

A Multi-resolution Multi-source and Multi-modal (M3) Transductive Framework for Concept Detection in News Video

Wang Gang

(M.Sc. National University of Singapore)

A THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
THE SCHOOL OF COMPUTING
NATIONAL UNIVERSITY OF SINGAPORE

2008

ACKNOWLEDGMENTS

The completion of this thesis would not have been possible without the help of many people to whom I would like to express my heartfelt appreciation.

First and foremost, I would like to thank my supervisor Chua Tat Seng for his great guidance and timely support in my PhD study. I am grateful that he gave me the invaluable opportunity to join the semantic concept detection project. This project revealed to me the significant gap between the performance of classical theories in real-life corpora and the user's information need. This forces me to learn how to think. During my study in NUS, Prof Chua was very kind and patient, supportive and encouraging, teaching me proper ways of doing research and helping me shape and reshape ideas and presentations in this thesis. It has been a great pleasure to have the opportunity to work with such a true expert in the field.

I would like to thank my other thesis committee members, A/P Kan Min Yen and A/P Ng Hwee Tou, for their invaluable assistance, feedback and patience at all stages of this thesis.

During my study in NUS, many Professors imparted me knowledge and skills, gave me good advices and help. I would like to thank Prof. Yuen Chung-Kwong, Prof. Tan Chew Lim, Prof. Kankanhalli, Mohan, Prof. Ooi Beng

Chin, A/P Yeo Gee Kin, A/P Wang Ye, A/P Leow Wee Kheng, A/P Sim Mong Cheng, Terence, A/P Sung Wing Kin, Ken. Thanks to all of them.

Thanks are also due to all the persons working and study in the multimedia search lab. Especially thank to Dr. Zhao Ming for sharing with me his knowledge and providing me some useful tools for my projects. Dr. Feng Hua Ming, Dr. Zhao Yun Long, Dr. Ye Shi Ren, Dr. Cui Hang, Dr. Lekha Chaisorn, Dr. Zhou Xiang Dong, Qiu Long, Mstislav Maslennikov, Xu Hua Xin, Shi Rui, and Yang Hui, for spending their time to discuss the project with me.

Thanks are also due to the School of Computing and Department of Computer Science for providing me with a scholarship and excellent facilities and environment for my research work.

Finally, the greatest gratitude goes to my parents for loving me, supporting me and encouraging me to be the best that I could be in whatever endeavor I choose to pursue.

ABSTRACT

We study the problem of detecting concepts in news video. Some existing algorithms for news video concept detection are based on single-resolution (shot), single source (training data), and multi-modal fusion methods under a supervised inductive inference; while many others are based on a text retrieval with visual constraints framework. We identify two important weaknesses in the state-of-the-art systems. One is on the fusion of multimodal features; and the other is on capturing the concept characteristics based on training data and other relevant external information resources.

In this thesis, we present a novel multi-resolution, multi-source and multi-modal (M3) transductive learning framework to tackle the above two problems. In order to tackle the first problem, we perform a multi-resolution analysis at the shot, multimedia discourse and story levels to capture the semantics in news video. The most significant aspect of our multi-resolution model is that we let evidence from different modal features at different resolutions support each other. We tackle the second problem by adopting a multi-source transductive inference model. The model utilizes the knowledge not only from training data but also from test data and other online information

resources. We first perform transductive inference in order to capture the distributions of data from both the observed (test) and specific (training) cases to train the classifiers. For those test data that cannot be labeled by transductive inference, our multi-source model brings in web statistics to provide additional inference on text contents of such test data to partially tackle the problem.

We test our M3 transductive model to detect semantic concepts using the TRECVID 2004 dataset. Experiment results demonstrate that our approach is effective.

TABLE OF CONTENTS

Chapter 1 Introduction	1
1.1 Motivation.....	1
1.2 Problem statement	4
1.3 Our approach.....	9
1.4 Main contributions.....	10
1.5 Organization of the thesis	11
Chapter 2 Background and Literature Review	13
2.1 Background	14
2.1.1 What is the concept detection task?	14
2.1.2 Why do we need detect semantic concepts?	15
2.2 Visual based semantic in the concept detection task	17
2.2.1 Low level visual features	18
2.2.2 Mid-level abstraction (detectors).....	20
2.3 Text semantics in the concept detection task	23
2.4 Fusion of multimodal features	28
2.5 Machine learning in the concept detection task	32
2.5.1 Supervised inductive learning methods	33
2.5.2 Semi-supervised learning	34
2.5.3 Transductive learning	35
2.5.4 Comparison of the three types of machine learning	37
2.5.5 Domain adaptation.....	39
2.6 Multi-resolution analysis.....	41
2.7 Summary	42
Chapter 3 System architecture	45
3.1 Design consideration.....	45
3.1.1 Multi-resolution analysis	45
3.1.2 Multiple sub-domain analysis.....	51
3.1.3 Machine learning and text retrieval.....	55
3.2 System architecture	58

Chapter 4 Multi-resolution Analysis	63
4.1 Multi-resolution features.....	63
4.1.1 Visual features.....	64
4.1.2 Text features.....	68
4.1.2.1 The relationship between text features and visual concepts.....	70
4.1.2.2 Establish the relationship between text and visual concepts.....	74
4.1.2.3 Word weighting.....	77
4.1.2.4 Similarity measure.....	78
4.2 The multi-resolution constraint-based clustering	84
Chapter 5 Transductive Inference	88
5.1 Transductive inference	88
5.2 Multiple sub-domain analysis.....	97
5.3 Multi-resolution inference with bootstrapping.....	100
Chapter 6 Experiment	102
6.1 Introduction of our test-bed	103
6.2 Test 1: Concept detection via single modality analysis	105
6.2.1 Concept detection by using text feature.....	105
6.2.2 Concept detection by visual feature alone.....	107
6.3 Test 2: Multi-modal fusion.....	109
6.4 Test 3: Encode the sub-domain knowledge.....	110
6.5 Test 4: Multi-resolution multimodal analysis.....	112
6.5.1 A baseline multi-resolution fusion system.....	112
6.5.2 Our proposed approach.....	116
6.6 Test 5: The comparison of M3 model with other reported systems.....	119
Chapter 7 Conclusion and Future Work	123
7.1 Contributions	124
7.1.1 A novel multi-resolution multimodal fusion model	124
7.1.2 A novel multi-source transductive learning model.....	125
7.2 Limitations of this work.....	126
7.3 Future work	127
Bibliography	131

LIST OF FIGURES

Figure 1.1: The concept “boat/ship” with different shapes and different colors.....	6
Figure 2.1: An example of detecting the concept “train”	15
Figure 2.2: False alarms and misses when performing matching using low-level features to detect the concept “boat/ship”.....	17
Figure 2.3: Captions: Philippine rescuers carry a fire victim in March 19 who perished in a blaze at a Manila disco.....	24
Figure 2.4: The association between faces and names in videos.....	25
Figure 2.5: The frequency of Bill Gates visual appearances in relation to his name occurrences.....	26
Figure 2.6: Different person X with different time distributions.....	26
Figure 2.7: The sentence separated by three shot boundaries causes the mismatch between the text clue and the concept “Clinton”.....	31
Figure 3.1: The ability and limitation of visual feature analysis at the shot layer.....	46
Figure 3.2: The ability and limitation of text analysis at the MM discourse layer.....	48
Figure 3.3: An example text analysis at the story layer.....	49

Figure 3.4: The distributions of positive data of 10 concepts from TRECVID 2004 in the training set.....	52
Figure 3.5: The characteristics of data from different domains may be different.....	54
Figure 3.6: An example of detecting concept “boat/ship” using two text analysis methods.....	57
Figure 3.7: The bootstrapping architecture.....	59
Figure 3.8: The multi-resolution transductive learning framework for each sub-domain data set.....	60
Figure 4.1: Examples of anchorperson shots in a news video.....	65
Figure 4.2: Commercial shots for a product in a news video.....	66
Figure 4.3: Examples of CNN financial shots.....	67
Figure 4.4: Examples of sports shots.....	67
Figure 4.5: The text clue “Clinton” co-occurred with the visual concept.....	71
Figure 4.6: An example of when the text clues appears, but the concept did not occur.....	71
Figure 4.7: An example of when the visual concept occurred, but we could not capture the text clues.....	72
Figure 4.8: Keyframes from shots and the topic vector in the story.....	73

Figure 4.9: An example of labeling a visual cluster by text information	75
Figure 4.10: An example where no text labels could be extracted from the image cluster.....	76
Figure 4.11: Two non-overlapping word vectors indicating a same concept “Clinton”	79
Figure 4.12: The Google search results using {Erskine Bowles, president, Lewinsky, white house} as a query.....	81
Figure 4.13: The Google search results using {Erskine Bowles, president, Lewinsky, white house} and “Clinton” as a query.....	82
Figure 4.14: The Google search results using {Clinton, Israeli, Prime Minister, Benjamin Netanyahu} and “Clinton” as a query.....	82
Figure 4.15: The Google search results using {Clinton, Israeli, Prime Minister, Benjamin Netanyahu} as a query.....	83
Figure 4.16: An example of using the cannot-link text constraints to purified visual shot clustering results.....	85
Figure 5.1: A traditional query expansion method that uses Web statistics.....	93
Figure 5.2: An example of our text retrieval model.....	95
Figure 5.3: A constraint based transductive learning algorithm.....	99
Figure 5.4: Our bootstrapping algorithm.....	101

Figure 6.1: The results of combining two types of text analysis 106

Figure 6.2: Two types of machine learning methods that detect concepts by using visual features alone.....108

Figure 6.3: Concept detection by using single modality versus multi-modality110

Figure 6.4: The systems with and/or without the use of sub-domain knowledge.....111

Figure 6.5: Results of single resolution fusion vs. multi-resolution fusion without using sub-domain knowledge.....114

Figure 6.6: Multi-resolution systems with and / or without sub-domain knowledge.....115

Figure 6.7: The result based on the shot layer analysis and different combinations of multi-resolution analysis.....117

Figure 6.8: Two types of multi-resolution fusion systems.....118

Figure 6.9: Comparison with other reported systems in TRECVID.....120

Figure 6.10: An example of our M3 transductive framework on the concept “train”.....122

Figure 7.1: Repeatedly labeling for similar images in the different videos....128

LIST OF TABLE

Table 2-1: Comparison of three types of learning approaches.....	38
Table 6-1: Ten semantic concepts used in TRECVID.....	104
Table 6-2: The setting parameters of the linear combination.....	113

Chapter 1

INTRODUCTION

1.1 Motivation

Our entrance into the information age has had significant impacts on our society. We have systematized the production of knowledge and amplified our brainpower. To use an industrial metaphor, we now mass-produce knowledge from information and this knowledge is the driving force of our economy. Thus, Naisbitt [1982] believed that the most important strategic resource is information. Therefore, more and more people have paid attention to the value of information and information dissemination. With the increasing value of information and the popularization of the Internet, the volume of information has been soaring ceaselessly. Based on the research by Lyman and Varian [2003], the world produced about five exabytes of new information in 2002, which is equivalent in size to the information contained in 37,000 new libraries

the size of the book collections in the US Library of Congress. Furthermore, the speed of producing information has the growth rate of 50% year on year.

With such a rapid growth of information, we can find that multimedia data play a more and more important role. In the 1990's, the major use of a computer is to count the numbers and process text data. It is reported in China Internet Network Information Center [1997] that in 1997, text-based web browsing, e-mail, ftp and telnet accounted for about 78.3%, 10.7%, 8.4 % and 1.6% of Internet traffic respectively. From such statistics, we can observe the fact that at that time the major modality in Internet traffic is text information. However, the advancement in computer processor, storage and the growing availability of low-cost multimedia recording devices have led to the explosive growth of multimedia data. Evans [2003] claimed that for BBC1 & BBC2 alone there were 700 hours of TV programs transmitted per week. Furthermore, in BBC alone, there were over 750,000 hours of television programs in the archive. It was reported in [Chang, 2007] that there are 31 million hours of TV programs produced each year. Since P2P was invented and widely used on Internet applications, more and more multimedia data has been transferred from one computer to another via the web. The statistics from an Internet study¹ shows that about 65% of Internet traffic was being taken up by transferring multimedia contents in 2007. Among them, about 73.79% is

¹ http://www.ipoque.com/media/internet_studies/internet_study_2007

video related content. With such huge volume of multimedia information, if such information is uncontrolled and unorganized, it becomes impossible to find them. In fact, researchers are even overwhelmed by the huge amount of technical data that it often takes more time to find out whether or not an experiment has been done than to do the experiment itself. Nasibitt and Aburdene [1990] claimed that: “we are drowning in information but starved for knowledge”.

In order to make use of such huge information, search engines like Google² provide a good solution to utilize text information resources. The success of text search engines whetted the appetite of users who hope to have similar abilities to search over large multimedia corpora. For example, in the early 21st century, many researchers such as [Chua et al. 2001] have a dream of building Video-On-Demand systems. Recently, news video retrieval sites such as Blinkx.com³ and Streamsage.com⁴ aim to aggregate news videos from multiple sources for retrieval. Such systems are based purely on the automatic speech recognition (ASR) text and are as effective as the quality of the ASR text. In particular, if the relevant video clips do not have the query text available, such video clips will not be retrievable. On the other hand, much false retrieval will occur for those irrelevant clips that contain the query text. Thus for effective

² <http://www.google.com>

³ <http://www.blinkx.com/>

⁴ <http://www.streamsage.com/>

management and retrieval of multimedia contents, we need to index the multimedia data at the higher semantic level, such as whether a shot contains person-X or object X etc. that frequently appear in queries. One example query is “find shots of Benjamin Netanyahu”. The target of our system is to find shots visually containing Benjamin Netanyahu in the given news video. In this example, the visual semantic concept is the visual appearance of Benjamin Netanyahu. However, it is impossible for humans to manually annotate concept X, as it is both error-prone and time consuming [Lin, Tseng, and Smith, 2003]. On average, the human annotator will use about 6.8x times that of the broadcast time to annotate news video properly. Therefore, there is an urgent need to automatically infer concept X.

1.2 Problem statement

The semantic concept detection task has attracted the attention of many researchers. One of the largest researcher communities to work under this topic is the TRECVID community [TRECVID, 2002-2007]. TRECVID is an annual benchmarking exercise, which encourages research in video information retrieval by providing a large video test collection, a set of topics, uniform methods for scoring performance, and a forum for organizations interested in comparing their results. The semantic concept detection task

began in 2002. The target of this task is to find whether the shot includes certain visual semantic concepts. Most participating groups have tackled the concept-detection task as either a shot-based supervised visual pattern classification problem [Naphade and Smith, 2004] or a text retrieval problem which combines text results with visual constraints [Yang et al. 2004]. In spite of these efforts, we are still far from achieving a good level of concept detection performance. Based on our analysis, we have identified two major weaknesses of current systems that should be addressed to enhance the performance.

- **Fusion of text and visual features**

Multimedia refers to the idea of integrating information of different modalities [Rowe and Jain, 2005] such as the combination of audio, text and images to describe the progress of news events in news video. As speech in news video is often the most informative part of the auditory source, we focus on the fusion of automatic speech recognition (ASR) text with visual features. However, there are errors in the ASR text and there often exist mismatches between text clues and visual contents at the shot layer [Yang et al. 2004]. On the other hand, it is very hard for detectors to use only visual features to detect whether such concepts exist in the shots. 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 1.1 shows examples of concept “boat” in news video with different shapes and colors. Thus, semantic concept detectors require a good fusion method to combine text and visual features. Although many efforts [TRECVID, 2002-2007] have been made, most of the existing systems fail to allow the evidence from text and visual features to support each other effectively.



Figure 1.1: The concept “boat/ship” with different shapes and different colors

- **Capturing the characteristic of the concepts via the training data and concept descriptions**

Many of the so-called concepts⁵ are abstract in that they focus on extracting the similarity of instances under these concept classes, while ignoring their differences. For the example in Figure 1.1, although different boats may have different colors and shapes, the boats have some common characteristics that are a watercraft of modest size designed to float or plane on water, and provide transport over it.

In general, there are two commonly used concept definition approaches. One is an example-based definition method, in which the examples are provided by the training data. Given a set of training data, the most widely used method is a supervised learning approach. However, such a type of learning requires the estimation of unknown function for all possible input items. This implies the availability of good quality training data, which must include the 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. One solution to obtaining 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

⁵ <http://en.wikipedia.org/wiki/Concept>

problem [Naphade and Smith, 2004]. The other concept definition method is a text description approach, where we use text to describe significant characteristics of the concept from its text description. For example, we can use “boat / ship: segment contains video of at least one boat, canoe, kayak, or ship of any type” to define the concept “boat / ship”. The concept text description is “boat / ship”. Thus, another widely used method to detect concepts is a text retrieval method such as [Hauptmann, et al., 2003, Yang et al., 2004, Chua et al., 2004, Campbell et al., 2006]. These methods regard words from concept text descriptions or some predefined keywords as the query and employ the text retrieval approach with query expansion techniques to capture the semantic concepts. However, the analysis in news video based only on text is effective only if the desired query concepts appear in both visual and text contents.

In general, we found that it is not easy to capture the characteristics of concept by using either type of definition methods. How we can make use of the knowledge from both definition methods to capture the characteristics of the concepts is an open problem.

1.3 Our approach

In this thesis, we propose a multi-resolution, multi-source and multi-modal (M3) transductive framework to tackle the above two problems. In our multi-resolution model, we first analyze different modal features at different resolutions such as the shot or story levels. When analyzing the evidence from each single resolution, we regard the evidence from other modalities at the other resolutions as contextual information or constraints to support the decision. Next, we fuse the evidence from different resolutions together according to the confidence. Based on such a framework, we allow the evidence from different modalities to support each other. In each resolution analysis, we adopt a transductive inference model. Such a model aims to capture the distributions of the training and test data well so that we have the knowledge to know when we can make an inference via training data. In order to tackle the limitation of the training data, our multi-source model brings the web statistics into the framework. Such web statistics are designed to capture the relationship between the text content in the test data and concept text descriptions via the web. Finally, we utilize the bootstrapping technique to make use of unlabeled data to boost the overall system performance. We test our M3 transductive model on the TRECVID 2004 dataset. The test results demonstrate that our M3 transductive framework is superior to the systems

based on text or visual information alone, and the reported multi-modal fusion frameworks.

1.4 Main contributions

In this thesis, we make the following contributions:

- **A novel multi-resolution multimodal fusion model**

Multimedia refers to the idea of integrating information of different modalities. As different modal features only work well in different temporal resolutions and different resolutions exhibit different types of semantics, we perform a multi-resolution analysis at the shot, multimedia discourse (or multi-sentence) 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 towards the multimedia discourse and story levels. More importantly, text and visual features in news video are coherent. In our multi-resolution multimodal fusion model, we let evidence from text and visual features support each other.

- **A novel multi-source transductive learning model with bootstrapping**

Different with traditional classifiers, the output of our novel multi-source transductive classifier has three possible states: positive, unknown and negative. The function of the new unknown state is similar to that of “0” between positive and negative numbers in mathematics. It suggests that in such cases, it is hard to assign a positive or negative label to these test data via the knowledge learned from the training data. To disambiguate test shots with unknown states, we integrate web statistics into the three transductive learners at different resolutions under our multi-resolution framework. Finally, we combine our M3 transductive learning framework with a bootstrapping technique to further process the test results with low confidence.

1.5 Organization of the thesis

The organization of this thesis is as follows:

Chapter 2 introduces background and related work about this topic. The chapter covers the background of concept detection; visual and text based inference; fusion of visual and text features, and machine-learning methods in the concept detection task.

Chapter 3 presents the architecture of multi-resolution, multi-source and multi-modal transductive learning framework. We provide a brief introduction of design consideration and the system architecture. The detailed discussion for each component will be covered in Chapters 4 and 5.

Chapter 4 covers the multi-resolution analysis at the shot, multimedia discourse, and story layer. We discuss the multi-resolution features, similarity measures and multi-resolution constraint clustering.

Chapter 5 discusses our multi-source transductive learning model. We first introduce the detail on transductive learning. We then expand our algorithm to using sub-domain knowledge. Finally, we combine our M3 transductive framework with a bootstrapping technique.

Chapter 6 reports the design of the experiments; measurements of the system performance and our experiment results with analysis.

Finally, Chapter 7 concludes the thesis with suggestions for future work.

Chapter 2

Background Studies and Literature Reviews

Semantic concept detection from multi-modal features enables high-level access to multimedia contents. In this chapter, we first introduce the background of concept detection. We then introduce the concept detection systems by using single modality such as visual or text features. Next, we report different strategies on the fusion of multi-modal features in the state-of-the-art systems. After that, different types of machine learning methods used in concept detection are covered. Finally, we introduce the background of multi-resolution analysis followed by a summary of the current problems with possible solutions.

2.1 Background

2.1.1 What is the concept detection task?

The purpose of a semantic concept detection task is to assign the appropriate semantic labels to a video clip based on visual appearance. Currently, the shot is the basic video unit in the benchmark TRECVID corpus. Figure 2.1 provides an example of the concept detection. Given a video shot, we can extract multi-modal features. In news video, the widely used multi-modal features are visual and text features. The text features come from the automatic speech recognizer, such as those shown in Figure 2.1. The visual features are color, texture, shape and so on. In addition, there are at least three types of knowledge that we can use in the semantic concept detection task. They are: knowledge from training data, knowledge from concept descriptions, and knowledge from external resources of information.

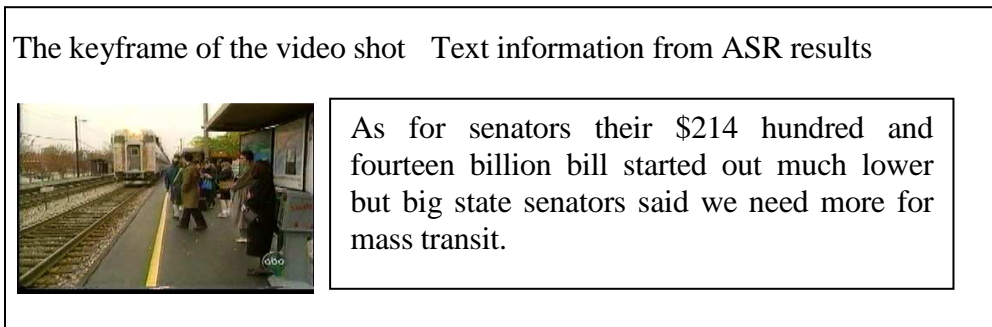


Figure 2.1: An example of detecting the concept “train”

More formally, the video concept detection task is defined as: given a set of predefined concepts $\vec{C} : [C_1, C_2 \dots C_n]$, develop a classifier to determine if the concept C_i appears visually in shot S_k .

