mid S) \quad (7)$$

where CI is the confidence index for the cluster that includes the test shot S.

As some clusters include only test data, we could not compute Equation (5). We adopt a multi-resolution analysis strategy and a web-based text smoothing approach to tackle the problem.

5. Visual analysis and transductive inference

At the shot layer, we capture the semantic of keyframes for each shot by using the low-level visual features as used in most other works. The visual features used includes 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 $f(f_1, f_2, f_3, ..., f_t)$. Such images are collected together by a transductive inference method.

In our design, the transductive inference is used to analyze both the visual and text features in our framework. It involves two stages. In stage 1, a series of clustering are applied as different inference hypotheses using an average-link clustering method [21]. Such a 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 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 the clustering process, one key aspect is the definition of similarity measure. We adopt the cosine similarity measure to compute the similarity between shot i and shot j as:

$$Cossim(i, j) = \frac{\sum_{k=1}^{t} (f_{ki} \bullet f_{kj})}{\sqrt{\sum_{k=1}^{t} f_{ki}^{\ 2} \bullet \sum_{k=1}^{t} f_{kj}^{\ 2}}} \quad (8)$$

Fig. 11. Two visual similar shots share the same concept "boat/ship"

The use of visual similarity measure is effective only for shots that are similar in contents and semantics such as in Figure 11. However, it tends to produce many misses (where shots with dissimilar visual content but share in semantic concepts such as the images in Figure 1) and false positive (vice versa). Figure 12 illustrate

examples of images with similar visual contents but different semantic concepts.

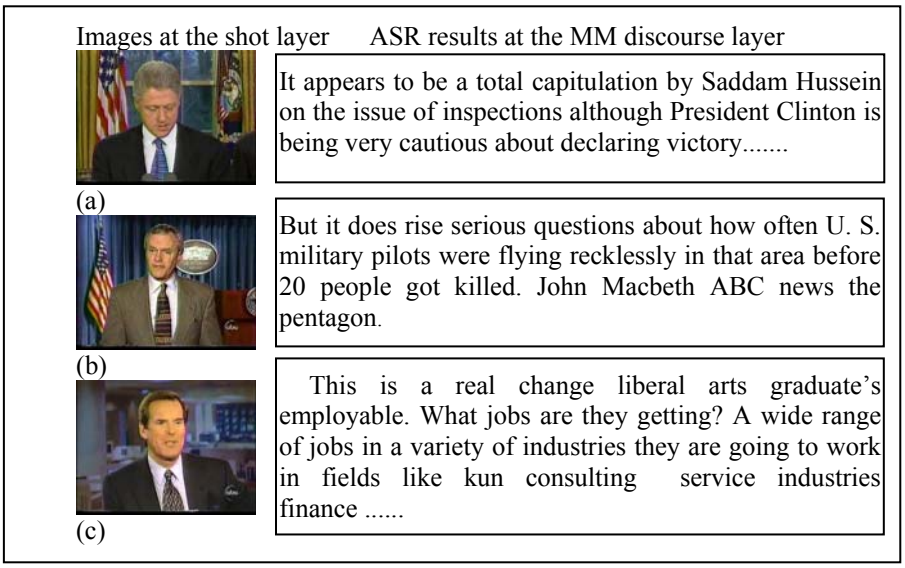

Fig. 12. Images with similar visual contents includes different semantic concepts

To tackle this problem, we employ the results of some useful mid-level detectors in news video such as the anchor person, commercial detectors [6], and add text constraints to purify the visual cluster result. Here, we leverage mid-level information at the shot level by adding the constraints of:

- Must-Link constraints
  If both shot i and shot j are detected as anchor person shots, then these two shots must be clustered together.
- Cannot-Link constraints
  If shot i and shot j are labeled as different genre types, such as when shot i is labeled as commercial and shot j as live reporting, then these two shots cannot be clustered together.

Because of the above visual middle level constraints, the images in Figure 12(a) of type speech and (c) type anchor person would not be clustered together. This could partially tackle the above problem. However, we can find that we still could not separate the visually similar speech type shots but depicting different speakers in Figure 12 (a) and (b). To tackle this problem, we need to incorporate the text cannot-link constraints. There are two cannot-link text constraints: one is from the MM discourse layer and the other come from the story layer. The text constraints eventually state that if the keyword representations at the MM discourse (or topics representations at the story) layer between shots i and j are very dissimilar, then shot i and shot j cannot be clustered together. The detail about text constraints and their use will be discussed in the next section. Together, the text constraints should be able to

resolve that the images in Figure 12(a) and 12(b) are different. The solution about this problem will be covered in the next section.