2.1.2 Why do we need to detect semantic concepts?

Semantic concept detectors are very important and fundamental to multimedia retrieval. There are at least two reasons that we need to detect semantic concepts. The first reason is that most users tend to express their information needs in terms of semantic concepts. An example query from the TRECVID⁶ is “Find shots of one or more buildings with flood waters around it/them”. Given such a natural language query, the query analysis model [Chua et al.

⁶ <http://www-nlpir.nist.gov/projects/tv2004/topics/topics.2004.xml>

2004] can transfer the users' information need from such a query to a Boolean Expression: "building" + "flood waters". If we have detected such concepts, we can employ a traditional retrieval method [Yates and Neto, 1999] to satisfy such queries. The second reason is due to the difference between multimedia retrieval and traditional text retrieval. Conventional text retrieval systems [Salton and McGill, 1983] only deal with simple data types, such as strings or integers. However, multimedia retrieval systems cannot rely on a single modal feature analysis such as visual or text matching alone. Figure 2.2 illustrates the problem of using only single modal feature matching to detect the concept "boat/ship". Suppose Figure 2.2 (a) is a query image for the concept "boat". Although there is high similarity between Figures 2.2 (a) and (b) in the low level feature space, the concept "boat" does not occur in Figure 2.2 (b). On the other hand, there is a large variation in the low-level visual feature spaces between Figures 2.2 (a) and (c), but Figure 2.2 (c) includes the concept "boat/ship". Similarly, Figures 2.2 (d) and (e) demonstrate the cases of false alarms and misses by keyword matching from the ASR results alone. The multimedia community calls this gap between the high-level semantics and the discrimination power from low-level features as the semantic gap [Hauptmann, 2005]. Thus, one important motivation for concept detection is to fuse evidence from different modalities from multimedia corpora to bridge the semantic gap.

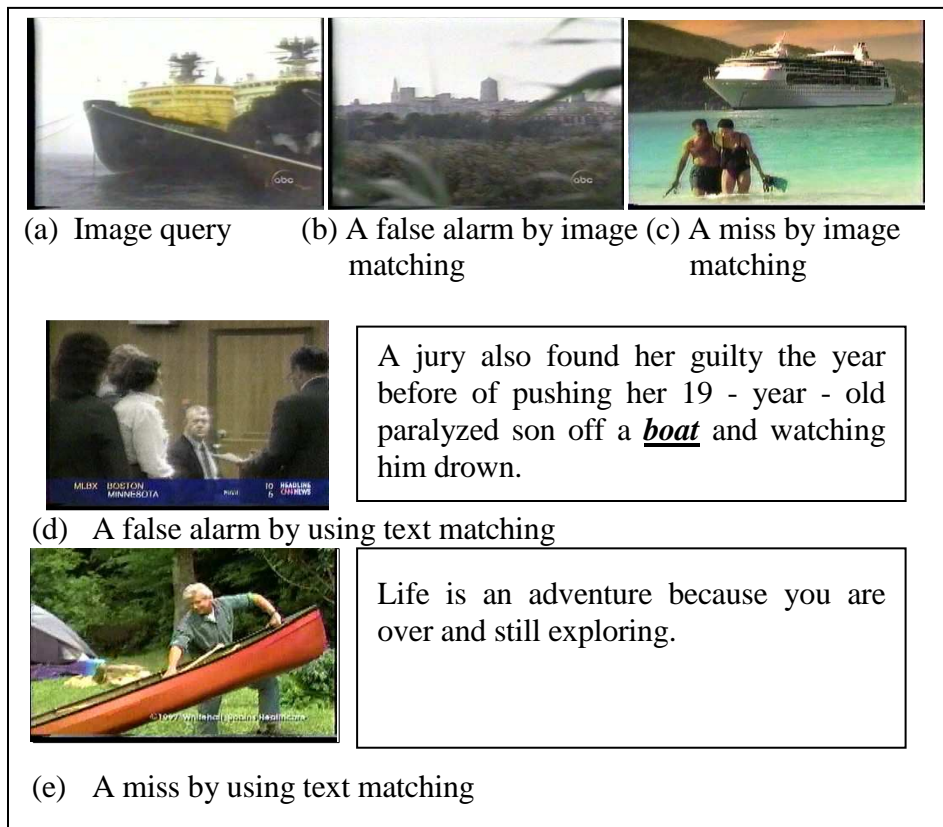


Figure 2.2: False alarms and misses when performing matching using low-level features to detect the concept “boat/ship”

2.2 Visual-based semantic in the concept detection task

Visual features are one of the most important classes of features in video analysis. In general, there are two types of visual features. One class is the set of low-level visual features such as color, textual and so on, and the other is the mid-level abstractions such as anchor person detectors, face detector and so on.

2.2.1 Low level visual features

In TRECVID, there are mainly three types of low-level image features that have been applied. They are the color [Stricker and Orengo, 1995], texture [Ohanian and Dubes 1992, Ma and Manjunath 1995] and shape [Amir et al. 2005].

Color has been shown to be the most widely used low-level visual features in TRECVID. This is because color provides strong clues that capture human perception. Many color models have been proposed such as the RGB, YUV, HSV, $L^*u^*v^*$ and L^*a^*b . One of the most effective and widely used representations of color-based feature is color moments [Stricker and Orengo, 1995]. As most of the information is concentrated in the first few moments, most researchers [TRECVID, 2002-2007] utilized only the first 3 moments, namely the mean, variance and skewness of the color distributions for each channel. Another type of color-based feature is color correlogram [Huang et al., 1997], which encodes the spatial correlation of pairs of colors.

Texture-based features are characterized by the spatial distribution of gray levels in a neighborhood [Jain, Kasturi, and Schunck, 1995]. In general, there are four types of methods to model texture [Tuceryan and Jain 1998]. They are

- Statistical methods, such as co-occurrence methods [Jain, Kasturi, and Schunck, 1995];

- Geometric methods, such as Voronoi tessellation features [Tuceryan and Jain 1990];
- Structural methods, such as texture primitives [Blostein and Ahuja 1989];
- Signal processing methods, such as Gabor filters and Wavelet models [Jain and Farrokhnia 1991].

In TRECVID evaluations, two widely used texture features are co-occurrence [Ohanian and Dubes 1992] and wavelet texture [Ma and Manjunath 1995].

Another type of low-level feature is shape. Various schemes have been proposed in the literature for shape-based retrieval, such as polygonal approximation of the shape [Schettini, 1994], image representation based on strings [Cortelazzo et al. 1994] and so on. However, such shape-based representation schemes are generally not invariant to large variations of image size, position and orientation. Thus, in TRECVID evaluations, the commonly used shape-based feature is edge-histogram layout [Amir et al. 2005].

2.2.2 Mid-level abstraction (detectors)

In addition to the low-level visual features, many mid-level detectors have also been built and widely used in news video processing. The most widely used mid-level detectors include face, anchorperson, and commercial, and shot genre detectors.

Because human activities are one of the most important aspects of news video, and many such activities can be deduced from the presence of faces, face detection is one of the most useful image processing technologies in the concept detection task [Hauptmann, 2005]. Yang, Kriegman, and Ahuja [2002] surveyed different methods to face detection. The methods usually made use of knowledge or machine learning based methods to detect face by facial features such as eyebrows, eyes, nose, mouth, skin color, and so on. Pham and Worring [2000] evaluated several reported methods currently available. They found that the method proposed by Rowley, Baluja, and Kanade [1998] performed the best. Such a neural network-based system is able to detect about 90% of all upright and frontal faces. The face detectors are used to detect people-related concepts. However, it should be noted that not many people-related concepts include only upright and frontal faces; hence, the use of frontal face detector has mixed performance in detecting people related concepts. One of the effective applications for face detectors is the

anchorperson detection. Anchorperson shots are graphically similar and occur frequently in a news broadcast. After obtaining results from face detectors, the systems in [Nakajima et al., 2002; Hauptmann et al. 2003] further detect anchorperson shots by combining text, audio, shot duration and visual features.

In news video, commercials are often inter-mixed with news stories. For efficient analysis of news video, the detection of commercials is essential. In general, there are three types of methods to detect commercials.

- Heuristic cutting marker methods

These methods of [Koh and Chua 2000] [Hauptmann and Witbrock, 1998] employed some special cutting markers such as black frames to detect commercials.

- Duplicate sequence methods

These methods [Chen and Chua, 2001] [Duygulu, Chen and Hauptmann, 2004] first detected candidate repeating keyframes and then construct the longest sub-sequence in detecting repeated commercials.

- Machine learning methods

The methods of [Duygulu, Chen and Hauptmann, 2004] employed a classifier such as a SVM to fuse the audio and visual features.

All methods have their strengths and weaknesses. No approaches can achieved the best in all situations. Thus researchers had been applying different methods in different applications.

Another type of visual-based detector is the shot genre detector. Such shot genre detectors [Chaisorn 2004; Snoek et al., 2004] divided news video into small sub-domain concepts such as live reporting, sports, finance, anchorperson, and so on. Researchers adopted knowledge engineering, machine learning, or their mixture to build such detectors.

The above mid-level detectors are widely used in news video processing. This is because:

- Such sub-domain data frequently occur in news video. For example, according to the statistics from Chaisorn [2004] in the TRECVID 2003 corpus, commercial shots and anchorperson shots account for about 40% and 9.5% of all the shots in news video, respectively.
- The performance of such sub-domain detectors is good. For example, Chaisorn [2004] claimed that the commercial detector achieves 99% in precision and over 95% in recall; the anchorperson detector achieves a performance of over 84.84% in precision and 87.6% in recall, and the overall accuracy of shot genres detectors is over 90%.

When obtaining such mid-level detectors, some researchers [Amir et al., 2005, Hauptmann, et al., 2005, Chua et al. 2004] made use of them to refine the results from general concept detectors.

In summary, in spite of many efforts that have been made to capture semantic concepts by visual features alone, except a few specific mid-level feature detectors in certain domains, the overall performance of the concept detection task is still unsatisfactory [TRECVID 2002-2007].

2.3 Text semantics in the concept detection task

Text information is another important information source in multimedia applications.

Rowe [1994] proposed to infer visual objects by using text semantics. In the paper, the author found that the primary subject noun phrase usually denotes the most significant information in the media datum or its “focus”. In the example of the image caption “Sidearm missile firing from AH-1J helicopter”, the “Sidearm missile” is the subject noun and “Ah-1J helicopter” is the prepositional phrase. Usually, we can expect to see a Sidearm missile firing in the image and we do not guarantee to find the helicopter in the image, because helicopter is in a preposition phrase and is secondary in focus. This image caption retrieval system was developed with the MARIE project for navy

aircraft equipment photographs. It was reported that the system could achieve 30 percent better precision and 50 percent better recall over a standard key phrase approach.

Sable et al. (2002) claimed NLP knowledge is useful in categorization based on text captions. Figure 2.3 shows an example. If we use the standard bag of words approach, we would associate the image with at least two categories:

- Rescuers → workers responding
- Victim → affected people

However, the predicate structure of the sentence emphasizes the rescuers and the ground truth made by human indicates that this image belongs to workers responding category.



Figure 2.3 Captions: Philippine rescuers carry a fire victim in March 19 who perished in a blaze at a Manila disco. <Taken from [Sable et al. 2002]>

However, we could not directly utilize such syntactic semantic technologies from the image caption retrieval to the concept-X detection task in news video. One of the important reasons is that the syntactic semantic analysis usually needs semantic parsers but they are designed for the grammatically correctly

written text. The other reason is that speech recognition text often contains too many errors that render the semantic parser ineffective. However, we can borrow the idea on using text focus to infer visual objects.

In news video processing, Satoh et al. [1997] suggested that co-occurrence relationship between name entities and concept person X is important in the “Name-It” project. Figure 2.4 shows two examples of the association between faces and names in videos.

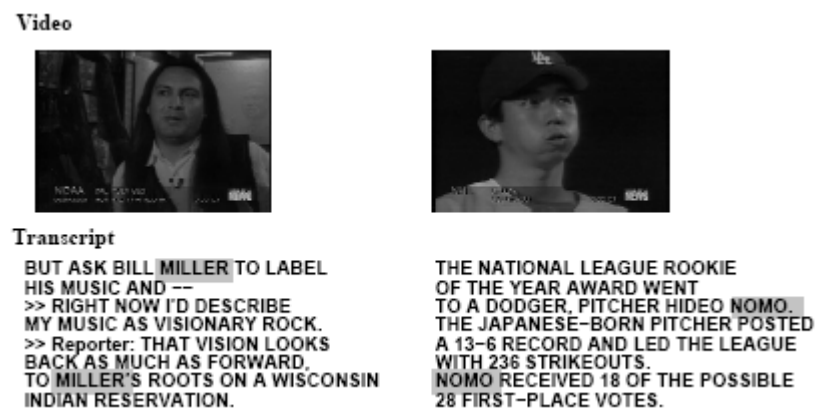


Figure 2.4: The association between faces and names in videos <Taken from Satoh et al. [1997]>

However, in many cases, we could not often find such a correlation relationship in a shot because of the mismatches between shot boundaries and text clues. Figure 2.5 shows the frequency of visual appearances of Bill Gates in relation to name occurrences, and Yang et al. [2004] used the Gaussian curves to capture the frequency distribution. However, Figure 2.6 shows that

different persons have different distance distributions, no matter whether we use the time-based or shot-based distance. Collecting such kinds of spatial distributions is a time-consuming task. It is also difficult to use such techniques in real applications. In any case such research suggests that text clues often have mismatches with the visual content.

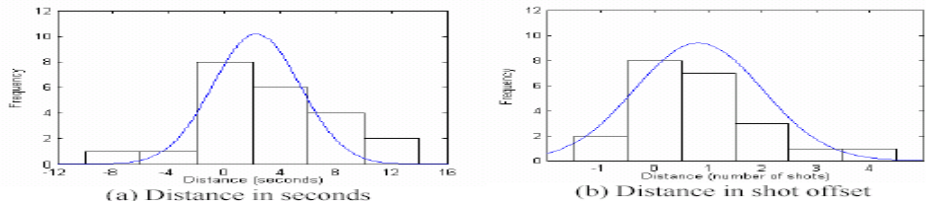


Figure 2.5: The frequency of Bill Gates visual appearances in relation to name occurrences. < Taken from Yang et al. [2004]>. We can find that there are time offset between visual appearance and name occurrences.

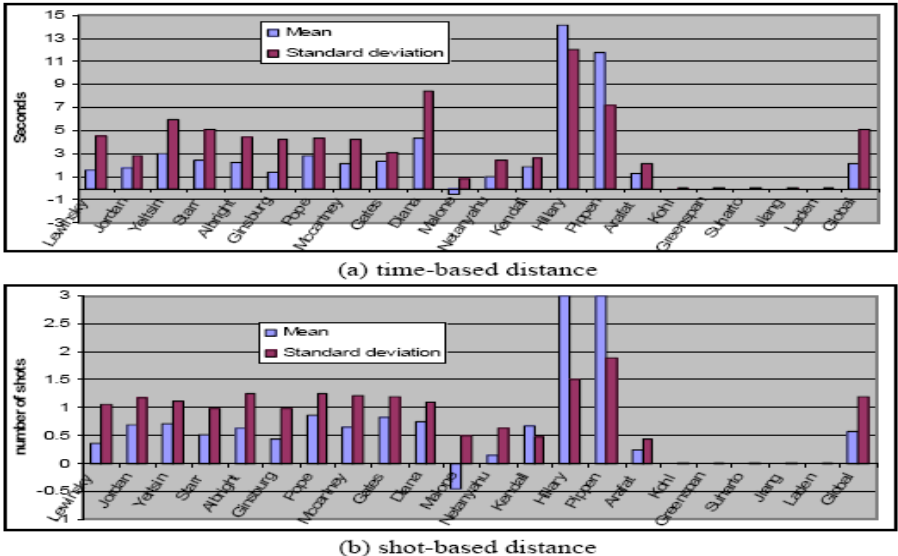


Figure 2.6 Different person X with different time distributions <Taken from Yang et al. [2004]>

There are two widely used methods to capture text semantics for general concepts in news video. One is text classification and the other is text retrieval. Text classification [Hauptmann et al., 2003] works for concepts that are transcribed with a specific and limited vocabulary such as the concept “Weather” in the 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 mismatch between text and visual contents at the shot layer and the difficulty in obtaining all typical training data. Text retrieval methods [Chua et al., 2004; Yuan et al., 2004; Campbell et al., 2006] regard words from concept text descriptions or some predefined keywords as queries and employ text retrieval with query expansion to find the related ASR transcriptions. After that, we can pinpoint the visual appearance based on the time information on the ASR results. Such methods are the only effective means when the training data is sparse and the text content in the test data includes the query word. However, in many cases, text clues in the test data do not contain the query words, and sometime not even appear in the expanded query word list.

Based on the above discussion, we found that both types of text analysis methods have their own strengths and weaknesses. Text classification captures the knowledge of training data. However, when the quality of training data is poor, the performance of the system is degraded. On the other hand, text

retrieval only captures the knowledge from concept text descriptions. Thus, when we could not find the text clues related to the query or when there are mismatches between text and visual appearance, the performance of the system will be poor. Hence, given a concept with some training data and the associated concept text descriptions, it is hard to know in advance which method is better.

In general, the analysis in news video based only on text is effective only if the textual descriptions of the desired visual concepts are well correlated.

2.4 Fusion of multimodal features

In general, there are three strategies to fuse the text and visual features in multimedia applications. They are:

- Strategy 1: First apply visual analysis to infer the concepts, and then employ the text semantic analysis.
- Strategy 2: First apply text analysis to infer the semantics, and then use the visual semantic analysis.
- Strategy 3: Jointly apply text and visual semantic analysis models to detect the concepts.

A lot of researchers adopted strategy 1 to fuse multi-modal features. For example, some image annotation algorithms, such as the translation model

[Duygulu et al., 2002], cross-media relevance model [Jeon, Lavrenko, and Manmatha, 2003], pattern-word association model [Xie et al., 2004] and so on, adopted the unsupervised visual analysis approaches to model images as the basis for annotation. For such a strategy, the performance of the systems is 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 their diverse appearances. Another example is the “Name it” project. The system built by Satoh et al. [1997] first detect faces, and then link name entities with those faces. However, in cases of errors (both misses and false alarm errors) in face detection, the later text analysis could not recover the errors from face detections.

Some researchers employed strategy 2 to detect semantic concepts. In the “Person X” detection project, Yang et al. [2004] first analyzed the text clues and they refined the results using visual constraints such as the filtering of anchorpersons. However, such a method will miss some relevant shots without appropriate text clues.

In TRECVID evaluations, many researchers adopted strategy 3 to combine multi-modal features. Snoek et al. [2006] summarized two general fusion approaches from different types of individual concept detection systems in TRECVID: 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 at the shot layer. In video analysis, a shot is one of the most widely used analysis units. It 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 to visual analysis but fails to capture the text semantics well. This is because the breaks from the shot boundaries occur often in the middle of a sentence. Figure 2.7 illustrates the problem of analyzing text using shot units, where the sentence is separated into three shot boundaries, which causes the mismatch between the text clue and the concept “Clinton”. To tackle this problem, Wilson and Divakaran [2008] proposed to detect scene changes by using training data under a supervised learning framework.

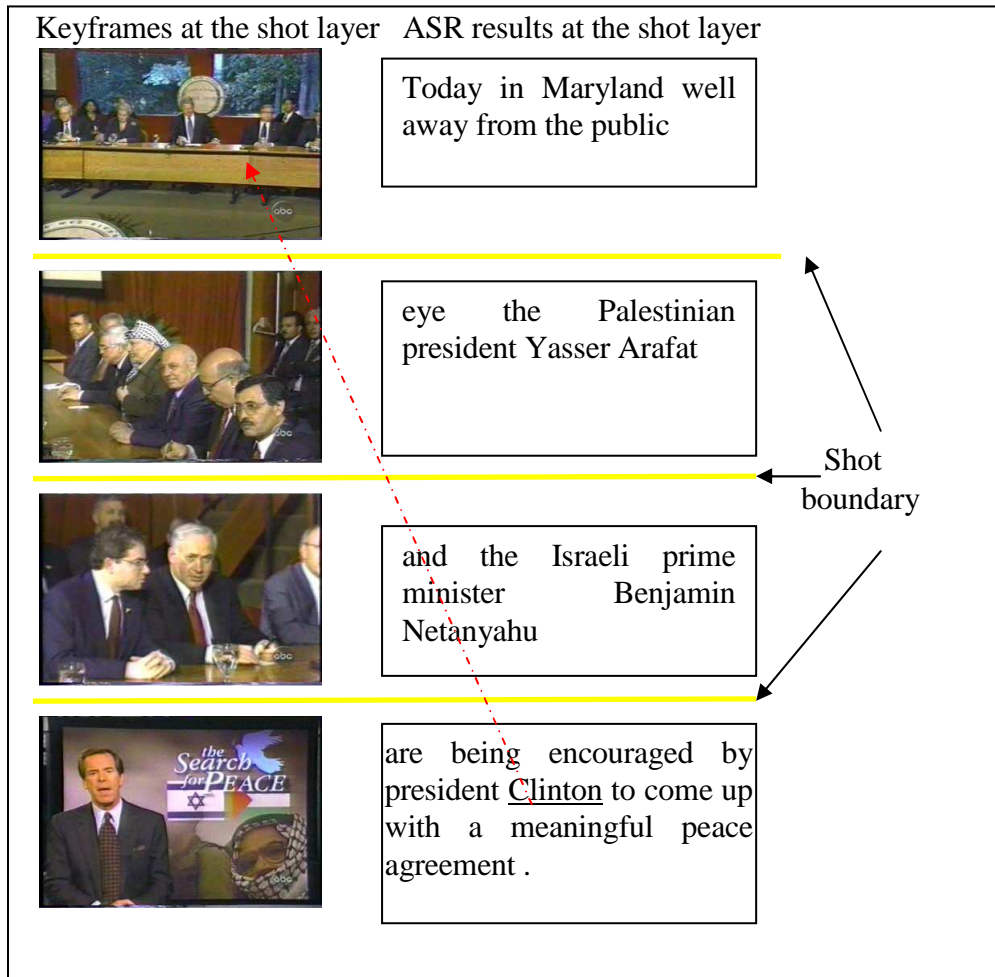


Figure 2.7: The sentence separated by three shot boundaries causes the mismatch between the text clue and the concept “Clinton”.

The late fusion scheme first reduces the unimodal features to separately learn the concept scores, and then integrate these scores to induce the concepts. The advantage of this approach is that it focuses on the individual strength of each modality. However, there is information loss in transferring from the original feature representations to scores. This brings about the potential loss of

correlation in the mixed feature space in the combination of scores from multi-modal feature analysis.

In summary, no matter which strategy is used, it is hard to let the evidence from different modalities support each other.

2.5 Machine learning in the concept detection task

Generally speaking, there are two approaches to describe the visual appearance of the concept in videos. One way is to adopt the rule-based approach. Several researchers [Yang, et al. 2004] captured semantic concepts by using a text retrieval approach with manual combination of some visual constraints. However, the drawbacks of such rule-based approaches are the lack of scalability and robustness. To overcome the problems, machine learning-based methods are widely used to detect semantic concepts in TRECVID evaluations. Given a set of training data, usually there are three types of machine learning inference methods. They are supervised inductive learning, semi-supervised learning and transductive learning. Most fusion approaches belong to supervised inductive learning. A number of researchers have adopted semi-supervised learning to fuse the multi-modal features. Few transductive learning algorithms have been used in the analysis of visual features. In this section, we will review the above three learning methods.

2.5.1 Supervised inductive learning methods

Supervised learning is an inference method, which first estimates the unknown dependency from training data and then uses the learned dependency model to predict outputs for future inputs. Most generic models [Snoek et al. 2006, Souvannavong et al. 2004, Naphade et al. 2002] in the semantic concept detection task employed supervised inductive learning methods such as Neural Networks [Amir et al. 2003], Hidden Markov Models [Huang, Wei and Petrushin, 2003], Support Vector Machines [Snoek et al., 2006], AdaBoost [Wu et al., 2003], Decision Trees [Hauptmann et al., 2002] and so on. However, according to the “No Free Lunch Theorem” [Duda, Hart and Stork, 2004], we could not prove in theory that any algorithm is better than the other learning algorithms. The reason is that the assumptions about the learning domains are relevant to the choice of the learning algorithm. In practice, we could observe the fact that more and more researchers have selected SVMs as their classifiers. This is because there are not enough training data in the TRECVID corpus and SVMs work well, especially when the training data is sparse. In addition, some researchers attempt to combine several learning methods together. For example, the best performance [Yuen et al. 2007] in TRECVID 2007 was achieved by combining of several learning methods such

as Stack SVM, RankBoost, and so on. However, how to select the learning methods and how to combine them are still open problems.

2.5.2 Semi-supervised learning

Semi-supervised learning [Zhu 2006] was proposed to use large amount of unlabeled test data together with the labeled training data to build classifiers. The co-training algorithm [Blum and Mitchell 1998, Pierce and Cardie 2001] is a typical semi-supervised learning method. The algorithms apply to learning problems that have multiple views, i.e., several disjoint subset of features, each of which is sufficient to learn the concepts of interest. Generally, semi-supervised learning algorithm includes two steps. First, they use a small-labeled training set to learn a classifier in each view. Then they bootstrap the views from each other by augmenting the training set with unlabeled samples acquired from the other views with high-confidence predictions. A number of researchers such as [Yan and Naphade 2005] adopted a semi-supervised learning method in semantic concept detection on news video corpus. Yan and Naphade [2005] proposed a semi-supervised cross feature learning method, which is a type of co-training algorithm. They removed one assumption of co-training algorithm that each view should be sufficient for learning by adding a validation set to monitor the performance of each view. The advantage of such

a semi-supervised learning is to use unlabeled test data to reduce the efforts of preparing the training data. In the report of Kender et al. [2004], they demonstrated that their semi-supervised learning method could achieve a good performance. However, it is still worse than the best supervised learning system developed in their own group. One of the reasons is that it is hard to make sure that the data in the validation set has the same distribution as those in the test set, especially when the corpus is large. Tian et al. [2004] pointed out that the unlabeled data helps only if the labeled and unlabeled data are from the same distribution. Otherwise, the unlabeled data may degrade the performance when it is added.

2.5.3 Transductive learning

Instead of obtaining a general hypothesis capable of classifying any “unseen” data under a supervised inductive learning framework, transductive learning [Marchenko, Chua and Jain, 2006] [Qi et al. 2007] [Yaniv and Gerzon, 2004] is concerned with directly classifying the given unlabeled data. Qi et al. [2007] proposed to further purify those hierarchical clustering results by a Gaussian Mixture Model (GMM) with an expectation maximization algorithm. A pure cluster is defined as the 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. Qi et al. assumed that they could find a good number of the Gaussian mixtures and the data distribution follows the Gaussian distribution. However, when the corpus is large, the above conditions are not always being satisfied. This is because there are many different types of data distributions, Gaussian distribution is only one of them and it is hard to estimate a good number of Gaussian distributions to model the data. In addition, only focusing on the purity of the cluster is not enough. This is because the performance of the system relies not only on precision, but also on recall. As we know, the purest state is that each shot is a cluster. However, it is not useful to make any inference, because it has the lowest recall. Therefore, the size of the cluster is another important factor for the performance of the classifiers. Marchenko et al. [2006] compared single-link, complete-link, average-link and k-means clustering approaches in the transductive learning framework. Their test results demonstrate that the results from the average-link clustering achieve the best results in her painting domain data set.

The above two methods only employ the visual features only. How to explore the correlation between text and visual features under a transductive framework is still an open problem.

In summary, 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 [Jain, Murty and Flynn, 1999]. 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. The other problem is how to analyze the unknown clusters, which are impure clusters or clusters that include only test samples.

2.5.4 Comparison of the three types of machine learning methods

Although much progress has been made in machine learning, how to capture the characteristics of semantic concepts has not been entirely successful. Table 2-1 contrasts the main features of the major machine learning approaches.

Table 2-1: Comparison of three types of learning approaches

	Assumptions	The size of training data	Inference ability
Supervised learning	<ul style="list-style-type: none"> ● Distribution of training set \approx test set ● Other specific assumptions⁷. 	As large as possible.	Make inferences for any input data via building a model based on training data.
Semi-supervised learning	<ul style="list-style-type: none"> ● Distribution of training set \approx test set ● The newly added unlabeled data assumptions⁸. 	Need some seed data.	Make inferences for any input data via building a model based on training data.
Transductive learning	No above assumptions, but it needs to map test data to training data.	Need typical data for the given corpus.	Any available test data set.

Both supervised and semi-supervised learning has different types of assumptions. The most important assumption is that the data distribution from

⁷ For example, the KNN algorithm works well in the cases that we could obtain a good “K” in advance. Bayesian inference builds the whole framework on the probability theory, which assumes to follow the “Law of large numbers” and most of possible cases in the test data should be covered in the training set.

⁸ At each iteration, semi-supervised learning will add some unlabeled data with high confidence of predicting labels into training data set. It assumes that the labeled and the new added unlabeled data are from the same distribution.

training set and test set is similar. On the other hand, transductive learning does not require the above assumption. In addition, supervised learning approaches need as much training data as possible in order to capture the characteristics of unknown test data. Semi-supervised learning attempts to tackle the large training data problem by incorporating unlabeled data with high confidence semantic labels into the training set. In order to achieve a good performance, it assumes such new unlabeled data comes from the same data distribution as that of the labeled training data. However, the assumption is not always satisfied. Transductive learning needs some typical training data for the given corpus, because it attempts to label test data by mapping it to training data, instead of a general model for any input data. However, one requirement of transductive learning is that it needs the test data set to be available in advance. This is because it does not have a model that can process any possible input.

2.5.5 Domain adaptation

Domain adaptation of statistical classifiers is the technique that arises when the data distribution in the test domain is different from that in the training domain. Most recent methods attempted to capture the data distributions in a new domain by using a large numbers of data from other domains and a

relative small amount of data in the new domain. Researchers will need to label some data in the new test domain as supplementary training data set to adapt the older classifiers. For example, given an existing classifier, Yang et al [2007] required a sufficient amount of labeled examples in the new dataset (test set) to learn the “delta function” between the original and the adapted classifier. Jiang and Zhai [2007] proposed to implement several adaptation heuristics to train a model from source domain (training) to predict the target domain (test). There are three heuristics in their proposal: (1) removing misleading training instances in the source domain; (2) assigning higher weights to labeled target instances than labeled source instances; and (3) augmenting training instances with target instances with predicted labels.

The assumption of such heuristics is that they have some labeled data in the test set. This may not be practical in many cases because there may not be much labeled data. Hence, if we just have very few labeled data from the test domain, how can we ensure that such data are typical?

2.6 Multi-resolution analysis

Davis and Bigelow [1998] provided a definition of multi-resolution models:

- An integrated family of two or more mutually consistent models of the same phenomena at different levels of resolution.

Generally, a researcher first takes data at the different resolutions to create a multi-resolution structure and then derives error metrics to help decide the best level of detail to use. The multi-resolution model is widely used in image processing, such as image pyramids [Wang and Li, 2002]. Such an approach first analyzes 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 [2000] and Li [2001] used a multi-resolution model to detect shot and story boundaries for video and text documents respectively. They used information at the lower resolution to locate the transition points and the higher resolution to identify the exact boundaries by finding the maximal path. Similarly, Slaney et al. [2001] proposed a multi-resolution analysis method to detect discontinuities in videos for story segmentation.

As far as we know, few multi-resolution models have been applied in the semantic concept detection task to fuse multi-modal features in the TRECVID corpus [TRECVID, 2002-2007]. Most current approaches, especially those

used in the large-scale TRECVID video concept detection and retrieval evaluations, such as [Chua et al., 2004], 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, and not used to reinforce the subsequent shot level analysis. 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 the 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 evidence from multi-resolution layers.

2.7 Summary

From the above discussion, we found that multimedia requires the integration of multi-modal features, and systems that emphasize only the use of single modality obtain poor performance. Recently, many systems focused on fusing the multi-modal features to detect concepts. Compared to the single modality based analysis, multi-modal fusion systems have reported better performance.

However, few works have been done to allow evidence from different modalities to support each other. Thus, Rowe and Jain [2005] listed the multimedia fusion as one of the SIGMM grand challenges:

“A third facet of integration and adaptation is the emphasis on using multiple media and context to improve application performance.”

Fusion of multi-modal features in news video to capture semantics has been carried out in many years. Most researchers employ machine learning approaches to perform fusion. Although much progress has been made in machine learning, the application of these technologies to news video concept detection has achieved mixed results. This is because machine learning systems are highly dependent on the quality of training data. In the concept detection task, the selection of training data is based on random sampling. We use a set of news video at the certain time period as training data and regard another set of news video at the other time period as test data. Due to the characteristics of news video, we could not always ensure the distribution of training data is similar as that in the test data. On the other hand, some researchers adopt text retrieval techniques to capture concepts in the news video. However, it is effective only when the textual descriptions of desired concepts are synchronized with the visual contents. Thus, we can find that both machine learning and text retrieval methods have their own strengths and weaknesses and their relative performance across multiple test corpora are

mixed. However, few works have integrated those two methods together and tried to let multi-modal features support each other.

In this thesis, we propose a method to integrate machine learning method and text retrieval together. It first adopts transductive inferences 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 concepts for the remaining ambiguous test samples by performing a multi-resolution analysis that incorporates web-based knowledge in a retrieval framework. The multi-resolution model makes use of different modal features at different resolution levels and introduces constraints from other levels when performing analysis at each resolution level. This model can let evidence from different resolutions support each other.

Chapter 3

SYSTEM ARCHITECTURE

In this chapter, we first briefly introduce our design consideration. We then report the system architecture.

3.1 Design consideration

In this section, we introduce the motivation of our M3 framework. In particular, we focus our discussion on three topics: the multi-resolution analysis, the multiple sub-domain analysis and the combination of machine learning and text retrieval.

3.1.1 Multi-resolution analysis

In order to fuse text and visual features effectively, we propose our multi-resolution transductive model. In our framework, we define three resolution

layers. They are the shot, multimedia (MM) discourse and story layer. A shot is an unbroken sequence of frames from one camera shot. The so-called multimedia (MM) discourse aims to capture the synchronization between visual features at the shot level and text features at the sentence level. A story provides the detailed information of an event. At each resolution, we make inferences by using a transductive inference to capture the knowledge from training data.



Figure 3.1: The ability & limitation of visual feature analysis at the shot layer

At the shot layer, our transductive inference is based on color, texture and edge visual features. Generally speaking, it puts similar images together and transfers the semantic labels from training to test data among similar images.

We adopt an average-link clustering method to cluster images from the whole corpus. The choice of the clustering results is selected based on Vapnik Combined Bound [Vapnik, 1998] in the transductive inference. Figure 3.1 provides some examples to demonstrate the ability and limitation of visual feature analysis at the shot layers to cluster shots sharing same concepts.

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 using text information. In order to tackle the mismatches between the text clues and visual contents at the shot layer, we define a new unit, namely the multimedia discourse unit. MM discourse boundaries may only occur at the co-occurrence of sentence and shot boundaries. In this work, we adopt the speaker change boundaries generated by a speech recognizer [Gauvain, Lamel, and Adda, 2002] as the pseudo-sentence boundaries. At the MM discourse layer, we capture the semantics mainly by extracting a group of words from the enclosing ASR text. The transductive inference attempts to use such words to capture semantics and transfer the semantic labels by finding similar text contents. However, as the ASR text at the MM discourse layer is insufficient to infer the linguistic variations and domain knowledge, we narrow this gap by exploring the relationship between concept text description and text contents in the test data via web co-occurrence. Figure 3.2 gives some examples to demonstrate the

ability and limitation of text feature analysis at the MM discourse layer.

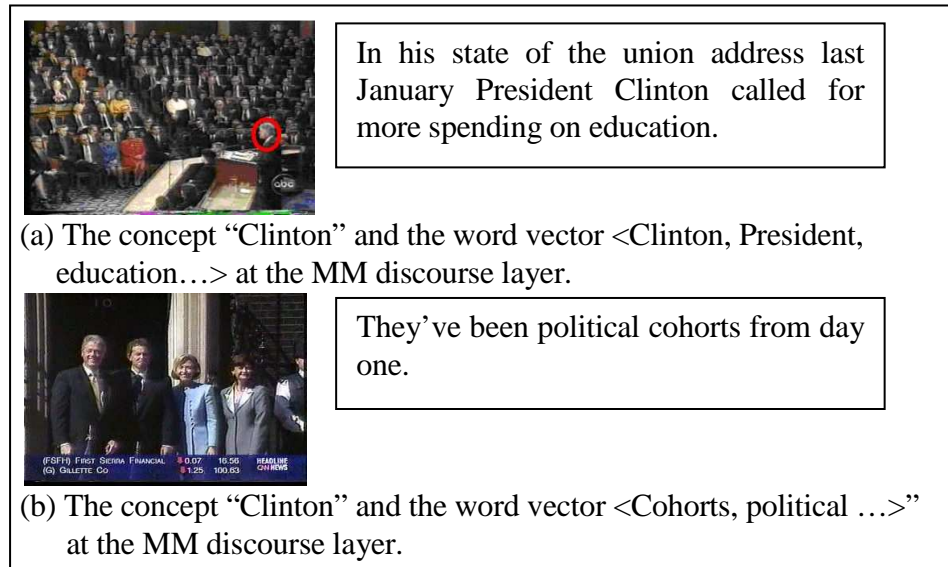


Figure 3.2: The ability and limitation of text analysis at the MM discourse layer

Although we can infer that the shot in Figure 3.2 (a) has high degree of relevance to the concept “Clinton” based on the word vector at the MM discourse layer, it is hard to make a decision in Figure 3.2 (b). In order to tackle the problem in Figure 3.2 (b), we incorporate text analysis at the story layer into the framework. The text analysis at the story layer is similar to that at the MM discourse layer. The main difference is that we attempt to capture the semantic concepts by exploring the relationship between the concept and the topics of a story. Here the topic refers to the main focus of a story. We employ the method developed in [Lin, 1997] to extract topics, which mainly

depends on a set of high frequency ASR words in a story. After extracting topics from a story, we form the topic vector for a story, which includes all topic terms at the story. For example in Figure 3.3, we can obtain the topic vector as {president, Clinton, Blair}. According to such a topic analysis, we can conclude that the enclosed shots (such as the shot in Figure 3.2 (b)) should have some degree of relevance to the concept “Clinton”.

Among these three resolutions, the shot layer analysis could achieve the highest precision, but the lowest recall. The performance of multimedia discourse layer analysis is in the middle at both precision and recall. The story layer analysis could obtain the best recall but the worst performance in precision. We adopt a bottom-up strategy to integrate three-layer analysis together to achieve higher performance.



Figure 3.3: An example text analysis at the story layer

As different modal features in news video represent the same events, we should let different modalities at different resolutions support each other. Thus,

we propose two constraints: the must-link and cannot-link constraints. The must-link constraints try to model the phenomenon that the precision goes down from the highest resolution layer (shot) to the lowest resolution layer (story). That is, the decision made at the higher resolution must be followed in the lower resolution analysis. Thus, 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 lower resolution analysis. The cannot-link constraints are designed to employ lower resolution text semantics to purify higher resolution results so that we have less chance to regard the low-level feature similarity (say visual similarity) as the high-level semantic similarity. That is, the semantic similarity depends not only on visual similarity, but also text similarity. If two given shots are only similar in visual feature space, these two shots cannot be clustered together. In our framework, when performing higher resolution analysis, we bring in cannot-link constraints from the lower resolution to leverage the higher resolution analysis.

The above discussion briefly introduces our bottom-up multi-resolution strategy. An alternative design is to consider a top-down multi-resolution strategy. However, because of the low discriminative power of the higher resolution analysis, it is hard to infer semantics from different resolutions by adopting a top-down multi-resolution strategy. For example, although we find

that a story (say the story in Figure 3.3) is related to the concept (for example, the concept “Clinton”), not all the shots in the story contain the concept. Moreover, it is hard to remove the noise shots from the shot lists in the story by performing the MM discourse and shot layer analysis. The situation at the MM discourse layer is similar to that of the story layer. Thus, we do not employ a top-down multi-resolution inference strategy.

3.1.2 Multiple sub-domain analysis

In concept detection, many researchers have developed mid-level detectors to supplement the low-level features such as the color and texture. Several researchers have reported good performance on a number of mid-level detectors in news video. For example, Chaisorn [2004] reported high accuracy of over 90% for shot genre detectors, such as anchorperson, live reporting, commercial, finance, etc. Techniques to detect sub-domain boundaries are relatively mature. Thus, the multiple sub-domains analysis has been used in the query-class dependent retrieval [Yan et al., 2004, Chua et al., 2004]. They first classified each user’s query into one of the predefined categories and then aggregated the retrieval results with query-class associated weights. This suggests that multiple sub-domain analysis should be effective. However, few

works have made use of the concept occurrence distributions from multiple sub-domains to enhance the concept detection task [TRECVID 2002-2007].

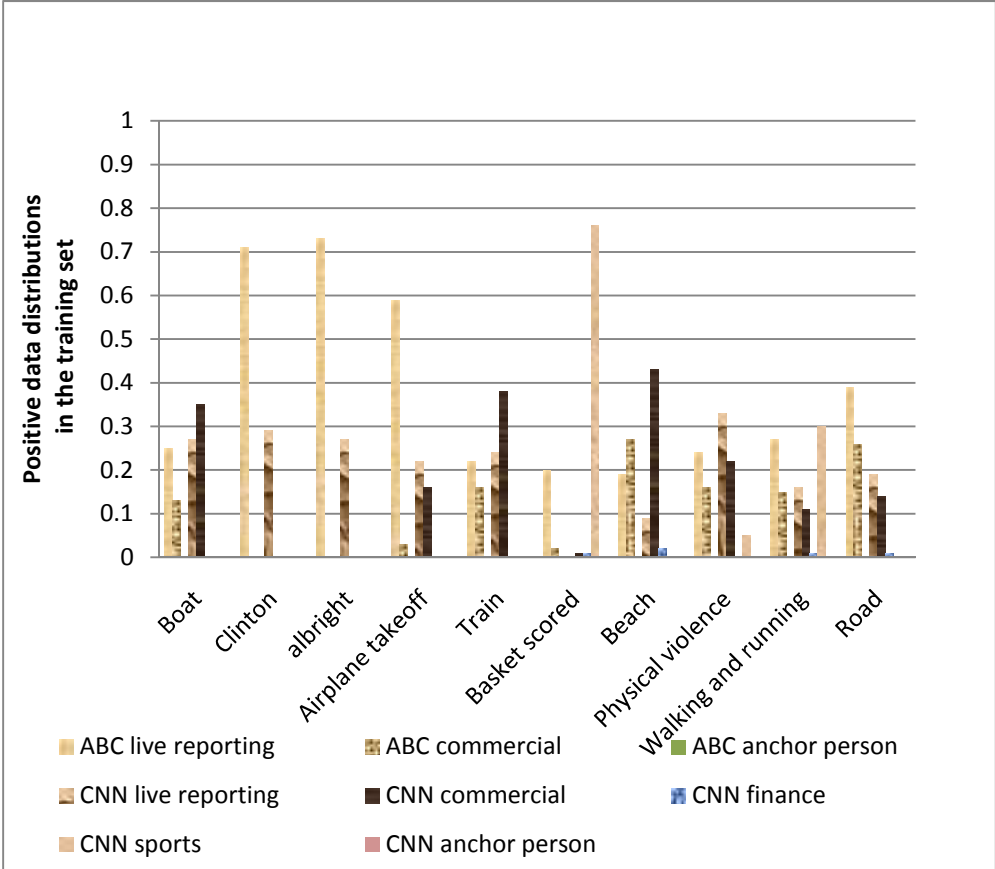


Figure 3.4: The distributions of positive data of 10 concepts from TRECVID 2004 in the training set.

In our work, we segment the news video corpus into the sub-domains of anchorperson, sports, finance, commercial, with the rest being placed under live reporting. The reasons for the above choice are that the detectors for the first four categories are well-defined [Chua et al., 2004], and the distributions

of concepts in those 5 categories are distinctive. We use 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 Figures 3.4, 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.