In stage 2 of transductive inference, a hypothesis is selected based on Vapnik Combined Bound [38] for determining the confidence of the series of clusters. That is, given a hypothesis $h \in H$ and unlabeled test set X_u, the predict risk of unlabeled samples is:

$$R_h(X_u) \le R_h(X_m) + \sqrt{(\frac{m+u}{u})(\frac{\tau + \log(C-1) + \ln\frac{1}{\delta}}{m})} \quad (9)$$

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

Based on the inference in the Vapnik bound, we can label the type 1 cluster as positive (P1) negative (N2) when the confidence is high, and unknown when the confidence is low. The unknown type 1 cluster together with Type 3 clusters are grouped as U1 (Unknown set).

The detail of transductive learning algorithm in our M3 framework is showed in Figure 13.

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) by using the cosine similarity measure 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; and $Sim(X_i, X_j)=1$ 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 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 calculate its Vapnik combined bound and select the hypothesis with the minimal Vapnik combined bound as our final clustering result.

Step 5: Label the test samples in the type 1 cluster by the training data in the same cluster, if the confidence is high and generate P, N and U sets of shots.

Fig. 13. A constraint based transductive learning algorithm

6. Text analysis at MM discourse and story layer

In this Section, we first introduce the inference at the MM discourse and story layers. We then discuss how to let different modalities support for each other. In addition, as the situation at the MM discourse is similar to that at the story layer. We first focus our discussion at the MM discourse and then highlight the differences at the story layer later.

6.1 Text inference at the MM discourse and story layer

After we obtain the text vector TV in Section 3 using Equation (1), we need to perform the term weighing. Here we employ a text weighting scheme based on tf.rf developed in [16] for text classification. Such a method measures the importance of a term based on its frequency (tf) and relevant frequency (rf). 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 only occurred in the test set. In order to tackle this problem, we leverage the web statistics to obtain other relevance information. The new weighting Equation is:

$$Weight(W_i) = tf * [CWI * \frac{\#(W_i, C_x)_{training}}{\#(W_i)_{training}} + (1 - CWI) * \frac{\#(W_i, C_x)_{web}}{\#(W_i)_{web}}] \quad (\mathbf{10})$$

We obtain $\#(W_i, C_x)_{training}$ and $\#(W_i)_{training}$ by counting the co-occurrence between terms Wi and Cx, and the occurrence of term Wi in the training data, respectively. We obtain $\#(W_i, C_x)_{web}$ by using the concept text description Cx together with term Wi as the query to Google search engine, and count the estimated number of hits that include the query terms. $\#(W_i)_{web}$ is computed in a similar manner. CWI is designed to balance the training data and web statistics. We estimate CWI in a way similar to CI using Equation (6), where T is the term frequency. That is if the term is of sufficiently high frequency in the training data, then the value of rf is based on the statistics in the training data. Otherwise, we will incorporate web statistics for smoothing. The resulting scheme considers all the words in the whole corpus instead of just words in the training data.

After term weighting, we obtain a new text vector TC. We use such new vectors as features to further cluster shots at the MM discourse layer by using the transductive learning algorithm in Figure 13. There are new two constraints for the MM discourse clustering. One is a must-link constraint from the shot layer. That is, if two shots i and j belong to one cluster in a visual feature clustering, it must be clustered together at the MM discourse layer too. This is the need from extracting text labels for images in the section 3. The other constraint is cannot-link constraint from the story layer,

which we will discuss in the next section. After employing the constraint based transductive learning algorithm, we will obtain a new shot cluster results $\{tcr_1, tcr_2, ..., tcr_n\}$ at the MM discourse layer. Similar to visual-based clustering at the shot layer, there still exist type 3 clusters and some type 1 clusters that include few training samples.

In order to process such unknown data, we again bring web statistics into our framework. Given a test shot S where $S \in U1$, and the cluster tcr_j that includes the test shot S, we use the text vector TC as an initial label vector of S. The semantic concept inference of S is defined as follows:

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

where

$$P_{Corpus}(C_x \mid S_{unit}) = \frac{NumofTrain\ ingShotsWi\ th(C_x)In(tcr_j)}{NumofTrain\ ingShotsIn\ (tcr_j)} \quad (12)$$

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

We obtain # (C_x,TC_i), and # (TC_i) in a similar manner as in Equation (10). The confidence index (CI) is defined as Equation (6).