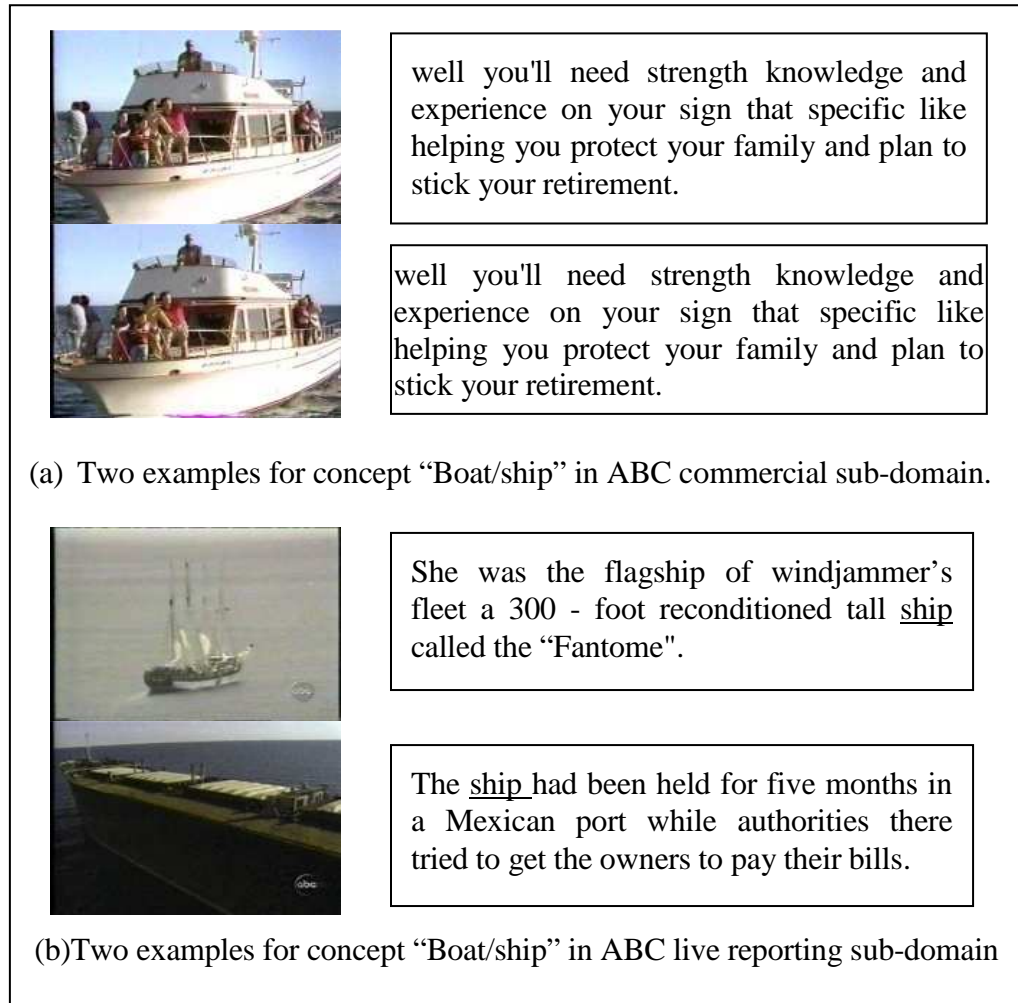


Figure 3.5: The characteristics of data from different domains may be different

In addition, the characteristics of data from different domains may be different.

From Figure 3.5, we find that shots sharing the same semantic concept for a product commercial usually have high similarity (or are even identical) 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 visual feature space. In order to capture the characteristics from different sub-domains, we need to analyze data from different sub-domains separately.

3.1.3 Machine learning and text retrieval

The common problem of 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 from the knowledge of 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, two news reports -- one about “September 11 event” and the other about “The progress of NATO invading Afghanistan” are given in the training and test data respectively. 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. Otherwise, we may have difficulties to assign the semantic label “violence” to the test data based on

training data. The other variation is called “mutation”, in which the concept can occur in unrelated events. 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. Thus, we should consider these two problems in our framework.

Although machine learning methods could learn some knowledge from training data, performance is highly dependent on the quality of the training data. On the other hand, text retrieval may be effective when the training data is sparse and text contents in the test data includes some query terms. For example, for test data 1 in Figure 3.6 (b), text retrieval could capture the concept “boat/ship”, because the query word “ship” appeared in the ASR transcript. At the same time, machine learning methods may fail, because of the large gap between training and test data. For test data 2 in Figure 3.6 (c) machine learning methods using text features can work well, because there exist clearly patterns between training and test data, 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, machine learning using text features and text retrieval approaches have their own strengths and we need to combine them to take advantage of their strengths in concept detection.




Images	The ASR results at the MM discourse layer	
	<p>Life is an adventure because you are over and still exploring.</p>	<p>Training data (a)</p>
	<p>The <u>ship</u> had been held for five months in a Mexican port while authorities there tried to get the owners to pay their bills.</p>	<p>Test data 1 (b)</p>
	<p>Life is an adventure because you are over and still exploring.</p>	<p>Test data 2 (c)</p>

Figure 3.6: An example of detecting concept “boat/ship” using two text analysis methods.

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

text retrieval model, which captures the relationship between the words describing events and concepts such as “boat / ship” via web statistics.

3.2 System architecture

In this section, we describe the architecture of our system. Figure 3.7 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 [Chua et al., 2004] to segment the corpus into sub-domain data sets. We then perform the multi-resolution, multi-source and multi-modal (M3) transductive learning model as shown in Figure 3.8 to detect the concepts in each of the sub-domain data set 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 repeat the M3 transductive inference in sub-domain data sets, where new test data are added into the training data.

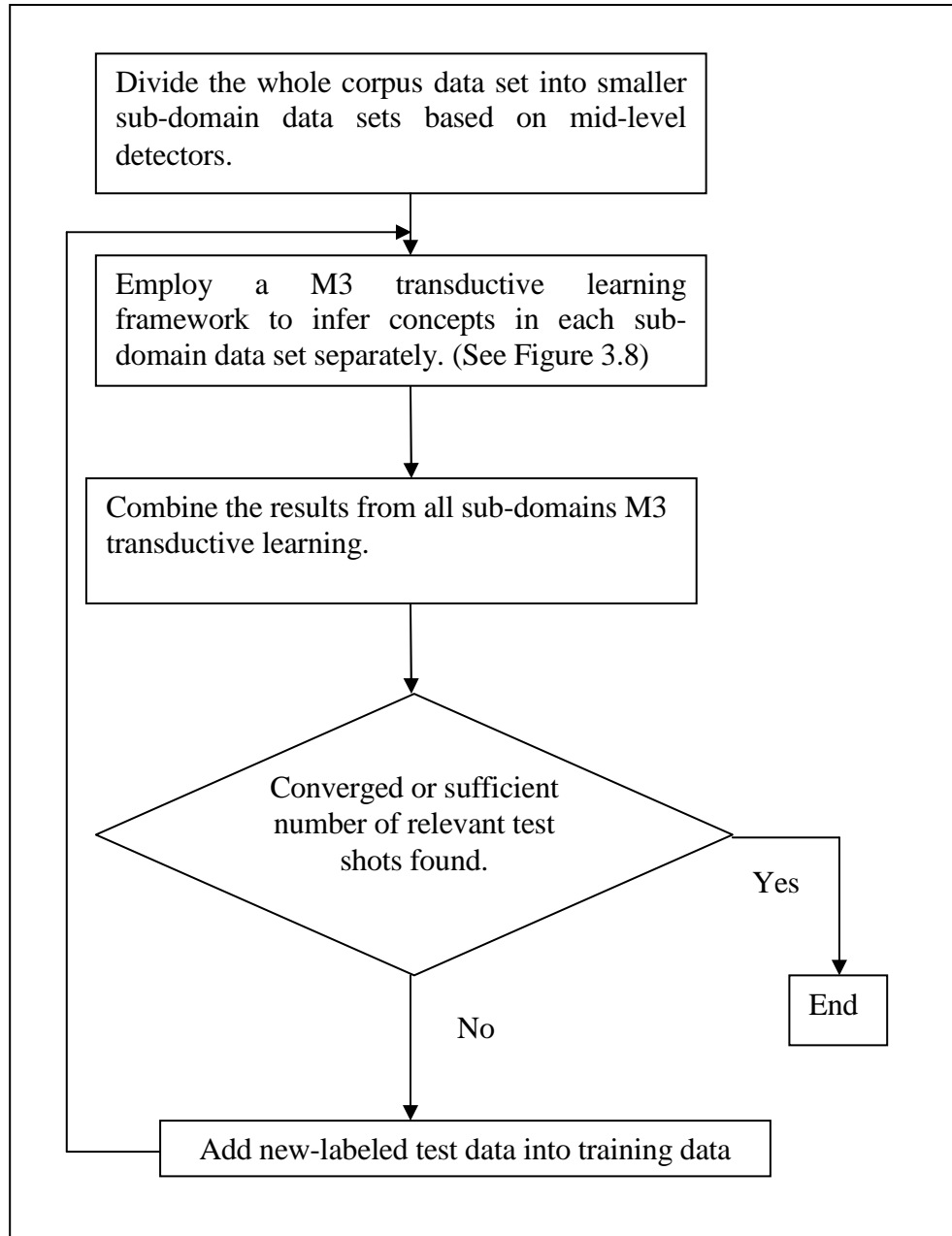


Figure 3.7: The bootstrapping architecture

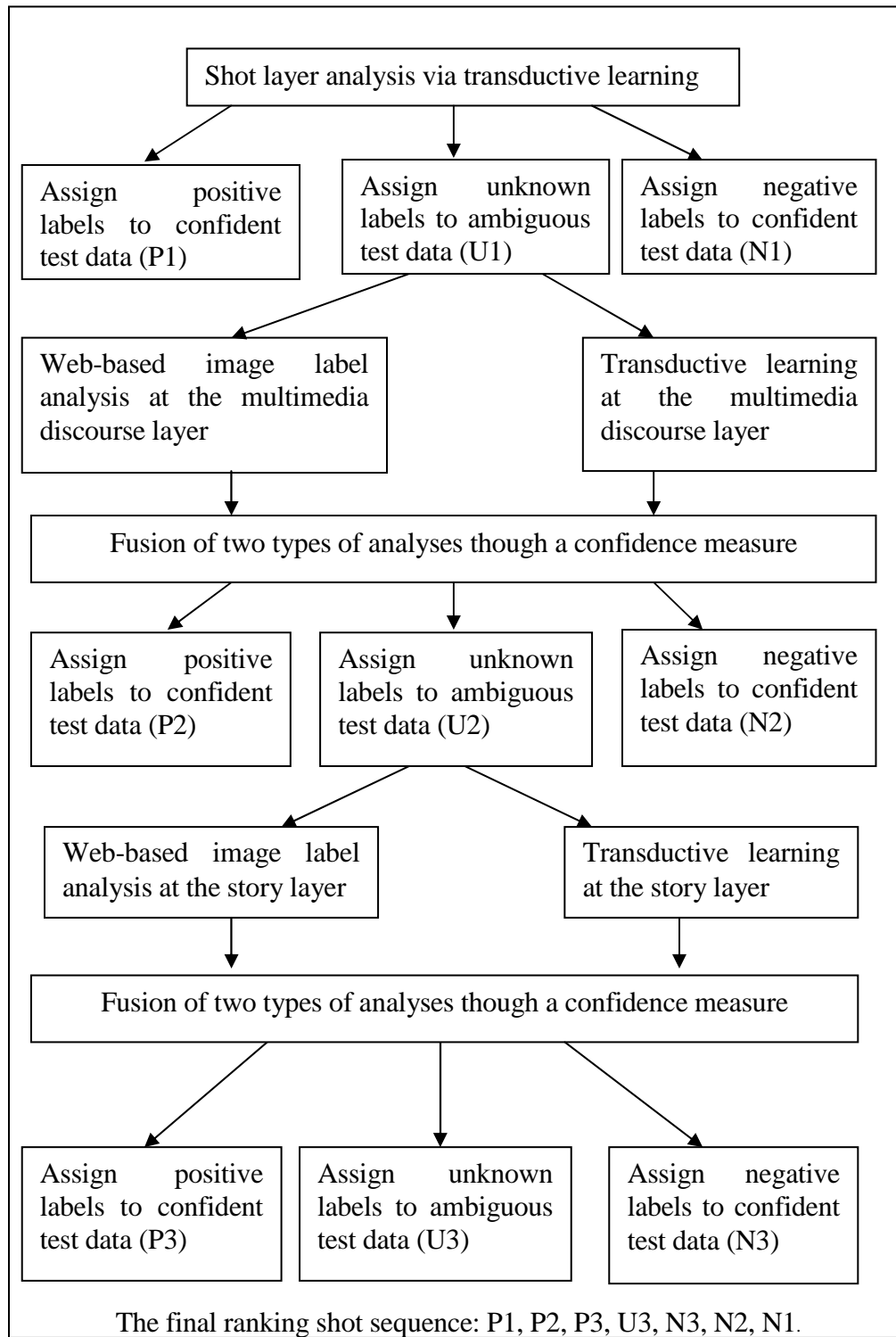


Figure 3.8: The multi-resolution transductive learning framework for each sub-domain data set

Figure 3.8 shows the details of the M3 transductive inference framework for each sub-domain. Our inference begins with the shot layer analysis. We infer the labels of test shots by clustering shots via a transductive learning framework. The confidence of our inference depends on the number of training data and its purity in any cluster. Based on the analysis, we divide the test data into three categories. The first two categories of test data can be labeled by the training data in the same cluster with high confidence. We use P1 and N1 to represent the set of positive and negative test shots, respectively. The other category of test data cannot be labeled as positive or negative with high confidence. We use U1 to represent these unknown shots. 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 automatically annotate such visual clusters by the word vector at the MM discourse layer. Two types of methods will be applied to make further inference. One is to capture the relationship between the word vector at the MM discourse layer and words from concept text descriptions via web statistics. The other is to further cluster shots by a transductive learning method based on the word vectors. After the analysis of MM discourse layer, we can divide the U1 set into two sets. One set is the labeled set: which includes a positive (P2) and a negative set (N2). The other

set is still the unknown set (U2).

For the U2 clusters, we label them using the topics extracted at the story layer.

We then perform a similar transductive inference as in the MM discourse layer to rank the U2 shots based on the story layer inference. After story layer analysis, the U2 data set is classified into P3, N3, and U3 sets. The final ranking of the shots is as follows: P1, P2, P3, U3, N3, N2, N1.

Chapter 4

MULTI-RESOLUTION ANALYSIS

In this chapter, we first introduce different types of features used in different resolution layers. We then report the multi-resolution constraints clustering.

4.1 Multi-resolution features

In this section, we first present visual features at the shot layer. Because there are few differences between our design with that of others at the shot layer, we only briefly present our approach here. One of the significant differences with other works is how we capture text semantics to help detect visual concepts. Thus, we pay more attention to discuss the text feature at the multimedia discourse and story layer.

4.1.1 Visual features

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 includes: Edge Histogram Layout (EHL: 8 dimension), Color Correlogram (CC: 64 dimension), Color Moments (CM: 225 dimension), Co-occurrence Texture (CT: 96 dimension) and Wavelet Texture Grid (WTG: 90 dimensions). For each shot, we extract the above visual features and generate a feature vector $f(f_1, f_2, f_3, \dots, f_t)$. As discussed in the previous chapters, the clustering processing is one of the core components in our transductive learning. One of the most important aspects in the clustering processing is the definition of the similarity measure. At the shot layer, we adopt the cosine similarity between feature vectors as the similarity between shot i and shot j :

$$\cos sim(i, j) = \frac{\sum_{k=1}^t (f_{ki} \cdot f_{kj})}{\sqrt{\sum_{k=1}^t f_{ki}^2 \cdot \sum_{k=1}^t f_{kj}^2}} \quad (4-1)$$

In addition to the above low-level visual features, we also employ four high performance mid-level detectors as follows:

- 1) The anchor person detector [Chua et al., 2004]

This is the most typical shot genre in news reports with one to two

anchorpersons appearing in the fixed background. Such shots normally contain at least one detected fronted face and always contain many repeating occurrences with similar images on the same day in news video. Examples of anchorperson shots in a news video from the TRECVID data are shown in Figure 4.1. It is reported [Chaisorn, 2004] that the anchorperson detector could achieve a performance of over 84.84% in precision and 87.6% in recall.



Figure 4.1: Examples of anchorperson shots from a news video

2) Commercial

Commercials are used to convey messages for selling products. The commercial shots typically contain fast changing shots and end with still images showing the company's logos or products. The commercial boundaries can normally be characterized by the presence of black frames, still frames and/ or audio silence [Koh and Chua 2000]. Sample keyframes of a product commercial are shown in Figure 4.2. Experiment results [Chaisorn, 2004] showed that the commercial detector could achieve 99% in precision and over 95% in recall.



Figure 4.2: Commercial shots for a product in news video

3) CNN finance

CNN finance is a special news program, which usually appears in the middle of CNN Headline News. Such finance shots normally begin with some special logo shots, include finance diagram shots and end with the beginning of the next commercial shots. Examples of the typical financial shots are shown in Figure 4.3. We use the tools in Chua et al. [2004] to label the CNN finance shots. Chaisorn [2004] claimed that the finance detector could achieve 100% in precision and 100% in recall in a small data set.



(a) A financial program logo



(b) A finance diagram shot

Figure 4.3: Examples of CNN financial shots

4) CNN sports

CNN sports is a special news program in CNN Headline News, which follows the CNN financial news. Such sports shots normally begin with some special logo shots and normally contain shots with noisy background and high motion activities. Examples of the typical sports shots are shown in Figure 4.4. We employ the tools in Chua et al. [2004] to label the CNN sports shots. Chaisorn [2004] reported that the performance of such a detector is high with accuracy of over 90%.



(a) A sports program logo



(b) An example of sports shot

Figure 4.4: Examples of sports shots

The above mid-level detectors are used to segment shots into different sub-domains. Except the above four types of shots, we assign the label “live-reporting” to the remaining shots. Because the styles in ABC World News Tonight and CNN Headline News are different, we divide them separately. Thus, we segment the TRECVID 2004 corpus into eight sub-domains, which are shown in Figure 3.4.

4.1.2 Text features

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 and closed captions (if they are available). Closed captions typically display a transcription of the audio portion of a program with punctuations. If closed captions are available, we use them to align ASR results to obtain the punctuation marks and reduce speech recognition errors from the ASR text. Otherwise, we regard the speaker change boundary as the pseudo punctuation mark to segment the ASR text. The text features used at the MM discourse layer are words. On the other hand, a story includes many sentences. Compared to the resource at the MM discourse layer, it provides a relatively rich set of linguistic information. We

are thus able to employ a topic extraction algorithm [Lin 1997] to capture text focus of the story. Compared to words, using topic terms to infer visual concept is more effective. This is because the topic usually refers to the high frequency visual concept in the story. We extract topic terms at the story layer. In general, there are three types of methods to capture topic semantics. They are statistics-based [Paice, 1990], knowledge-based [Hahn, 1990] and hybrid [Hearst, 1994]. Among these techniques, only the word frequency counting method, which belongs to the statistical methods, 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 [Shibata and Kurohashi, 2006] adopted knowledge-based approaches to identify topics in some specific domains, such as cooking instruction videos. However, as far as we know, few researchers adopted topic identification techniques to support concept detection in an open domain such as news video [TRECVID, 2002-2007].

Both the word vector at the MM discourse layer and topic term vector at the story layer are used to represent the text content of individual entity at both layers. We denote such text vector as $T(w_1, w_2, \dots, w_n)$. On the other hand, we need to model the text content of multiple terms at the cluster level, which was denoted as $TC(w_1, w_2, \dots, w_n)$. This is needed to build the linkage between text and visual features, which we will discuss in Section 4.1.2.2. Although there

are two types of text vectors, we employ the same text similarity measure and weighting scheme. Thus, in discussion on such issues, we use one notation T to represent both types of text vectors.

In this section, we first discuss the relationship between text features and visual semantics. We then present the method to establish the linkage between text and visual features. After that, we report our weighting scheme. Finally, we introduce our web-based concept similarity measure.

4.1.2.1 The relationship between text features and visual concepts

The relationship between text features and visual concepts at the MM discourse and story layer is not exactly the same, but similar. Spatially, the text descriptions do not always co-occur with the visual concepts at the shot level. Following, we discuss the situations at the MM discourse and story layer respectively.

In general, there are four types of relationships between keyword-based text semantics and shot-based visual semantics at the MM discourse layer.

- a) Type 1: We could infer the visual concept based on the text clues.

Figure 4.5 shows an example where we find text clue word “Clinton”, while the visual content showing “Clinton” simultaneously.

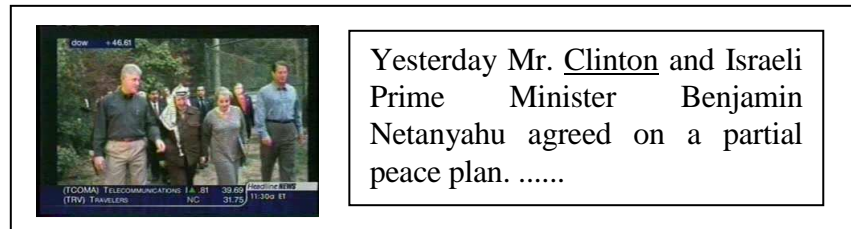


Figure 4.5: The text clue “Clinton” co-occurred with the visual concept.

- b) Type 2: We could find related text clue words, but the visual concept is not present. Figure 4.6 shows an example in which the keyword “Clinton” appears in the ASR transcripts, but we could not find the semantic concept “Clinton” occurring in the shot.

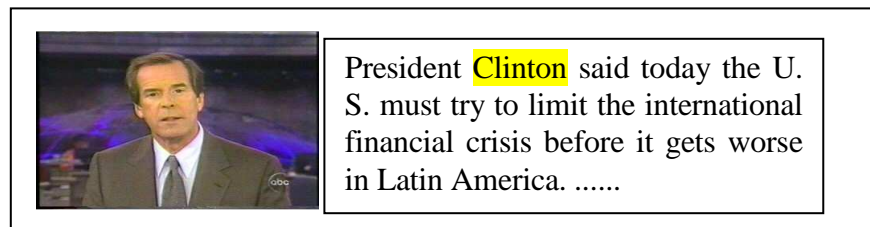


Figure 4.6: An example of when the text clues appeared, but the visual concept did not occur.

- c) Type 3: The visual concept is present but the related text clue words are absent. Figure 4.7 shows an example in which the concept occurs in the shot, but it is difficult to capture the text clues.

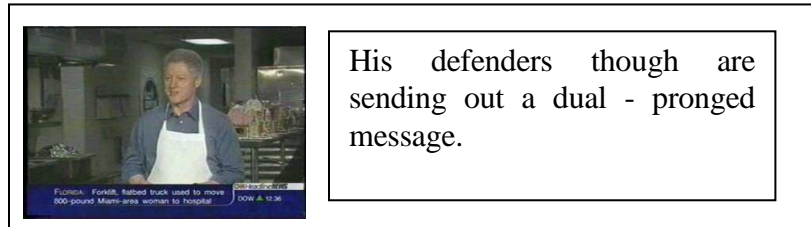


Figure 4.7: An example of when the visual concept occurred, but we could not capture the text clues.

- d) Type 4: Both text clue words and visual features are not available. Because it is not useful in concept detection, we will not discuss such a case in detail.

The relationship between text features and visual concepts at the story layer is similar as that at the MM discourse layer.

At the story layer, we attempt to capture the semantic concepts by exploring the relationship between the concept and the topics of a story. Generally speaking, at the story layer, if topics include the concept, we usually can find the visual semantic in the story. Figure 4.8 shows an example. However, we observe that not all the shots in the story include the concept “train”. Thus, how to establish the relationship between topics and visual content at the story layer is an important problem too.



Figure 4.8: Keyframes from shots and the topic vector in the story

4.1.2.2 Establish the relationship between text features and visual concepts

The above analysis highlights one challenge. That is how to find the terms (words and topic terms) from ASR transcripts to describe the image content. In our framework, we first cluster visually similar images together, which is done by a transductive learning algorithm. We then label the clusters by using the following approach.

For all terms appearing in the visual shot-based cluster (vcr_i), we first remove all stop terms [Salton and McGill, 1983]. We then assign the weight for the remaining term W_k using the following equation:

$$P(W_k, vcr_i) = \frac{NumofShotsInTheClusterIncludes(W_k)}{NumofShotsInTheCluster} \quad (4-2)$$

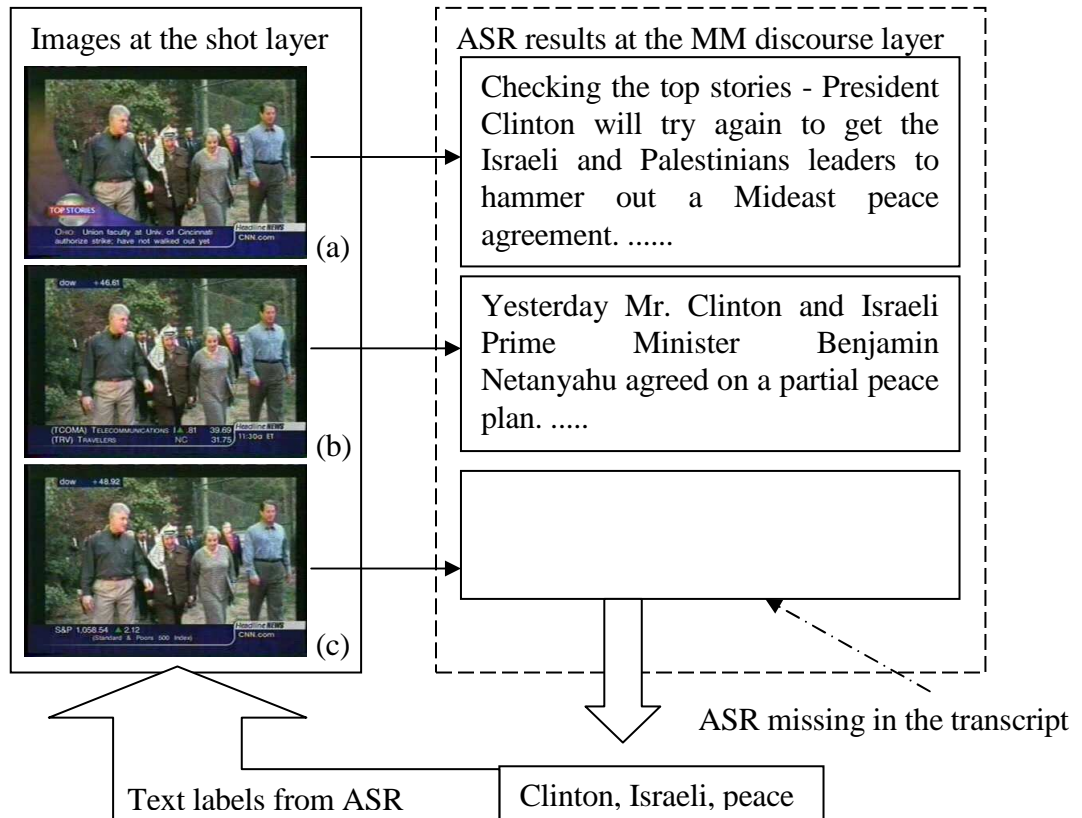


Figure 4.9: An example of labeling a visual cluster by text information

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

Figure 4.9 gives an example of visual cluster labeling by words at the MM discourse layer. The visual cluster result vcr_i is labeled by a word vector $TC_i = \{\text{Clinton, Israeli, peace}\}$.

However, in some cases, we could not find any text labels. This is because the

ASR words in a cluster exhibits large diversity. Figure 4.10 shows such an example. Because of such a characteristic, we could partially tackle the problem in Section 4.1.2.1 type 2, in which no labels for the cluster can be found. Hence MM discourse layer could not return any matching.

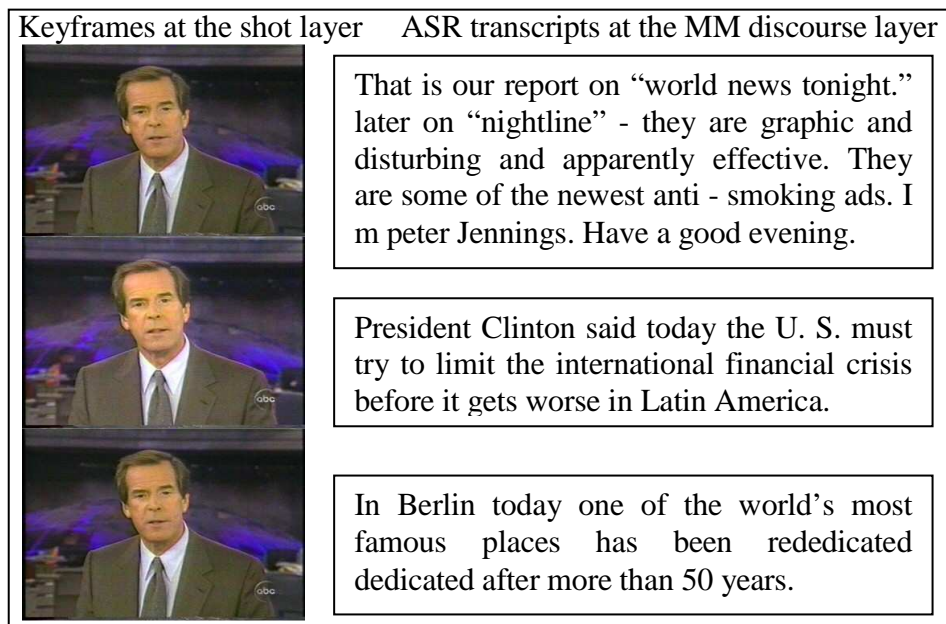


Figure 4.10: An example where no text label could be extracted from the image cluster.

However, Equation (4-2) could not solve the problem in Figure 4.7 at the MM discourse layer. In order to tackle this problem, we add text analysis at the story layer into the framework. The analysis at the story layer provides a global view to decide whether the shot includes the concept. At the story layer, we build the linkage between topic terms and visual contents in the same manner as that at the MM discourse layer, which uses Equation (4-2).

4.1.2.3 Word weighting

To improve the effectiveness of web search and text based transductive inference, we need to select a few dominant terms in the text vector. Here we employ a text-weighting scheme based on tf.rf developed in [Lan, et al. 2006] for text classification. 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 the test data; while the relevance frequency rf in the tf.rf approach does not consider terms only occurring 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:

$$Weight(W_i) = tf * [\alpha_w * \frac{\#(W_i, C_x)_{training}}{\#(W_i)_{training}} + (1 - \alpha_w) * \frac{\#(W_i, C_x)_{web}}{\#(W_i)_{web}}] \quad (4-3)$$

We obtain $\#(W_i, C_x)_{training}$ and $\#(W_i)_{training}$ by counting the co-occurrence between terms W_i and C_x , and the occurrence of term W_i in the training data, respectively; we obtain $\#(W_i, C_x)_{web}$ by using the concept text description C_x (such as “Clinton”, “Boat”, etc.) together with term W_i 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. α_w is designed to balance the training data and web statistics. We defined it as follows:

$$\alpha_w = \begin{cases} \text{Log}_{(\alpha+1)}(1 + tf) & \text{tf} < \alpha \\ 1 & \text{Otherwise} \end{cases} \quad (4-4)$$

where tf is the term frequency in the whole corpus and α is a predefined threshold. That is if the term is of sufficiently high frequency in the training data, 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.

4.1.2.4 Similarity measure

Computing semantic distances between two text vectors is one of the most important issues in the text-based clustering. Given two vectors, most systems adopt the cosine similarity measure. However, the cosine similarity measure considers the degree of the word overlapping. However, the same concept may be expressed in different word vectors that share few words. Figure 4.11 illustrates an example relating to concept “Clinton”. That is, the concept “Clinton” can occur in the different events such as the sex scandal and Middle East peace talk. However, due to different focuses, two reports may include

the same semantic concept but share few words in common. Thus, if we were to employ the cosine similarity, we could not detect that these two multimedia discourses are similar in the semantic view of the concept, say “Clinton” in Figure 4.11.

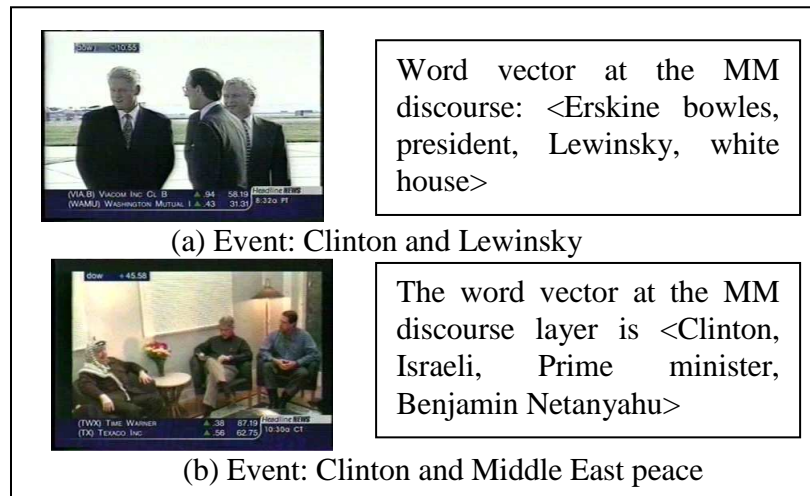


Figure 4.11: Two non-overlapping word vectors indicating a same concept “Clinton”

To overcome this problem, we propose a new web-based concept similarity measure. Such a method can assign a high similarity score to those word vectors with few or even non-overlapping words. For example: the text vector 1: < Erskine Bowles, president, Lewinsky, White House> should infer the occurrence of the concept “Clinton”; and the text vector 2: <Clinton, Israeli, Prime minister, Benjamin Netanyahu> should also infer the presence of the concept “Clinton”. Although there are no words that overlap, because of the high co-occurrence on the web statistics between the word vectors and the

concept text description “Clinton”, both word vectors indicate a high probability of the occurrence of concept “Clinton”. On the other hand, if a high amount of words overlap between two text vectors, such a method will be assigned a high similarity score too.

The definition of such a similarity measure is:

$$Sim_{unit}(T1, T2) = 1 - | P_{web}(C_x | T1) - P_{web}(C_x | T2) | \quad (4-5)$$

where T1, T2 are text vector instances, which is made of terms at the MM discourse or story layers. C_x is the concept text description (say “Clinton”).

The basic idea of this formula is that the similarity between two vectors is based on not only their contents, but also the co-occurrence relationship between the text content and the concept text descriptions. Because there are large amount data on the Web, we capture the co-occurrence relationship by counting web statistics, which is similar to the work of [Tan et al., 2008].

We obtain $P_{web}(C_x | T)$ in Equation (4-5) as follows:

$$P_{web}(C_x | T) = \frac{P(C_x, T)}{P(T)} = \frac{\frac{\#(C_x, T)}{\sum_{i=1}^n \#(T_i)}}{\frac{\#(T)}{\sum_{i=1}^n \#(T_i)}} = \frac{\#(C_x, T)}{\#(T)} \quad (4-6)$$

We obtain $\#(C_x, T)$, $\#(T)$ is computed in a similar manner as the variables in Equation (4-3).

In the following, we will demonstrate our new concept-based similarity measure step by step using the example in Figure 4.11.

Step1: We use the word vector {Erskine Bowles, president, Lewinsky, White House} as a query to Google search engine, which is shown in Figure 4.12.



Figure 4.12: The Google search results using {Erskine Bowles, president, Lewinsky, white house} as a query.

From the Google search results, we can find that there are approximately 839 documents that satisfy the query in the Google collection.

Step2: We need use the word vector T1 {Erskine Bowles, president, Lewinsky, White House} together with the concept description word “Clinton” as a query again to Google search engine, which is shown in Figure 4.13.

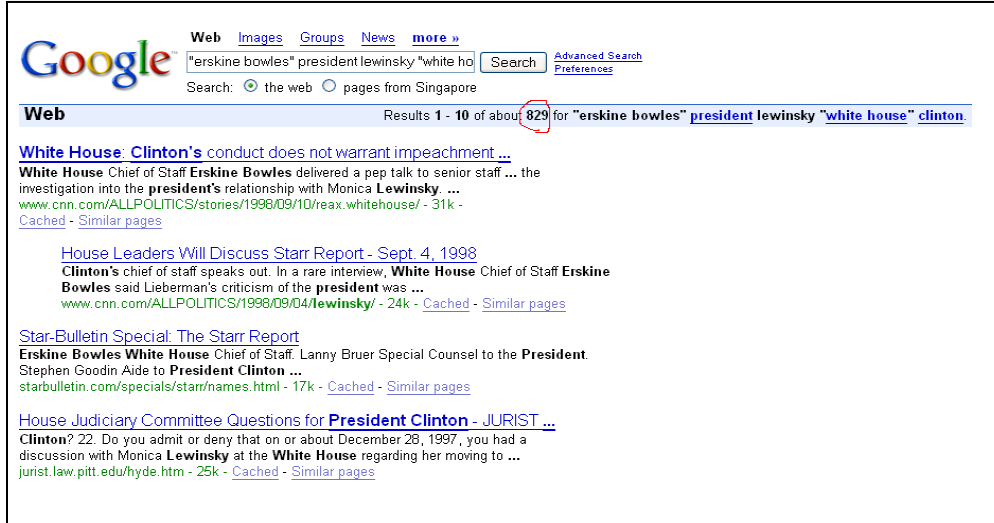


Figure 4.13: The Google search results using {Erskine Bowles, president, Lewinsky, White House} and “Clinton” as a query.

For this a query, we can obtain approximately 829 documents.

Step 3: We process the word vector T2 {Clinton, Israeli, Prime Minister, Benjamin Netanyahu} in a similar manner. The results are shown in Figure 4.14 and 4.15 with the same number of retrieval documents of 11,600.



Figure 4.14: The Google search results using {Clinton, Israeli, Prime Minister, Benjamin Netanyahu} as a query.

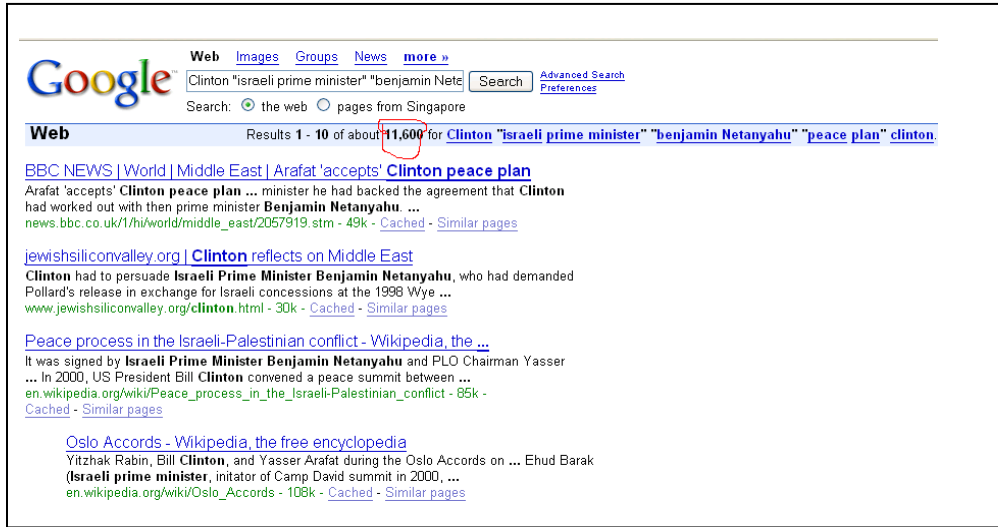


Figure 4.15: The Google search results using {Clinton, Israeli, Prime Minister, Benjamin Netanyahu} and “Clinton” as a query.

Based on the above retrieval results and by applying Equation (4-6), we obtain

$$P(\text{Clinton}|\text{T1}) = \frac{829}{839} = 0.988$$

$$P(\text{Clinton}|\text{T2}) = \frac{11600}{11600} = 1$$

We can substitute these two web-based similarity values in Equation (4-5) to compute the similarity between the two MM discourse text segments, and obtain the following result

$$\text{Sim}(S1, S2) = 1 - |0.988 - 1| = 0.988$$

Based on the above results, we can conclude that these two text segments are similar.

4.2 The multi-resolution constraint-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 characteristics 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 on constraints from different resolutions. At the shot layer, we first employ a visual based must-link constraint. That is, if both shot i and shot j are detected as anchorperson shots, then these two shots must be clustered together. After that, text constraints from lower resolutions are used to provide the cannot-link constraints that avoid the clustering of semantically dissimilar shots together. Figure 4.16 illustrates the cannot-link constraints from the 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 4.16. 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 are designed to avoid clustering the shots with a high visual similarity, but a low semantic similarity. 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 together, 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.

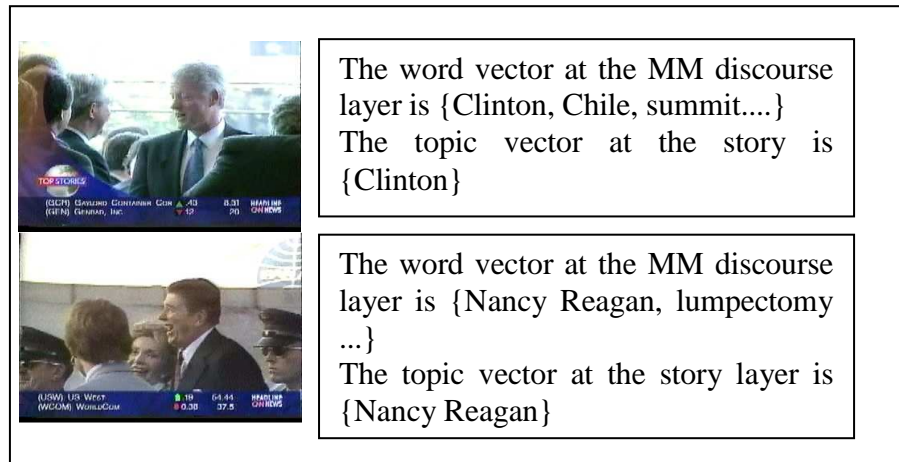


Figure 4.16: An example of using the cannot-link text constraints to purify the visual shot clustering results: the above images have high visual similarity, but when we consider text information, we find they have low possibility of including the same concept. Thus, the cannot-link constraints ensure that they are not clustered together.

Second, when using the text features at the MM discourse layer to further cluster the results from the first step, we make use of the must-link and cannot-

link constraints, to ensure high quality clusters too. We use the must-link constraints derived from visual shot clustering to ensure that two highly “visually” similar shots that were gathered in the shot layer analysis are remained clustered together 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 layer 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. That is, given two shots $S(i)$ and $S(j)$ with high text similarity at the MM discourse layer, however, 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 use the text features at the story layer to further cluster the shots based on the results at the MM discourse layer by utilizing the must-link constraints. 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 analysis.

Finally, we rank the results based on the shot, MM discourse and story layers. Because the cluster results at the shot layer is the purest, we have the highest confidence for the results and we assign them with 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 are able to derive good ranking of shots for each concept by using the full range of features.

CHAPTER 5

TRANSDUCTIVE INFERENCE

In this chapter, we first discuss the transductive algorithm. We then report our transductive-based cross-domain adaptation algorithm. Finally, we combine our M3 transductive inference with a bootstrapping technique.

5.1 Transductive inference

The transductive inference is used to analyze both the visual and text features at the different resolutions in our framework. It has two important functions. One is to capture knowledge from the training data. The other is to capture the distributions of training and test data well so that we have the knowledge to know when we can make an inference via training data. The differences between our design with the other works are listed as follows:

- We propose a new state “unknown” in transductive learning. The so-called unknown category of test data is that cannot be labeled as positive or negative with high confidence. If a test data belongs to such an unknown category, we can explore the relationship between concept text description and text content in the test data by using web co-occurrence to help infer the semantic label.
- We propose a novel multi-resolution based transductive learning inference. Thus, we could let the clusters in transductive learning approximately be as pure and as large as possible.

Generally speaking, transductive learning involves three stages. In stage 1, a series of clustering are applied as different inference hypotheses by using a constraint-based average-link clustering method at each resolution, which is discussed in Section 4.2. 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 stage 2 of transductive inference, a hypothesis is selected based on Vapnik combined error bound [Vapnik, 1998] to determine the confidence of the series of clusters. The basic idea of such a error bound is to minimize the inference risk in the test data. 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:

$$R_h(X_u) \leq R_h(X_m) + \sqrt{\left(\frac{m+u}{u}\right) \left(\frac{\tau + \log(C-1) + \ln \frac{1}{\delta}}{m}\right)} \quad (5-1)$$

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). $R_h(X_m)$ is the total number of positive and negative training data in the same clusters.

In stage 3, we label the test sample in the selected hypothesis by using the training data in the same cluster. We can label the type 1 cluster as positive (P) or negative (N) 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 U (Unknown set). That is, given a test shot S, appearing in a

visual based cluster vcr_i containing both training and test data, we compute the probability of C_x appearing in the cluster, or $P(C_x | S)$, as:

$$P(C_x|S) = \frac{P(C_x,S)}{P(S)} \approx \frac{\text{NumofTrainingShotsWith}(C_x)\text{IntheCluster}(vcr_i)}{\text{NumOfTrainingShotIntheCluster}(vcr_i)} \quad (5-2)$$

However, some clusters may include very few training data instance, 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 as follows:

$$CI = \begin{cases} \text{Log}_{(\lambda+1)}(1 + TD) & \text{TD} < \lambda \\ 1 & \text{Otherwise} \end{cases} \quad (5-3)$$

where TD represents the number of training data in a cluster and λ is the predefined threshold.

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

$$\text{Score}(S) = CI * P(C_x | S) * \log_2[1 + P(C_x | D_i)] \quad (5-4)$$

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, the value of $P(C_x | D_i)$ may be zero. This causes the score from Equation (5-4) for the test data in such domains is zero. Thus, it is difficult for us to rank them. In

order to tackle the problem, we employ an add-one smoothing method [Jurafsky and Martin, 2000] to estimate the probability of concept C_x in sub-domain D_i as:

$$P(C_x | D_i) \approx \frac{ShotWith(C_x)in(D_i) + 1}{ShotsIn(D_i)InTheTraining + 1} \quad (5-5)$$

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 (5-2). Thus, we adopt a multi-resolution analysis strategy and a web-based text retrieval approach to tackle this problem. The so-called multi-resolution strategy is to make an inference according to shot, MM discourse and story in order. In the following section, we mainly introduce our web-based text retrieval approach.