At the MM discourse layer, the text inference is carried out as follows:

a) If $Score(C_x \mid S_{MD_t}) > \alpha$, we label it as positive data and put it into the P2 shot set.
b) If $Score(C_x \mid S_{MD_t}) < \delta$, we label it as negative data and put it into the N2 shot set.
c) Otherwise, we assign an unknown label to it and put it into U2 set for the story layer inference.

The difference between the story layer analysis and the MM discourse analysis is listed as follows:

1) The test shot S for story layer analysis belongs to U2 instead of U1 set.
2) The analysis target is topics instead of keywords.
3) After performing the transductive learning, we put all the results into the SR set and rank the results based on the value of $Score(C_x \mid S_{st})$.

6.2 Different modalities support for each other

Multimedia refers to the integration of different modalities. This statement has been reflected in one of the SIGMM grand challenges [28]: "A third facet of integration and adaption is the emphasis on using multiple media and context to improve application performance."

In our framework, there are two strategies to let different modalities at different resolutions support each other. One is the multi-resolution inference as outlined in Figure 10; while the other is constraints from different resolutions. Figure 12 illustrates the importance of such constraints. If we measure the similarity between two shots by global visual features alone, three images may have some degrees of similarity as shown in Figure12. However, when we consider its context text information, we can know that Figure12 (a) is related to the concept "Clinton" and the others are irrelevant to "Clinton". Such an example demonstrates the importance of constraints from different resolutions. In our framework, when performing higher resolution analysis, we bring in cannot-link constraints from the lower resolution to leverage the higher resolution analysis. When performing the lower resolution analysis, we incorporate the must-link constraints from the higher resolutions such that the shots clustered by a higher resolution shot layer analysis must be put in the same cluster at the low resolution analysis. From the above two strategies, we attempt to separate images with different semantic concepts but similar appearance and group images with the same semantic content but diverse appearance.

In order to tackle the problem in Figure 12, we add text constraints to purify the higher resolution clustering results. The text constraints come from the measure of homogeneity of text semantics.

There is one MM discourse layer text constraint for visual-based shot clustering, where the text-based Cannot-Link constraint is defined as:

Given two shots S(i) and S(j) with high visual similarity, if $Sim_{MD}[S(i), S(j)] < \delta_1$ then shots i and j cannot be clustered together, where MD is the text similarity at the MM discourse layer. In other words, if the similarity between two shots based on text analysis at the MM discourse layer is not sufficiently high, then the two shots cannot be clustered together.

In order to compute the above similarity, we built a word vector for each image at the MM discourse layer. The word vector is composed of all the non-stop words from ASR transcripts. As different word vectors may express the same concept, we propose a new web-based concept similarity measure. Such a method can assign a high similarity score for those word vectors with few or even no-overlapping words. On the other hand, if there is high word overlapping between two word vectors, such a method will be assigned a high similarity score too.

The definition of such a similarity measure is:

$$Sim_{unit}(T1, T2) = 1 - |P_{web}(C_x \mid T1) - P_{web}(C_x \mid T2)| \quad \textbf{(14)}$$

where T1, T2 are text feature vectors, which is made of keywords and topics at the MM discourse or story layers. C_x is the word from the concept text descriptions. We obtain $P_{web}(C_x \mid T)$ in Equation (14) as follows:

$$P_{web}(C_x \mid T) = \frac{\#(C_x, T)}{\#(T)} \quad \textbf{(15)}$$

We obtain # (C_x,T), # (T) in a similar manner as in Equation (10). Because there is a limitation on the number of terms in the query for most search engines, we employ the method in Equation (10) to select a few dominant terms in the text feature vector as query.

At the story layer, there is a similar constraint for visual-based clustering at the shot layer and keyword vector based clustering at the MM discourse layer. That is, if $Sim_{st}[S(i), S(j)] < \delta_2$ then the two shots cannot be clustered together at the shot and MM discourse layer respectively. Figure 14 shows the importance of the story layer constraints.