At the MM discourse layer, we bring text retrieval 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 semantic. In our text retrieval model, we make use of the web statistics. The difference between our design and the other works are shows as follows:

- There are some works in text retrieval that used Web statistics to expand the query words. Suppose we have a query word “Clinton”, the query expansion method may find many co-occurrence words for

“Clinton”, such as “President”, “Lewinsky”, and so on. Because the number of expansion words is limited, some words such as “Albright” may be not in the query expansion list. If we use the query and its expansion words in Figure 5.1 to do text retrieval, the document 1, 2, 3 may have similar relevance scores for the query “Clinton”. In fact, if a human being reads such documents, usually we can draw the conclusion that Document 1 is related to “Clinton” and Documents 2 and 3 are not. Thus, simply expanding query words via web statistics could not tackle the problem.

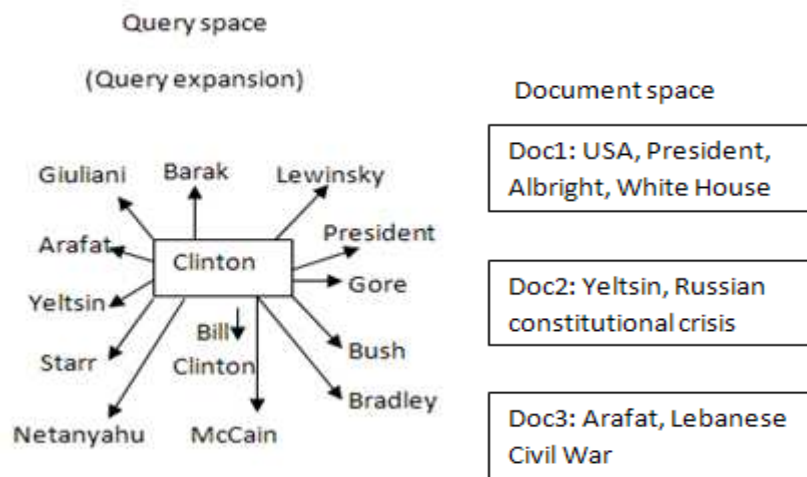


Figure 5.1: A traditional query expansion method that uses Web statistics

- In our retrieval model, we consider all text clues in the documents as a whole group, instead of a bag of independent words. Figure 5.2 demonstrates the idea in our design. Given a query, say, “Clinton”, we

did not expand a lot of related words. This is because usually one query may have many co-occurrence words. For example, the query “Clinton” is at least related to the following terms such as “Lewinsky scandal”, “Middle-East peace agreements” and so on. If we did not consider the retrieval targets (document space), it is hard to make a decision. Thus, some expansion words may fail to occur in the real target corpus. On the other hand, the list of query expansion words may fail to include some related query words in the target corpus. In order to tackle this problem, the basic idea of our design is to find the relationship between the query word and the text content in the target corpus. Generally, there are three steps. First, we establish the linkage between text features and visual clusters to obtain text labels (text clues) in the target news video corpus (test set). Next, because we know the date information of the news video in the TRECVID corpus, we can find documents on the web corpus with the same date as the target news corpus. These web documents would include all the text clues, such as “USA, President, Albright, White House”. Third, we check how many searched documents include the query “Clinton”. Based on the above statistics, we can estimate the relevance of the documents by using both the query words (concept text descriptions) and contents in the target corpus. Due to web redundancy, we can

approximately estimate the relationship between query and contents in document space. Thus, we could infer that the possibility of concept “Clinton” occurred in document 1 is higher than that in documents 2 and 3 for the case in Figure 5.1.

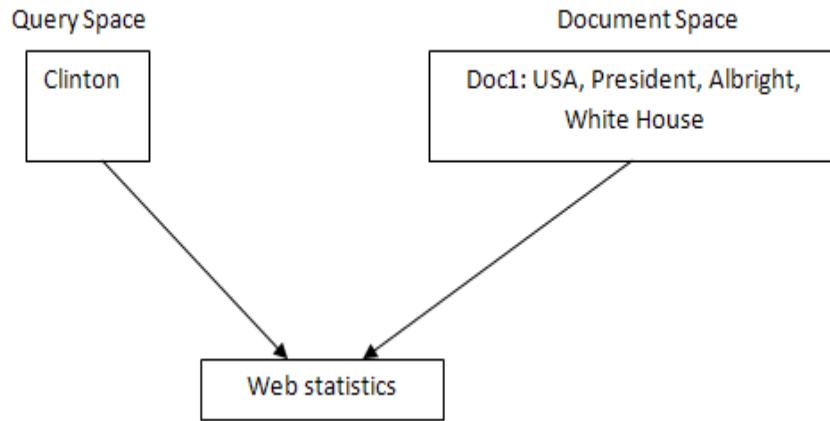


Figure 5.2: An example of our text retrieval model

Based on our discussion, the inference at the MM discourse layer is defined as follows: given a test shot S , we can find a MM discourse cluster $mmcr_j$, which includes the test shot S . The text label vector for the cluster is TC , which is obtained by Equation (4-2).

$$\text{Score}(C_x|S) = [CI * P_{\text{corpus}}(C_x|S) + [1 - CI] * P_{\text{web}}(C_x|TC)] * \log_2[1 + P(C_x|D_i)] \quad (5 - 6)$$

$$\begin{aligned}
P_{\text{Corpus}}(C_x|S) &= \frac{P_{\text{corpus}}(C_x, S)}{P_{\text{corpus}}(S)} \\
&\approx \frac{\frac{\text{NumofTrainingShotsWith}(C_x)\text{In}(\text{mmcr}_j)}{\text{TotalTrainingshots}}}{\frac{\text{NumofTrainingShotsIn}(\text{mmcr}_j)}{\text{TotalTrainingshots}}} \\
&= \frac{\text{NumofTrainingShotsWith}(C_x)\text{in}(\text{mmcr}_j)}{\text{NumofTrainingShotsIn}(\text{mmcr}_j)} \quad (5 - 7)
\end{aligned}$$

$$\begin{aligned}
P_{\text{web}}(C_x|TC) &= \frac{P_{\text{web}}(C_x, S)}{P_{\text{web}}(S)} \\
&= \frac{\frac{\#(C_x, TC)_{\text{web}}}{\text{TotalNumofWeb}}}{\frac{\#(TC)_{\text{web}}}{\text{TotalNumofWeb}}} \\
&= \frac{\#(C_x, TC)_{\text{web}}}{\#(TC)_{\text{web}}} \quad (5 - 8)
\end{aligned}$$

We obtain $\#(C_x, TC_i)_{\text{web}}$, $\#(TC_i)_{\text{web}}$ in a similar manner as that in Equation (4-3), CI is the confidence index and D_i is a sub-domain data set.

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

After each layer's 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 $Score(C_x | S_{\text{layer}}) > \alpha_{\text{layer}}$, we label it as positive data and put it into the P shot set.

- b) If $Score(C_x | S_{layer}) < \delta_{layer}$, 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.

5.2 Multiple sub-domain analysis

In this section, we discuss how to encode multiple sub-domain knowledge in our transductive inference framework. As discussed in the previous chapter, the sub-domain information is important. If we ignore the characteristics of sub-domain data or train a model via mixture of different sub-domain sets, we may lose information specific to the sub-domains and degrade the performance of the system. On the other hand, if we segment training data into several small data sets and use them separately, we have to face a problem of imbalanced distribution of training data in certain segments of sub-domains.

However, the existing cross-domain adaptation algorithms could not tackle the problem. This is partly because they adopted supervised learning approaches. 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 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, unless we manually label more data in the test set.

Here, we develop a pseudo-Vapnik combined error bound transductive learning approach to partially tackle this problem without additional manually assigning labels in the test data. As we have discussed in the previous section, 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, the function of Vapnik combined error bound is to select a cluster hypothesis. If there is very few or even no positive data, it is hard to compute the term $R_h(X_m)$ in Equation (5-1) accurately. To tackle this problem, we develop a pseudo-Vapnik combined error bound adaptation algorithm. Given that there is insufficient training data 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-domains with imbalanced 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 $X = \{X_1, X_2, \dots, X_{m+u}\}$;
 A training set with semantic labels $\{(X_1, Y_1), \dots, (X_m, Y_m)\}$.

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

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

Step 3: Place each sample in X as its own cluster, creating the list of clusters $C: C = c_1, c_2, \dots, c_{l+u}$

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

End while

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 5.3: A constraint based transductive learning algorithm

5.3 Multi-resolution inference with bootstrapping

We employ the bootstrapping technique to further process the unknown test results from our initial M3 transductive learning. 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, it is not always true. In order to reduce the errors from newly added unlabeled data, we propose a new bootstrapping algorithm, which is shown in Figure 5.4. The basic idea is that we use test data with high inference confidence to rerank the data in the unknown clusters from our initial M3 transductive model. The main differences between our approach and the other bootstrapping works [Feng et al. 2004] 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 lower value as compared to the labeled training data.
- (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 the shot layer; $i=2$ for MM discourse layer; and $i=3$ for story layer. Index j records the number of iteration 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 $K=C$, where C is a constraint (say $C=50$). 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 sequence in $[P(1)(0), P(2)(0), P(3)(0), \dots, P(i)(j)]$ is above the user's requirement or data propagation has converged, we stop the program.

Else, go to 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 sets $P(1)(j+1), P(2)(j+1), P(3)(j+1)$ and three negative sets $N(1)(j+1), N(2)(j+1), N(3)(j+1)$. The other set is still the unknown set $U(3)(j+1)$.

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

Step 6: Update j and K for next iteration, $j=j+1$; $K=K+C$; Go to step 2

Figure 5.4: Our bootstrapping algorithm

CHAPTER 6

EXPERIMENTS

In this chapter, we first introduce the corpus. We then report some baseline results from single modal systems, multi-modal fusion, sub-domain based multi-modal fusion, and sub-domain based multi-modal multi-resolution fusion. After that, we report the result from our multi-resolution, multi-source, multi-modal transductive learning framework. Finally, we compare our results with the reported systems on this corpus.

6.1 Introduction of our test-bed

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 Table 6-1. Although many works have been done in other TRECVID corpus, few works tackled the two problems discussed in this thesis. The two problems are: (a) how to let the evidence from text and visual features support each other to detect concepts, and (b) how to capture the characteristics of concepts via training data and concept descriptions. Most researchers focused on how to model good visual features such as parts-based object detection model [Zhang and Chang, 2005], SIFT features [Snoek et al. 2006] and concept relationship [Chang et al. 2006]. In order to evaluate how we tackle the two problems without affecting the other issues, such as machine translation errors, good visual features, concept relationship modeling and so on, we adopt the TRECVID 2004 data set and extract the common visual features and ASR results.

Table 6-1: Ten semantic concepts used in TRECVID

1	2	3	4	5
Boat	Madeleine Albright	Bill Clinton	Train	Beach
6	7	8	9	10
Basket Scored	Airplane takeoff	People walking and running	Physical violence	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 metric used in TRECVID 2004. The value of MAP is the mean of the average of precisions over all relevant judged shots. Hence, it combines precision and recall into one performance value. Let $p^k = \{i_1, i_2, \dots, i_k\}$ 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 the MAP for the ten concepts is defined as:

$$MAP = \frac{1}{10} \sum_{C_i=1}^{10} \left[\frac{1}{R} \sum_{k=1}^A \frac{R \cap p^k}{k} \varphi(i_k) \right] \quad (6-1)$$

where the indicator 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.

6.2 Test 1: Concept detection via single modality

analysis

In this section, we evaluate the performance of using the single modal feature, text or visual so that we can observe the description power of individual modality analysis.

6.2.1 Concept detection by using text feature

We first 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}⁹ as the classifier. The other is text retrieval, which we adopt a state of the art retrieval system [Cui et al. 2004] with query expansion techniques using external knowledge. For completeness, we also explore the combinations of both methods using the following equation:

$$Score(S) = \alpha * Score_{IR}(S) + (1 - \alpha) * Score_{TCL}(S) \quad (6-2)$$

⁹ <http://svmlight.joachims.org/>

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

Figure 6.1 lists the results based on text classification and retrieval at the shot, MM discourse and story layer respectively. We experiment with different values of α ranges from 0 to 1, and report only the increment of α at 25% interval.

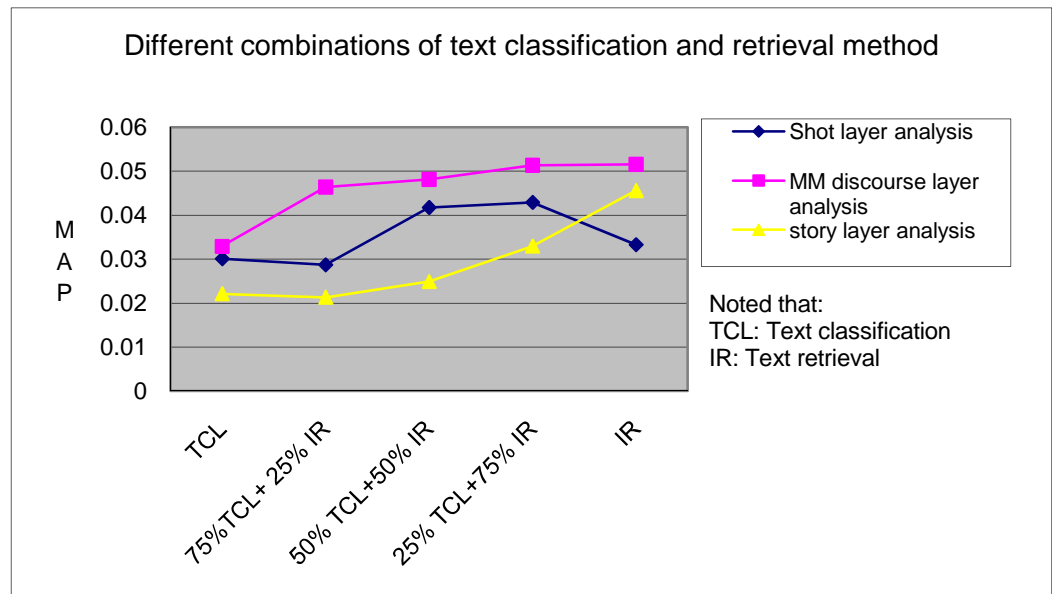


Figure 6.1: The results of combining two types of text analysis

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 only obtain fragmented text clues;

whereas systems based on the story boundaries tend to cover a large number of shots and hence could obtain higher recall, but lower precision.

- 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 [Naphade and Smith, 2004] and text retrieval method tends to perform better than the text classification method under such circumstances.
- Although we tried different setting for the combinations of text classification and retrieval method, no combinations 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 the strengths and weaknesses of different methods in detail.

6.2.2 Concept detection by visual feature alone

We employ two types of machine learning methods to detect concepts by using visual features. One is a supervised learning method, which is based on the SVM^{light}. The other is a transductive learning method, in which we adopt

the method discussed in Chapter 5. The results of comparing SVM against transductive learning are listed in Figure 6.2.

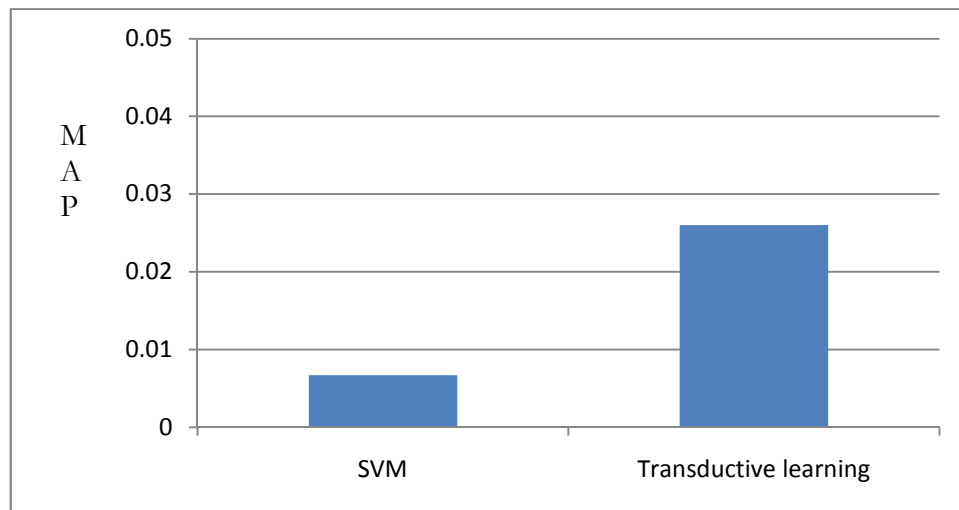


Figure 6.2: Two types of machine learning methods that detect concepts by using visual features alone

From the figure, we can derive the following observations:

- The result based on transductive learning is better than that of SVM approach. This is partly because transductive learning could capture the distribution of whole corpus so that we can obtain a better hypothesis.
- In addition, if we make analysis based on the shot layer using only visual features alone, we achieve a very low MAP of 0.026, which is much lower than that achievable using the text retrieval method (at MAP=0.051, See Figure 6.1). This shows that when the performance

of the ASR results are good such as ABC and CNN news transcriptions, the use of text is superior to that of using only the visual feature. Hence, text analysis helps in visual analysis.

6.3 Test 2: Multi-modal fusion

We employ the early fusion and late fusion [Snoek et al. 2006] to detect concepts by using text and visual feature at the shot layer. These are the state of the arts systems. Figure 6.3 shows results based on the early and late fusion, respectively.

From the figure, we can draw the following conclusions:

- The performance of concept detection by multi-modal fusion is better than those of single modal detectors. This is because multi-modal features provide more information than single modality alone. However, how to fuse multi-modal features affects the performance of the system. We note that the performance of the late fusion is better than that of early fusion strategy.
- The result from the text retrieval system at the MM discourse layer (the best text analysis result) is comparable to those from early and late fusion approaches. In fact, the results are only slightly worse than that using the late fusion strategy. This suggests that the choice of a good unit to analyze

features is a very important issue.

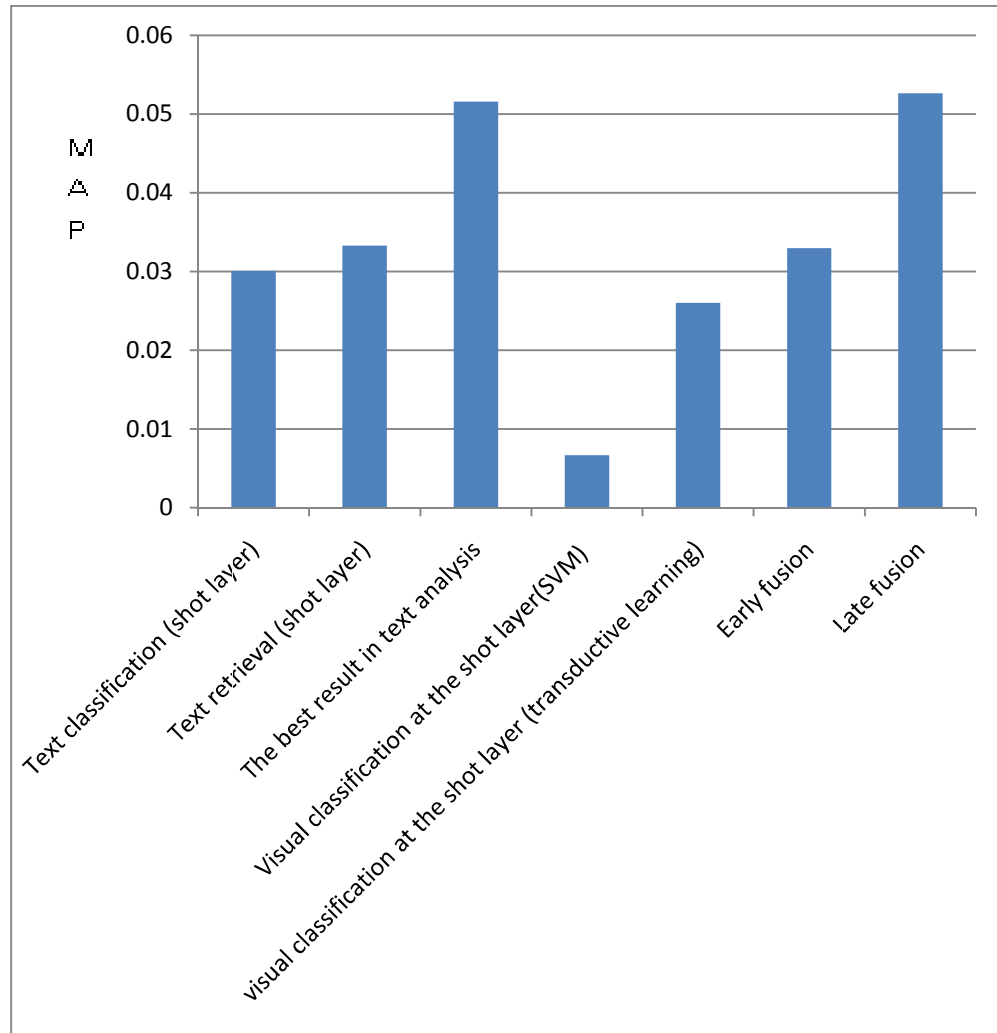


Figure 6.3: Concept detection using single modality versus multi-modality

6.4: Test 3: Encode the sub-domain knowledge

As we discussed earlier, sub-domain knowledge is an important information source. We encode it into the early and late fusion framework by using the

following equation.

$$\text{Weight}(S) = \text{Positive}(\text{Score}(S)) * P(C_x|D_i) \quad (6-3)$$

where $P(C_x|D_i)$ is defined as same as Equation (5-5), which encodes sub-domain knowledge and $\text{Score}(S)$ is the result from the SVM classifier. The definition of the function $\text{Positive}()$ is as follows:

$$\text{Positive}(\text{Score}(S)) = \text{Score}(S) + \text{offset} \quad (6-4)$$

Because the score from a SVM can be positive or negative, in order to rank the test shots by combining the knowledge from sub-domain and multi-modal feature space, we add a constant (offset=10) to let all the value to be positive.

Figure 6.4 presents the results of fusion system with and/or without encoding the sub-domain knowledge.

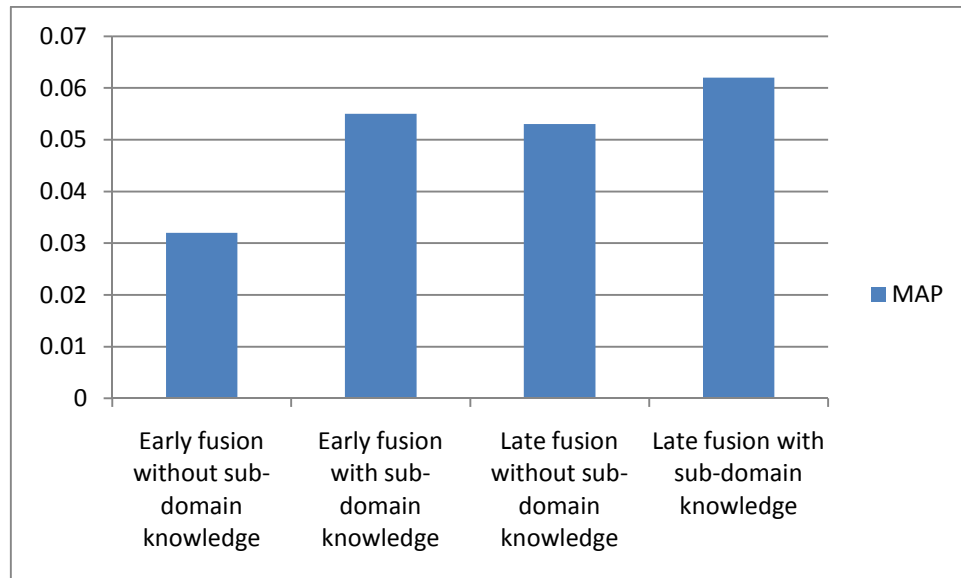


Figure 6.4: The systems with / without sub-domain knowledge

From the figure, we find that if we encode the sub-domain knowledge into the frameworks, we can achieve significant improvement in both the results of using early fusion and late fusion. This suggests that the use of sub-domain knowledge is beneficial to the concept detection task.

6.5 Test 4: Multi-resolution multimodal analysis

In the previous section, we observe that the performance of multimodal analysis is usually better than that of single modal analysis. However, how to fuse multimodal features is still a problem. We believe different modal features work well at the different resolutions and different resolutions have different types of semantics. Thus, we should fuse multimodal features at different resolutions. In this section, we first introduce a baseline multi-resolution system and then report our M3 transductive framework results.

6.5.1 A baseline multi-resolution fusion system

We build a baseline multi-resolution multimodal system to demonstrate the effective of our proposed multi-resolution fusion method. At each resolution, we select the method with the best performance in the previous experiment to perform concept detections. At the shot layer, we adopt transductive learning using visual features. At the MM discourse and story layers, we employ text

retrieval methods. We combine the results from such three layers via a linear combination using Equation (6-5).

$$Weight(S) = \alpha * Score_{Shot} + \beta * Score_{MMdiscourse} + \gamma * Score_{story} \quad (6 - 5)$$

where $Score(S)_{shot}$ is the inference score at the shot layer by using a transductive learning, which is defined in Equation (5-2); $Score(S)_{MMdiscourse}$ and $Score(S)_{story}$ are the inference scores at the MM discourse and story layer by using text retrieval. We define them as follows:

$$Score(S)_{unit} = \frac{\sum_{i=1}^n Rel(word_i)}{n} \quad (6 - 6)$$

where a unit can be a MM discourse and story, n is the number of terms in the query expansion list and the function Rel (word) is defined the relevance of query words from the system [Cui et al. 2004].

Table 6-2: The setting parameters of the linear combination

	Shot layer (α)	MM discourse layer (β)	Story layer (γ)
Fusion System 1	0.25	0.25	0.5
Fusion System 2	0.25	0.5	0.25
Fusion System 3	0.5	0.25	0.25

Figure 6.5 presents the results of different fusion schemes based on the

parameter setup as shown in Table 6.2. The results presented do not make use of sub-domain knowledge.

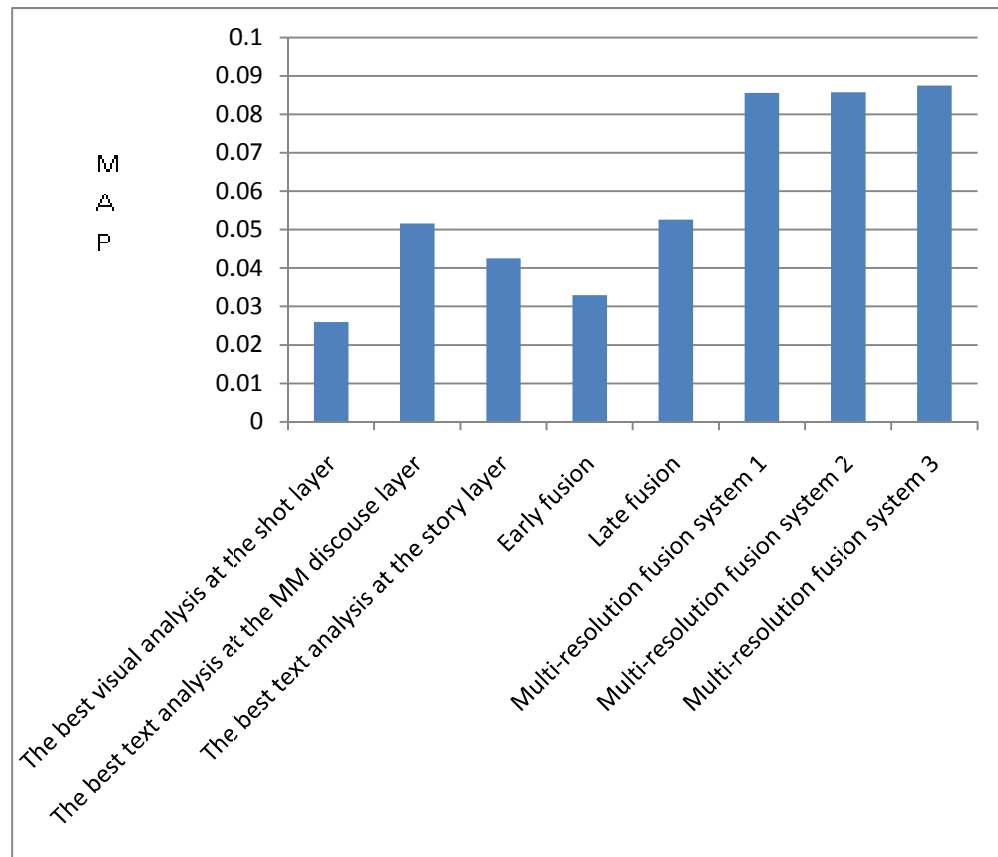


Figure 6.5: Results of single resolution fusion vs. multi-resolution fusion without using sub-domain knowledge

From the figure, we can derive the following observations:

- The performances of multi-resolution fusion systems are better than obtained using a single resolution analysis. The best result of a single resolution at MAP of 0.053 comes from the late fusion strategy (see

Figure 6.3). On the other hand, all the results obtained from multi-resolution fusion systems are above the MAP of 0.086. The best result among the multi-resolution fusion systems come from system 3, with an MAP of 0.087. This result amounts to over 64% improvement in MAP performance as compared to the best of single resolution analysis method based on the late fusion scheme.

In order to evaluate the effects of sub-domain knowledge, we conduct a further experiment that encodes the sub-domain knowledge into the multi-resolution framework. We compute the weight of each test shot by using the following equation.

$$Weight(S) = \left(\alpha * Score_{shot} + \beta * Score_{MM_{discourse}} + \gamma * Score_{story} \right) * P(C_x|D_i) \quad (6-7)$$

where $P(C_x|D_i)$ is defined as same as Equation (5-5) and the definitions of the three occurrence of the function “Score” are described by Equation (6-5).

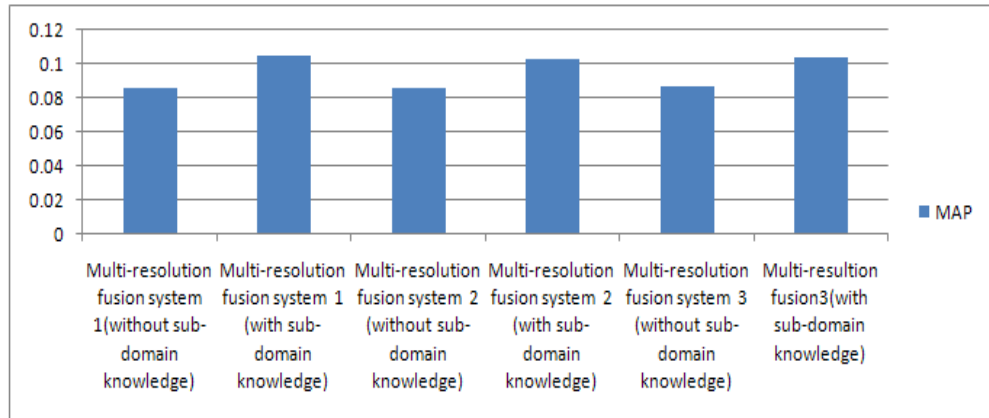


Figure 6.6: Multi-resolution systems with and/or without sub-domain knowledge

Figure 6.6 presents the results of multi-resolution analysis with and without the use of sub-domain knowledge. From the figure, we observe that the use of sub-domain knowledge could general improve the performance of the systems from the MAP of 0.086 to 0.105, which a 21% relative improvement. This confirms that the use of sub-domain knowledge is effective in the concept detection task.

6.5.2 Our proposed approach

We employ our M3 transductive framework as discussed in Chapters 3, 4, 5. In particular, we perform three experiments: (a) Transductive learning based on shot layer visual analysis without text; (b) shot layer + MM discourse layer analysis; (c) our full M3 model with story layer analysis, and (d) our full

model + bootstrapping. The results are shown in Figure 6.7.

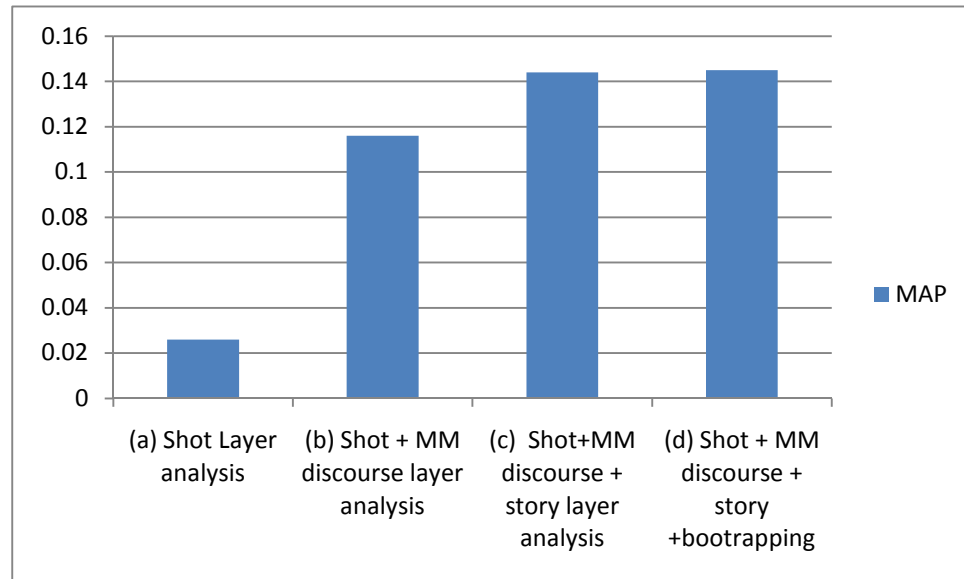


Figure 6.7: The result based on the shot layer analysis and different combinations of multi-resolution analysis

From the figure, we can observe that the performance of multi-modal fusion is better than that of single modal analysis. This is demonstrated in runs (b) and (c) that incorporate text semantics. In particular, run (b) which incorporates text features at the MM discourse layer achieves a substantially improved result at MAP of 0.116; while the better result is achieved when we perform the full multi-resolution analysis at the shot, MM discourse and story level, with a MAP of 0.144. In addition, run (d) 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 [Hull,

1993] ($p < 0.05$). This shows that the bootstrapping method is feasible.

Figure 6.8 compares our M3 model and the baseline multi-resolution linear fusion systems with the use of sub-domain knowledge.

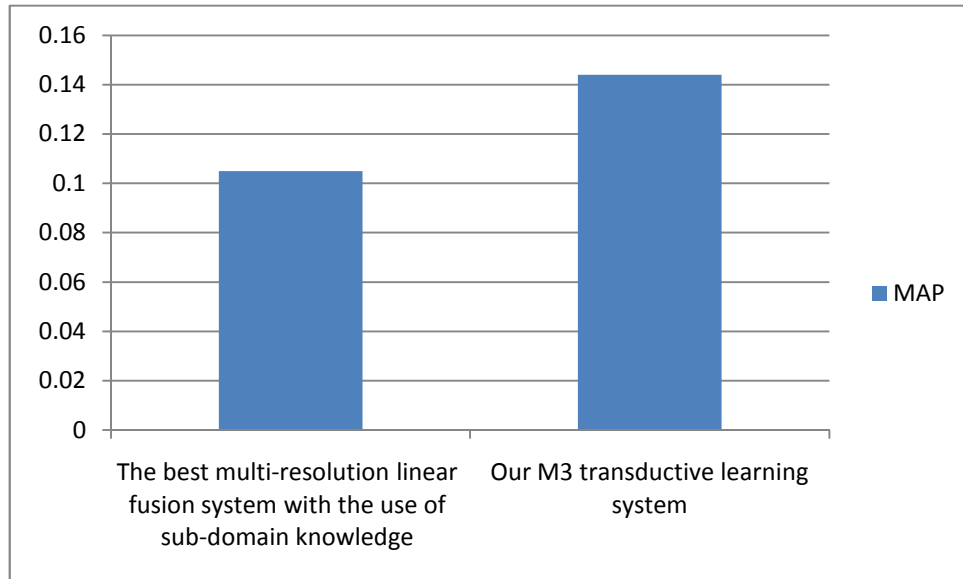


Figure 6.8: Two types of multi-resolution fusion systems

From Figure 6.8, we found that our M3 transductive model achieves an improvement of 37% over the best baseline multi-resolution system. We are able to achieve the better performance, mainly because:

- a) We employ the must-link and cannot-link constraints and multi-resolution inference structure so that we can let evidence from different resolutions support each other. On the other hand, without such multi-resolution strategies, there were many false alarms and misses from the

transductive learning and text retrieval inference. These cause the performance of the baseline system to be worse than our M3 model.

- b) We combine transductive inference and web-based text retrieval model at the MM discourse and story layer so that we can exploit the training, concept text description and web statistics at each resolution. On the other hand, for the baseline system, it is hard to combine the SVM results and text retrieval results. Also, without encoding the knowledge from training data, the performance of the system will be degraded.

6.6 Test 5: The comparison of M3 model with other reported systems

In order to compare our results with other reported systems, we tabulate the results of all reported systems [TRECVID 2004] that have completed all ten concepts in Figure 6.9. We can divide the reported systems into three categories. They are text retrieval systems, machine learning system using text and visual features and the linear combination of text retrieval and machine learning based systems. In Figure 6.9, we include the results of our four systems as listed in Figure 6.7.

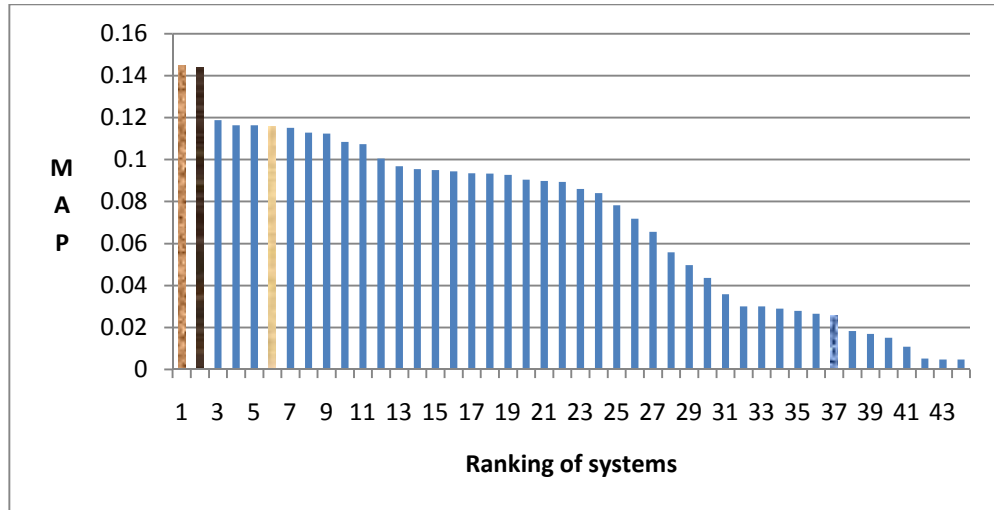


Figure 6.9: Comparison with other reported systems in TRECVID.

From Figure 6.9, we also observe that our four combinations of M3 models tested in Figure 6.7 are ranked as 1st, 2nd, 6th, and 37th, respectively. As compared to the best reported system which is ranked 3rd in Figure 6.9, our M3 transductive framework is able to achieve more than 22% improvement in MAP performance. This clearly demonstrates that our M3 model is superior.

From our prior analysis, we found that these current systems have the following two problems:

- a) It is difficult for the current systems to allow the evidence from different modalities to support each other.
- b) The performance of these supervised inductive inference approaches 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 degrade drastically. On the other hand, the text retrieval based system failed to make use of knowledge in the training data. Thus, the ability to combine both advantages from machine learning method and text retrieval is an important problem.

We believe that our M3 framework provides a novel multi-resolution solution to integrate multimodal features naturally that partially tackles problem 1 as follows:

- It allows the visual analysis to support text analysis. For example, if we were to rely on just text analysis without visual clustering at the shot level to group visually relevant shots we would have captured some false positive shots such as those illustrated in Figure 4.6, and missed some relevant shots such as shown in Figure 4.9 (c).
- It permits text analysis to leverage on visual detection. For example, if we were to rely on just visual analysis, shots with high visual similarity but large semantic variance will not be grouped together such as in Figure 4.16. In addition, without text analysis, shots with large visual variance but sharing the same concept are not detected.

In addition, our multi-source transductive model provides a novel solution to combine the training and external web knowledge as illustrated in Figure 6.10.

We observe that there are significant difference between training and test data

on the concept “train” from both the text and visual features. Thus, we failed to find similar shots in Figure 6.10 (b) and (c) from the training data by using transductive inference. In these cases, the external knowledge from the web statistics played a dominant role in capturing the evidence {train, freight, storm} and {train, conductor, jelly, country} for the above two examples at the MM discourse and story layer respectively. This is the reason that the performance of our system is better than those using training data only.

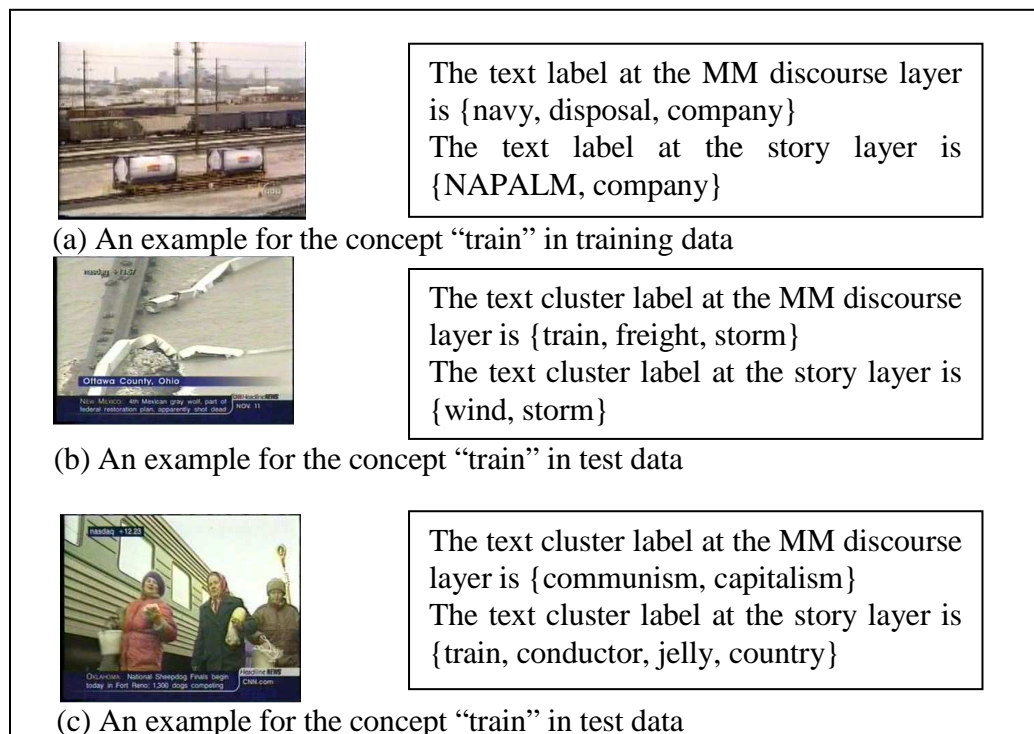


Figure 6.10: An example of our M3 transductive framework on the concept “train”

CHAPTER 7

CONCLUSION AND FUTURE WORK

In this thesis, we have identified two important weaknesses in current research on semantic concept detection. We proposed a multi-resolution, multi-source and multimodal transductive learning framework to tackle these problems. In this chapter, we recap our contributions of the research and summarize them in the next subsections. We will then discuss the limitations of this work. Finally, we conclude the thesis with an outline of future work.

7.1 Contributions

Although research on semantic concept detection has been carried out for many years, the analysis based on multi-resolution and the combination of the knowledge from the training labels, concept text descriptions and web statistics via a transductive learning model has been relatively recent. In this thesis, we make the following contributions:

- **A novel multi-resolution multimodal fusion model**
- **A novel multi-source transductive learning model**

7.1.1 A novel multi-resolution multimodal fusion model

We developed a multi-resolution model to fuse of text and visual features in news video to detect semantic concepts. The multi-resolution model contributes to the field of multimedia processing. Fusion of multimodal features has been developed for many years. However, we found that most of the existing works only focus on single resolution (usually at the shot layer) fusion. Such efforts suffered from the mismatch between text and visual features at the shot layer.

In contrast, we perform our analysis at the shot, multimedia discourse and story layers to tackle the mismatch that occurs when using resolution at the shot layer. Furthermore, our multi-resolution model can capture different types of semantics at different resolutions. More importantly, our framework allows evidence from text and visual features to support each other. Thus, our framework achieves the aim of processing multimedia content in a unified framework, instead of processing multimodal features independently.

7.1.2 A novel multi-source transductive learning model

This work contributes to machine learning in multimedia applications. Most current efforts detect concepts by using machine learning methods, text retrieval methods and their combinations. However, both the first two types of methods have their strengths and weaknesses. Without deep analysis of their characteristics, it is hard to obtain a good result by the combination. In this thesis, we propose a multi-source transductive learning model to combine the two methods together. It leverages on both of the training data and test data, along with others sources of information to reduce the reliance on training data. Transductive learning captures the data distribution of the corpus and makes an inference based on training data. Text retrieval, as a smoothing method, is employed to process the test data with unknown labels.

Based on the evaluation results, our system outperforms the reported systems.

7.2 Limitations of this work

This thesis has contributed in narrowing the semantic gap. However, it has several limitations that need to be addressed.