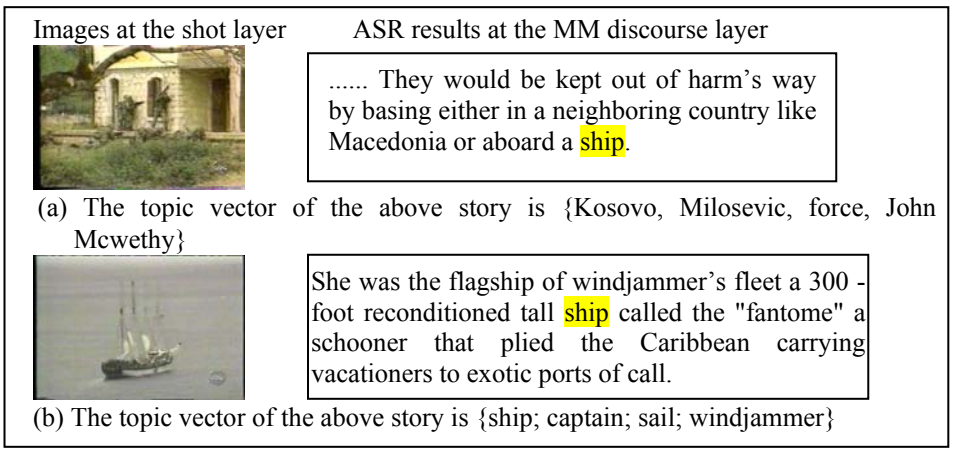

Fig. 14. An example of story information at detecting concept "boat/ship".

In Figure14, we find that the ASR transcripts of both shots (a) and (b) contain the text clue "ship" and it is above the threshold in Equation (1) in their related visual cluster, hence "ship" is assigned as a label to both keyframes. This causes these two shots to have a high similarity when we perform the MM discourse analysis. However, if we consider the context information at the story layer based on their topic vectors, we find that Figure14 (a) is not irrelevant to concept "boat/ship".

7. Experiment

In this section, we first introduce the test-bed and measurement of the experiments. We then compare different combinations of text retrieval and classification methods. Finally, we compare our M3 transductive model with the reported systems.

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 TECVID 2004 semantic concept task. The concepts are

"boat", "Albright", "Clinton", "train", "beach", "basket scored", "airplane takeoff", "people walking/running", "physical violence" and "road".

The performance of the system is measured using the mean average precision (MAP) based on the top 2000 retrieved shots for all ten concepts. This is the same as the evaluation used in TRECVID 2004. The MAP combines precision and recall into one performance value. Let $p^k = \{i_1, i_2, ..., i_k\}$ be a ranked version of the answer set A. At any given rank k, let $R \cap p^k$ be the number of relevant shots in the top k of p, where R is the total number of relevant shots. Then MAP for the ten concepts C_i is defined as:

$$MAP = \frac{1}{10}\sum_{C_i=1}^{10}[\frac{1}{R}\sum_{k=1}^{A}\frac{R \cap p^k}{k}\varphi(i_k)] \quad (16)$$

where the indictor function $\varphi(i_k) = 1$ if $i_k \in R$ and 0 otherwise. Because the denominator k and the value of $\varphi(i_k)$ are dominant, it can be understood that this metric favors highly ranked relevant shots.

7.1 Test1: Comparison on use of text features

We first evaluate the use of purely text feature in concept detection. We investigate different combinations of text retrieval and classification methods. For each method, we consider the scope of text features for the shot to be: (a) within the shot boundaries; (b) within the MM discourse boundaries; and (c) within the story boundaries. The text semantic analysis belongs to two methods. One is text classification, which we adopt the SVM^{light} [32] as the classifier. The other is text retrieval, which we adopt a state of the arts retrieval system [7] with query expansion techniques using external knowledge. For completeness, we also explore the combination of both methods using the following equation:

$$Score(S) = \alpha * Score_{IR}(S) + (1-\alpha) * Score_{TC}(S) \quad (17)$$

where IR is the score of retrieval method and TC is the score of the corresponding classification method.

Figure 15 lists the results based on text classification and retrieval at the shot, MM discourse and story layer respectively. We use different value of α range from 0 to 1. From the Figure, we can derive the following observations:

- The systems based on the MM discourse boundaries perform the best for both classification and retrieval methods. The main reason is that systems based on the shot boundaries could obtain only fragmented text clues; whereas systems based on the story boundaries could cover a large number of shots and thus could obtain higher recall, but lower precision. However, the MAP measure pays more attention to precision than recall.