- **Transductive learning is a very time consuming process.**

Because transductive learning exhaustively analyzes all possible relationships between training and test data instead of using training data alone, it is a very time consuming process. In order to speed up transductive learning, we should carry out further theoretical studies especially in mathematical optimization. In fact, many groups complained about the large computational efforts of multimedia analysis, even under the supervised learning frameworks. For example, Snoek et al. [2004] estimated that the processing of the entire TRECVID 2004 data set would have taken over 250 days on the fastest available sequential machine at that time. Cao et al. [2006] claimed the same problem in their report. They estimated that they need over 600 days for one computer to complete their algorithms for semantic concept detection. The efficiency problem in concept detection task is still a big obstacle.

- **Multi-label annotation**

It is an important trend to use relationships among concepts to help concept detection. However, in real news reports, new concepts such as names, events and so on are always occurring. The relationship among concepts may also change over time. Although in our framework, we do not pre-define the list of concepts, each non-stop-word in the corpus can represent a concept. We attempt to encode the relationship of concepts using web statistics. However, we still have difficulty in making use of concept relationships in visual content and fusing them with the text component.

7.3 Future work

We summarize our plan for future research.

- **Integration of corpus knowledge with the knowledge from human annotations and manually built encyclopedia**

Under current frameworks, the preparation of training data and analysis of corpus data via transductive learning is independent. The random sampling or time period selection of training data causes at least the following problems:

- Training data usually cannot include all typical scenarios in the test data. Thus, there is a large gap between the distributions of data in training and testing.
- In multimedia processing, the problem of imbalanced training data usually exists. That is, the number of negative label training samples is significantly larger than that of the positive samples. In fact, only those negative training samples that are similar to those positive training samples are useful.
- Without the support of automatic data analysis tools, human annotators have to repeatedly assign labels to similar multimedia contents in different videos as shown in Figure 7.1. Such efforts are time-consuming and error-prone.



Figure 7.1: Repeatedly labeling for similar images in the different videos.

Because the interactive combination of automatic data analysis and manual labeling is another new and challenging task, we did not

incorporate it in our M3 framework. Thus, in the near future, we plan to develop a new interactive concept detection system. With the help of automatic corpus data analysis such as clustering together with the use of active learning approach, we can identify typical data for manual labeling. This will reduce the amount of annotation efforts and boost system performance.

- **Optimization of transductive learning algorithm**

We plan to explore techniques to speed up the transductive learning. Current transductive learning algorithm assumes that the entire test data is given. However, we often need to process new data, after we have processed existing test corpus data. Thus, how to incrementally make use of the results in old test data, instead of re-computing everything from scratch, is another optimization problem. We will explore an efficient algorithm for transductive learning.

- **Encoding concept relationships in concept detection**

Encoding concept relationships is an important problem in concept detection. There are two types of concept relationships. One is static and the other is dynamic. Current systems only encode the static relationship among concepts. However, the contents of news are always dynamically changing. We should consider dynamic concept relationship in the concept detection framework.

- **Develop more effective visual features and visual models**

Because our goal is to detect visual semantics in the video, it is important to develop effective visual features and models to capture the visual semantics.

BIBLIOGRAPHY

- A. Amir, G. Iyengar, C. Y. Lon, C. Dorai, M. Naphade, A. Natsev, C. Neti, H. Nock, I. Sachdev, J. Smith, Y. Wu, B. Tseng, and D. Zhang, “IBM Research TRECVID 2003 Video Retrieval System”, Proceedings of TRECVID 2003, Gaithersburg, MD, November 2003. Available at: <http://www-nlpir.nist.gov/projects/tvpubs/tv3.papers/>.
- A. Amir, G. Iyengar, J. Argillander, M. Campbell, A. Haubold, S. Ebadollahi, F. Kang, M. R. Naphade, A. (P.) Natsev, J. R. Smith, J. Tešić, and T. Volkmer, “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/>.
- R. Baeza-Yates and B. Ribeiro-Neto, “Modern Information Retrieval”, ACM Press, Addison-Wesley, 1999.
- D. Blostein and N. Ahuja, “Shape from Texture: Integrating Texture-element Extraction and Surface Estimation”, IEEE Transaction on Pattern Analysis and Machines Intelligence vol 11, pp. 1233-1251, 1989.
- A. Blum and T. Mitchell, “Combining Labeled and Unlabeled Data with Co-Training”, Proceedings of the Workshop on Computational Learning Theory, pp. 92-100, 1998.

- M. Campbell, S. Ebadollahi, D. Joshi, M. Naphade, A. Natsev, J. Seidl, J. R. Smith, K. Scheinberg, J. Tešić, L. Xie and A. Haubold, “IBM Research TRECVID-2006 Video Retrieval System”, Proceedings of TRECVID 2006, Gaithersburg, MD, November 2006. Available at: <http://www-nlpir.nist.gov/projects/tvpubs/>.
- J. Cao, Y. Lan, J. Li, Q. Li, X. Li, F. Lin, X. Liu, L. Luo, W. Peng, D. Wang, H. Wang, Z. Wang, Z. Xiang, J. Yuan, W. Zheng, B. Zhang, J. Zhang, L. Zhang, and X. Zhang, “Intelligent Multimedia Group of Tsinghua University at TRECVID 2006”, Proceedings of TRECVID 2006, Gaithersburg, MD, November 2006. Available at: <http://www-nlpir.nist.gov/projects/tvpubs/>.
- L. Chaisorn, “A Hierarchical Multi-Modal Approach to Story Segmentation in News Video”, Ph.D. thesis in National University of Singapore, 2004.
- S. F. Chang, “Advances and Open Issues for Digital Image/Video Search”, Keynote Speech at International Workshop on Image Analysis for Multimedia Interactive Services, 2007. Available at: <http://www.ee.columbia.edu/%7Esfchang/papers/talk-2007-06-WIAMIS-Greeceprint.pdf>.
- S. F. Chang, W. Hsu, W. Jiang, L. Kennedy, D. Xu, A. Yanagawa, and E. Zavesky, “Columbia University TRECVID-2006 Video Search and High-Level Feature Extraction”, Proceedings of TRECVID 2006. Available at: <http://www-nlpir.nist.gov/projects/tvpubs/>.

- 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, pp. 1005-1008, 2005.
- L. P. Chen and T. S. Chua, "A Match and Tiling Approach to Content-based Video Retrieval", Proceeding of IEEE International Conference on Multimedia and Expo, pp. 301-304, 2001.
- China Information Center, "The 1st Statistical Survey Report on the Internet Development in China", 1997. Available at: <http://cnnic.cn/download/2003/10/13/93603.pdf>.
- 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.
- T. S. Chua, S. Y. Neo, K. Y. Li, G. Wang, R. Shi, M. Zhao, and H. Xu., "TRECVID 2004 Search and Feature Extraction Task by NUS PRIS" Proceedings of (VIDEO) TREC 2004, Gaithersburg, MD, November 2004. Available at: <http://www-nlpir.nist.gov/projects/tvpubs/>.
- T. S. Chua, C. H. Goh, B. C. Ooi, and K. L. Tan, "A Replication Strategy for Reducing Wait Time in Video-On-Demand Systems", Journal of Multimedia Tools Application, 15(1): pp. 39-58, 2001.

- G. Cortelazzo G. A. Mian, G. Vezzi and P. Zamperoni, “Trademark Shapes Description by String Matching Techniques”, Pattern Recognition vol 27, pp. 1005-1018, 1994.
- 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>.
- P. K. Davis and J. H. Bigelow, “Experiments in Multiresolution Modeling (MRM)”, 1998. Available at <http://www.rand.org/publications/MR/MR1004/>.
- R. O. Duda, P. E. Hart and D. G. Stork, “Pattern Classification”, Wiley Interscience, 2004.
- P. Duygulu, K. Barnard, J. de Freitas and D. Forsyth. “Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary”, Proceedings of European Conference on Computer Vision, vol.4, pp. 97-112, 2002.
- P. Duygulu, M. Y. Chen, and A. Hauptmann, “Comparison and Combination of Two Novel Commercial Detection Methods”, Proceedings of the 2004 International Conference on Multimedia and Expo (ICME' 04), vol. 2, pp. 1267 – 1270, 2004.

- J. Evans, "The Future of Video Indexing in the BBC", 2003. Available at:
<http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html>.
- H. M. Feng, R. Shi, and T. S. Chua, "A Bootstrapping Framework for Annotating and Retrieving WWW Images", Proceeding of the 12th ACM International Conference on Multimedia, pp. 960-967, 2004.
- J. L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System", Speech Communication, 37(1-2) pp. 89-108, 2002.
- U. Hahn, "Topic Parsing: Accounting for Text Macro Structures in Full-text Analysis", Information Processing and Management, 26 (1): pp. 135-170, 1990.
- A. Hauptmann, "Lessons for the Future from a Decade of Informedia Video Analysis Research", Proceedings of the 4th International Conference on Image and Video Retrieval, pp. 1-10, 2005.
- A. Hauptmann, R. Yan, Y. Qi, R. Jin, M. Christel, M. Derthick, M. Y. Chen, R. Baron, W. H. Lin, and T. D. Ng, "Video Classification and Retrieval with the Informedia Digital Video Library System", 2002. Available at:
<http://www-nlpir.nist.gov/projects/tvpubs/>.
- A. Hauptmann, R. V. Baron, M. Y. Chen, M. Christel, P. Duygulu, C. Huang, R. Jin, W. H. Lin, T. Ng, N. Moraveji, N. Papernick, C. G. M. Snoek, G. Tzanetakis, J. Yang, R. Yan, and H. D. Wactlar, "Informedia at TRECVID 2003: Analyzing and Searching Broadcast News Video", Proceedings of

TRECVID 2003, Gaithersburg, MD, November 2003. Available at:
<http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html#2003> .

A. Hauptmann, M. Y. Chen, M. Christel, W. H. Lin, R. Yan, and J. Yang
“Multi-Lingual Broadcast News Retrieval”, Proceedings of TRECVID
2006. Available at: <http://www-nlpir.nist.gov/projects/tvpubs/>.

A. Hauptmann and M. Witbrock, “Story Segmentation and Detection of
Commercials in Broadcast News Video”, Advances in Digital Libraries
Conference, pp. 168-179, 1998.

M. A. Hearst, “Context and Structure in Automated Full-Text Information
Access”, Ph.D. thesis, University of California at Berkeley, 1994.

W. Hsu, S. F. Chang and C. W. Huang, “Discovery and Fusion of Salient
Multi-modal Features towards News Story Segmentation”, IS&T/SPIE
Symposium on Electronic Imaging: Science and Technology SPIE Storage
and Retrieval of Image/Video Database, pp. 244-258, 2004.

J. Huang, S. R. Kumar, M. Mitra, W. J. Zhu, and R. Zabih., "Image Indexing
Using Color Correlogram", Proceedings of Computer Vision and Pattern
Recognition, pp. 762-768, 1997.

X. Huang, G. Wei, and V. Petrushin, “Shot Boundary Detection and High-
Level Features Extraction for TRECVID 2003”, Proceedings of TRECVID
2003, Gaithersburg, MD, November 2003. Available at <http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html>.

- 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, pp. 329-338, 1993.
- A. K. Jain and F. Farrokhnia, "Unsupervised Texture Segmentation Using Gabor Filters", Pattern Recognition, 24: pp.1167-1186, 1991.
- 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
- R. Jain, R. Kasturi and B. G. Schunck, "Machine Vision", published by the MIT Press and McGraw-Hill 1995.
- J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic Image Annotation and Retrieval Using Cross-media Relevance Models", Proceedings of the 26th Annual International ACM SIGIR Conference, pp. 119-126, 2003.
- D. Jurafsky and J. H. Martin, "Speech and Language Processing", published by Prentice-Hall Inc 2000.
- J. R. Kender, C. Y. Lin, M. Naphade, A. P. Natsev, J. R. Smith, J. Tešić, G. Wu, R. Yan, D. Zhang, J. O. Argillander, M. Franz, G. Iyengar, A. Amir, and M. Berg, "IBM Research TRECVID 2004 Video Retrieval System", Proceedings of (VIDEO) TREC 2004, Gaithersburg, MD, November 2004. Available at: <http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html>.

- C. K. Koh and T. S. Chua, "Detection and Segmentation of Commercials in News Video". Technical Report, School of Computing, National University of Singapore 2000.
- 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.
- Y. Li "Multi-resolution Analysis on Text Segmentation", Master Thesis, National University of Singapore 2001.
- C. Y. Lin, "Robust Automated Topic Identification", Ph.D. Thesis, University of Southern California 1997.
- Y. Lin, "TMRA-Temporal Multi-resolution Analysis on Video Segmentation", Master thesis, 2000, National University of Singapore.
- C. Y. Lin, B. Tseng, and 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>.
- P. Lyman and H. Varian "How Much Information", available at: <http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/>.
- W. Y. Ma and B. S. Manjunath, "A Comparison of Wavelet Transform Features for Texture Image Annotation", Proceedings of International Conference on Image Processing, pp. 2256-2259, 1995.

- Y. Marchenko, T. S. Chua, and R. Jain, "Transductive Inference Using Multiple Experts for Brushwork Annotation in Paintings Domain", Proceedings of the 14th ACM International Conference on Multimedia, pp. 157–160, 2006.
- J. Naisbitt, "Megatrends: Ten New Directions Transforming Our Lives", Warner Books, 1982.
- J. Naisbitt and P. Aburdene, "Megatrends 2000: The Next Ten Years Major Changes in your Life and World", Sidgwick & Jackson, 1990.
- Y. Nakajima, .D. Yamguchi, H. Kato, H Yanagihara, and Y. Hatori, "Automatic Anchorperson Detection from an MPEG Coded TV Program", Proceedings of International Conference on Consumer Electronics, pp. 122–123, 2002.
- M. Naphade, I. Kozintsev and T. Huang, "A Factor Graph Framework for Semantic Video Indexing", IEEE Transactions on Circuits and Systems for Video Technology, pp. 40-52, 2002.
- M. R. Naphade and J. R. Smith, "On the Detection of Semantic Concepts at TRECVID", Proceedings of the 12th ACM International Conference on Multimedia, pp. 660-667, 2004.
- P. P. Ohanian and R. C. Dubes, "Performance Evaluation for Four Classes of Texture Features", Pattern Recognition, 25(2), pp. 819-833, 1992.

- C. D. Paice, “Constructing Literature Abstracts by Computer: Techniques and Prospects”, *Information Processing and Management*, 26 (1) pp. 171-186, 1990.
- T. V. Pham and M. Worring, “Face Detection Methods: A Critical Evaluation”, Technical Report 2000-11, Intelligent Sensory Information Systems, University of Amsterdam, 2000.
- D. Pierce and C. Cardie, “Limitations of Co-Training for Natural Language Learning from Large Datasets”, *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP-2001)*, Association for Computational Linguistics Research, pp. 1-10. 2001.
- G. J. Qi, X. S. Hua, Y. Song, J. H. Tang, and H. J. Zhang, “Transductive Inference with Hierarchical Clustering for Video Annotation”, *International Conference on Multimedia and Expo*, pp. 643 – 646 2007.
- G. J. Qi, X. S. Hua, Y. Rui, J. H. Tang, T. Mei, and H. J. Zhang, “Correlative Multi-Label Video Annotation”, *Proceedings of ACM International Conference on Multimedia*, pp. 17–26, 2007.
- 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 pp. 3-13, 2005.

- N. C. Rowe, "Inferring Depictions in Natural Language Captions for Efficient Access to Picture Data", *Information Process & Management* vol 30, No 3, pp. 379-388, 1994.
- H. A. Rowley, S. Baluja, and T. Kanade, "Neural Network-based Face Detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 20, Issue 1, pp. 23–38, 1998.
- C. Sable, K. McKeown, and K. W. Church, "NLP Found Helpful (at least for One Text Categorization Task)", *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, vol 10, pp. 172 – 179, 2002.
- G. Salton and M. J. McGill, "Introduction to Modern Information Retrieval", McGraw-Hill, 1983.
- S. Satoh and T. Kanade, "'Name-It: Association of Face and Name in Video", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 368-373, 1997.
- R. Sehetini, "Multicolored Object Recognition and Location", *Pattern Recognition Letters*, vol 15, pp. 1089-1097, 1994.
- T. Shibata and 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.

- M. Slaney, D. Ponceleon, and J. Kaufman, "Multimedia Edges: Finding Hierarchy in All Dimensions", Proceedings of the 9th International Conference on Multimedia, pp. 29-40, 2001.
- C. G. M. Snoek, D. C. Koelma, J. van Rest, N. Schipper, F. J. Seinstra, A. Thean, and M. Worring, "The MediaMill TRECVID 2004 Semantic Video Search Engine", Proceedings of the 2nd TRECVID Workshop, Gaithersburg, USA, 2004. Available at: http://staff.science.uva.nl/~cgmsnoek/pub/UvA-MM_TRECVID2004.pdf.
- C. G. M. Snoek, J. C. Van Gemert, Th. Gevers, B. Huurnink, D. C. Koelma, M. Van Liempt, O. De Rooij, F. J. Seinstra, A. W. M. Smeulders, A. H. C. Thean, C. J. Veenman, and M. Worring, "The MediaMill TRECVID 2006 Semantic Video Search Engine", Proceedings of TRECVID2006. Available at: <http://www-nlpir.nist.gov/projects/tvpubs/>.
- C. G. M. Snoek, M. Worring, J. C. V. Gemert, J. Geusebroek, and A. W. M. Smeulders, "The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia", Proceedings of the 14th ACM International Conference on Multimedia, pp. 421–430, 2006.
- F. Souvannavong, B. Merialdo, and B. Huet, "Eurecom at Video-TREC 2004: Feature Extraction Task", Proceedings of TRECVID 2004, Gaithersburg, MD, November 2004. Available at: <http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html>.

- M. A. Stricker and M. Orengo, "Similarity of Color Images", Proceedings on Storage and Retrieval for Image and Video Databases (SPIE) pp. 381-392, 1995.
- Y. F. Tan, E. Elmacioglu, M. Y. Kan and D. W. Lee, "Efficient Web-Based Linkage of Short to Long Forms", International Workshop on the Web and Databases (WebDB), Vancouver, Canada, June 2008.
- 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.
- TRECVID (2002-2007): "Online Proceedings of the TRECVID Workshops", available at: <http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html>.
- M. Tuceryan and A. K. Jain, "Texture Segmentation Using Voronoi Polygons", IEEE Transaction on Pattern analysis and Machines Intelligence, 12, pp. 211-216, 1990.
- M. Tuceryan and A. K. Jain, "Texture analysis", The Handbook of Pattern Recognition and Computer Vision (2nd edition), pp. 207-248, published by World Science Publishing Co. 1998. Available at: <http://www.cs.iupui.edu/~tuceryan/research/ComputerVision/texture-review.pdf>.
- V. N. Vapnik, "Statistical Learning Theory", Wiley Interscience New York. pp. 120-200, 1998.

- J. Z. Wang and J. Li, "Learning-Based Linguistic Indexing of Pictures with 2-D Multi-resolution Hidden Markov Models", Proceedings of the 10th International Conference on Multimedia, pp. 436-445, 2002.
- K. W. Wilson and A. Divakaran, "Broadcast Video Content Segmentation by Supervised Learning", in Multimedia Content Analysis: Theory and Applications, Ed. Ajay Divakaran, Springer 2008.
- L. Wu, Y. Guo, X. Qiu, Z. Feng, J. Rong, W. Jin, D. Zhou, R. Wang, and M. Jin, "Fudan University at TRECVID 2003", available at: <http://www-nlpir.nist.gov/projects/tvpubs/>.
- L. Xie, L. Kennedy, S. F. Chang, A. Divakaran, H. Sun, and C. Y. Lin, "Discovering Meaningful Multimedia Patterns with Audio-visual Concepts and Associated Text", IEEE International Conference on Image Processing (ICIP 2004), Singapore, vol 4, Issue 24-27 pp. 2383—2386, 2004.
- R. Yan and M. R. Naphade, "Semi-supervised Cross Feature Learning for Semantic Concept Detection in Video", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 657-663, 2005.
- R. Yan, J. Yang, and A. Hauptmann, "Learning Query-Class Dependent Weights for Automatic Video Retrieval", Proceedings of the 12th ACM International Conference on Multimedia, pp. 548–555, 2004.

- 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, pp. 270-278, 2004.
- J. Yang, R. Yan and A. Hauptmann, "Cross-Domain Video Concept Detection Using Adaptive SVMs", In Proceedings of the 15th Annual ACM International Conference on Multimedia, pp. 188-197, 2007.
- M. H. Yang, D. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 1, pp. 34-58, 2002.
- R. E. Yaniv, and L. Gerzon, "Effective Transductive Learning via PAC-Bayesian Model Selection", Technical Report CS-2004-05, IIT, 2004.
- J. Yuan, W. Zheng, Z. Tong, L. Chen, D. Wang, D. Ding, J. Wu, J. Li, F. Lin, B. Zhang, "Tsinghua University at TRECVID 2004: Shot Boundary Detection and High-Level Feature Extraction", Proceedings of TRECVID 2004, Gaithersburg, MD, November 2004. Available at: <http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html>.
- J. Yuan Z. Guo, L. Lv, W. Wan, T. Zhang, D. Wang, X. Liu, C. Liu, S. Zhu, D. Wang, Y. Pang, N. Ding, Y. Liu, J. Wang, X. Zhang, X. Tie, Z. Wang, H. Wang, T. Xiao, Y. Liang, J. Li, F. Lin, and B. Zhang, "THU and ICRC

at TRECVID 2007”, Proceedings of TRECVID 2007. Available at:

<http://www-nlpir.nist.gov/projects/typubs/>.

D. Zhang and S. F. Chang, “Learning Random Attributed Relational Graph for Part-based Object Detection”, ADVENT Technical Report #212-2005-6 Columbia University, May 2005.

X. J. Zhu, “Semi-Supervised Learning Literature Survey”. Available at:

<http://pages.cs.wisc.edu/~jerryzhu/research/ssl/semireview.html>.

Publication List

1. Gang Wang and Tat-Seng Chua, “Capturing Text Semantics for Concept Detection in News Video”, In *Multimedia Content Analysis, Signals and Communication Technology*, Springer Science +Business Media LLC 2009.
2. Gang Wang, Tat-Seng Chua, and Ming Zhao, “Exploring Knowledge of Sub-domain in a Multi-resolution Bootstrapping Framework for Concept Detection in News Video”, *Proceedings of ACM International Conference on Multimedia*, pp. 249-258, 2008.
3. Tat-Seng Chua, Shi-Yong Neo, Hai-Kiat Goh, Ming Zhao, Yang Xiao, and Gang Wang, “TRECVID 2005 by NUS PRIS”, In *TRECVID 2005*.
4. Tat-Seng Chua, Shi-Yong Neo, Keya Li, Gang Wang, Rui Shi, Ming Zhao, Huaxin Xu, “TRECVID 2004 Search and Feature Extraction Task by NUS PRIS”, In *TRECVID 2004*.