- The performance of text retrieval system is superior to that of the text classification system. This is because we usually face the sparse training data problem in TRECVID data [24] and text retrieval method tends to perform better than text classification method under such circumstance.
- Although we tried different setting for the combinations of text classification and retrieval method, no combination could outperform the text retrieval systems. On the other hand, the performance of some combinations may be worse than the results from the text classification system. This suggests that if we want to combine different text analysis methods, we have to know in detail the strengths and weaknesses of different methods.

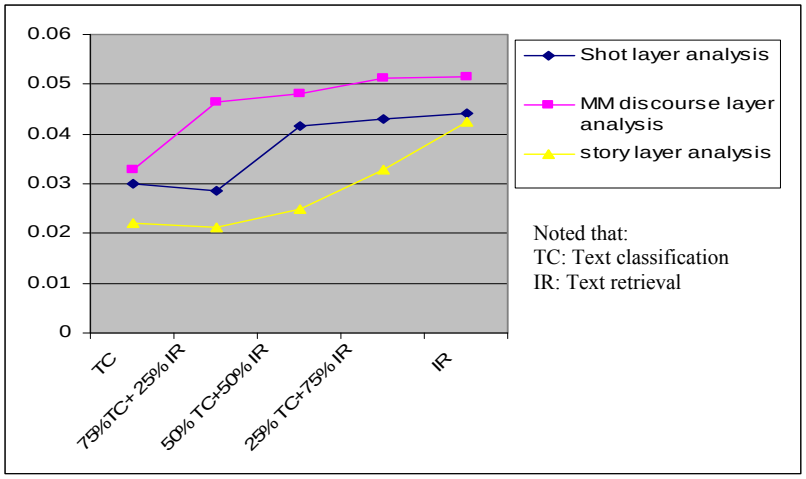

Fig. 15. The results of combination two types of text analysis

7.2 Test2: Multi-resolution multimodal analysis

Next, we employ the text retrieval method in combination with multimodal features in a M3 transductive framework as discussed in V and VI. In particular, 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. In order to compare our results with other state-of-the-arts systems, we tabulate the results of all reported systems that have completed all the ten concepts in Figure 16.

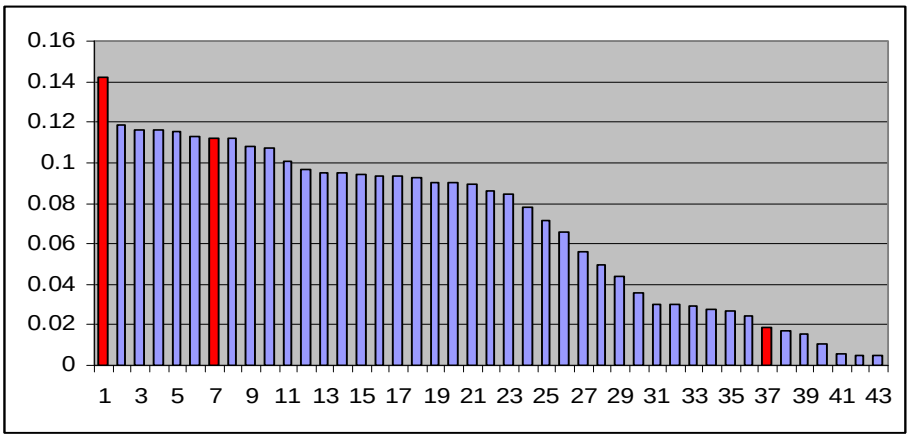

Fig. 16. The comparison with other systems in TRECVID.

From the Figure, we can observe that if using only the shot layer visual analysis without text, we could achieve only very low MAP of 0.024, which is much lower than that achievable using the text retrieval method (See Figure 15). This shows that the use of text will help visual analysis. This demonstrated in run (b) that incorporates text semantics at the MM discourse layer, in which we could achieve a substantially improved result at MAP of 0.112. However, the best result is achieved when we perform full multi-resolution analysis at the shot, MM discourse and story level, with a MAP of 0.142.

Compare to Figure 15, we find that the performance of our M3 transductive inference is significantly better than that of purely text analysis. This is because:

- Visual analysis could support text analysis. For example, if we were to rely just on text analysis, without visual clustering at the shot level to group visually relevant shots we would have captured some false positive shots for such as those illustrated in Figure 4, and missed some relevant such shots as shown in Figure 6(c).
- Our M3 transductive framework provides a novel solution to combine the classification and retrieval methods so that we could capture both relevant test shots in Figure 9(b) and (c).

From the Figure16, we also observe that our three combinations of systems ranked as 1st, 7th, and 37th. Compare to the best reported system ranked 2nd in Figure 16, our M3 transductive framework achieved more than 19% improvement in MAP performance. This is because most current systems are single resolution (shot), single source (training data) and multi-modality fusion methods under a supervised inductive inference framework. Usually there are at least two problems in such a framework.

- In most current systems, it is difficult to allow the evidences from different modalities to support each other.

- The performance of such a supervised inductive inference is highly dependent on the size and quality of training data. If the quality of training data is not good, the performance of the systems will decline significantly.

In our design, we propose a multi-resolution model to tackle the first problem. It emphasizes on using multi-modality features and their context to improve the performance. Our multi-source transductive can partially tackle the second problem by analyzing the data distribution between training and test data and integrating the external information sources under a retrieval framework. It is partially effective when the training data is not effective.

8. Conclusion and future work

Although research on semantic concept detection has been carried out for several years, the study on analyzing text semantics for concept detection has been relatively recent. This paper outlines a M3 transductive learning model. In the multi-resolution model, we emphasized on the techniques to employ different types of text semantics at different resolutions to support visual concept detection. In our multi-source transductive model, we proposed a novel approach to combine the classification and retrieval methods. The experimental results demonstrated that our approach is effective.

The work is only the beginning. Further research can be carried out as follows:

- We will further study on how to include visual information to improve the performance in extracting topics at the story layer.
- We will further improve the performance on building linkage between visual features and ASR transcript.
- We should further study visual analysis and improve the performance, because text analysis is just a supplementary source to support visual analysis.

Reference

[1] A.Amir et al, "IBM research TRECVID 2003 video retrieval system", available at: http://www-nlpir.nist.gov/projects/tvpubs/ tv3.papers/

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

[3] M. Campbell et al, "IBM Research TRECVID-2006 Video Retrieval System", Proceedings of TRECVID 2006, Gaithersburg, MD, November 2006 available at: http://www-nlpir.nist.gov/projects/ tvpubs/

[4] S.F. Chang, "Advances and Open Issues for Digital Image/Video Search", Keynote Speech at International Workshop on Image Analysis for Multimedia Interactive Services, available at : http://www.ee.columbia.edu/%7Esfchang/papers/talk-2007-06-WIAMIS-Greece-print.pdf

[5] 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

[6] T.S.Chua et al, "TRECVID 2004 Search and Feature Extraction Task by NUS PRIS", Proceedings of (VIDEO) TREC 2004, Gaithersburg, MD, November 2004
[7] H. Cui, K. Li, R. Sun, T.-S. Chua and M.-Y. Kan. National University of Singapore at the TREC-13 Question Answering Main Task. Proceeding of TREC-13, 2004 available at: http://lms.comp.nus.edu.sg/papers/Papers/text/trec04-Notebook.pdf
[8] P.Duygulu, K.Barnard, J.de Freitas, and D.Forsyth. "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary" Proceeding of European Conference on Computer Vision, volume 4 pp.97-112, 2002
[9] J.L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System." Speech Communication, 37 (1-2) pp89-108, 2002.
[10] U. Hahn "Topic parsing: accounting for text macro structures in full-text analysis" Information Processing and Management, 26 (1): pp.135-170, 1990
[11] A.Hauptmann, et al, "Informedia at TRECVID 2003: Analyzing and Searching Broadcast News Video", Proceedings of (VIDEO) TREC 2003, Gaithersburg, MD, November 2003, available at: http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html#2003
[12] M. A. Hearst. "Context and Structure in Automated Full-Text Information Access". PhD thesis, University of California at Berkeley, 1994.
[13] http://www.ipoque.com/media/internet_studies/internet_study_2007
[14] J.Jeon, V.Lavrenko, and R.Manmatha, "Automatic image annotation and retrieval using cross-media relevance models", In proceedings of the 26th Annual International ACM SIGIR Conference pp. 119 - 126, 2003
[15] D. Jurafsky and J. H. Martin, "Speech and language processing", published by Prentice-Hall Inc, 2000.
[16] M. Lan , C L Tan and H B Low "Proposing a new term weighting scheme for text categorization", Proceedings of the 21st National Conference on Artificial Intelligence, AAAI-2006
[17] Y. Li "Multi-resolution analysis on text segmentation", Master thesis, National University of Singapore, 2001
[18] C.Y. Lin, B. Tseng, J.R. Smith " Video Collaborative Annotation Forum: Establishing Ground-Truth Labels on Large Multimedia Datasets", 2003 available at: http://www-nlpir.nist.gov/projects /tvpubs / tv.pubs.org.html#2003
[19] C.Y.Lin, "Robust Automated Topic Identification" Ph.D. Thesis, University of Southern California 1997
[20] Y.Lin, "TMRA-Temporal Multi-resolution Analysis on Video Segmentation", Master thesis, National University of Singapore, 2000.
[21] 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
[22] J.R. Kender, et al, "IBM Research TRECVID 2004 Video Retrieval System", Proceedings of (VIDEO) TREC 2004, Gaithersburg, MD, November 2004
[23] Y. Marchenko ,T.S. Chua and R. Jain "Transductive inference using multiple experts for brushwork annotation in paintings domain ", Proceedings of the 14th ACM Multimedia, pp. 157 – 160, 2006
[24] M. R. Naphade, J. R. Smith , "On the detection of semantic concepts at TRECVID", Proceedings of the 12th ACM Multimedia, pp660-667, 2004
[25] C. D. Paice "Constructing literature abstracts by computer: Techinques and prospects", Information Processing and Management, 26 (1) pp. 171-186, 1990
[26] G.J. Qi, X.S.Hua, Y. Song, J.H.Tang, H.J. Zhang, "Transductive Inference with Hierarchical Clustering for Video Annotation" Proceedings of International Conference on Multimedia and Expo, pp. 643 – 646, 2007
[27] N.C. Rowe "Inferring depictions in natural language captions for efficient access to picture data", Information Process & Management Vol 30 No 3. pp379-388,1994

[28] L. A. Rowe and R. Jain, "ACM SIGMM Retreat Report on Future Directions in Multimedia Research", ACM Transactions on Multimedia Computing, Communications, and Applications, Volume 1, issues 1, pp3-13, 2005
[29] T.Shibata S.Kurohashi, "Unsupervised topic identification by integrating linguistic and visual information based on Hidden Markov Models", Proceedings of the International Association for computational linguistics conference pp.755-762, 2006
[30] M.Slaney, D.Ponceleon, J.Kaufman, "Multimeida Edges: Finding Hierarchy in all Dimensions", Proceeding of the 9th International Conference on Multimedia, pp. 29-40,2001
[31] C. G.M. Snoek, M. Worring, J. C. V. Gemert, J.M. Geusebroek, and A. W.M. Smeulders, "The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia", Proceedings of the 14th ACM Multimedia, pp.421 – 430, 2006.
[32] SVM^{light}, available at: http://svmlight.joachims.org/
[33] TRECVID (2005-2006): "Online Proceedings of the TRECVID Workshops", available at http://www-nlpir.nist.gov/projects/tvpubs/ tv.pubs.org.html
[34] V.N.Vapnik, "Statistical learning theory", Wiley Interscience New York. pp120-200,1998
[35] J. Z. Wang and J Li , "Learning-Based Linguistic Indexing of Pictures with 2-D MHMMs", Proceedings of the 10th International Conference on Multimedia, pp. 436-445,2002
[36] K.W.Wilson and A.Divakaran, "Broadcast Video Content Segmentation by Supervised learning",
[37] J. Yang, A. Hauptmann, M. Y. Chen, "Finding Person X: Correlating Names with Visual Appearances", Proceedings of International Conference on Image and Video Retrieval (CIVR'04), Dublin City University, Ireland, July 21-23, 2004
[38] R. E. Yaniv, and L.Gerzon, "Effective Transductive Learning via PAC-Bayesian Model Selection.", Technical Report CS-2004-05, IIT, 2004.
[39] J. Yuan et al "Tsinghua University at TRECVID 2004: Shot Boundary Detection and High-Level Feature Extraction", Proceedings of TRECVID 2004, Gaithersburg, MD, November 2004 http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html