An Information-Theoretic Framework towards Large-Scale Video Structuring, Threading, and Retrieval

Winston H. Hsu

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Graduate School of Arts and Sciences

Columbia University

2007

ABSTRACT

An Information-Theoretic Framework towards Large-Scale Video Structuring, Threading, and Retrieval

Winston H. Hsu

Video and image retrieval has been an active and challenging research area due to the explosive growth of online video data, personal video recordings, digital photos, and broadcast news videos. In order to effectively manage and use such enormous multimedia resources, users need to be able to access, search, or browse video content at the semantic level. Current solutions primarily rely on text features and do not utilize rich multimodal cues. Works exploring multimodal features often use manually selected features and/or ad hoc models, thus lacking scalability to general applications. To fully exploit the potential of integrating multimodal features and ensure generality of solutions, this thesis presents a novel, rigorous framework and new statistical methods for video structuring, threading, and search in large-scale video databases.

We focus on investigation of several fundamental problems for video indexing and retrieval: (1) How to select and fuse a large number of heterogeneous multimodal features from image, speech, audio, and text? (2) How to automatically discover and model mid-level features for multimedia content? (3) How to model similarity between multimodal documents such as news videos or multimedia web documents? (4) How to exploit unsupervised methods in video search to boost performance in an automatic fashion?

To address such challenging problems, our main contributions include the following: First, we extend the Maximum Entropy model to fuse diverse perceptual

features from multiple levels and modalities and demonstrate significant performance improvement in broadcast news video segmentation. Secondly, we propose an information-theoretic approach to automatically construct mid-level representations. It is the first work to remove the dependency on the manual and labor-intensive processes in developing mid-level feature representations from low-level features. Thirdly, we introduce new multimodal representations based on visual duplicates, cue word clusters, high-level concepts, etc. to compute similarity between the multimedia documents. Using such new similarity metrics, we demonstrate significant gain in multi-lingual cross-domain topic tracking. Lastly, to improve the automatic image and video search performance, we propose two new methods for reranking the initial video search results based on text keywords only. In the image/video level, we apply the information bottleneck principle to discover the image clusters in the initial search results, and then rerank the images based on cluster-level relevance scores and the occurrence frequency of images. Such method is efficient and generic, applicable to reranking of any initial search results using other search approaches, such as content-based image search or semantic concept-based search. In the multimedia document level, building on the multimodal document similarities, we propose a random walk framework for reranking the initial text-based video search results. Significant performance improvement is demonstrated in comparison with text-based reranking methods. In addition, we have studied application and optimal parameter settings of the power method in solving the multi-modal random walk problems. All of our experiments are conducted using the large-scale diverse video data such as the TRECVID benchmark data set, which includes more than 160 hours of broadcast videos from multiple international channels.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank those who have made my thesis possible and enriched my life at Columbia.

First and foremost, I would like to thank my advisor, Prof. Shih-Fu Chang, who has been a great source of support and guidance throughout my research at Columbia. Prof. Chang insists on rigorous thinking in research and consistently inspires novel perspectives in challenging problems. He always points to critical insights during the discussions, guides me through the perplexing setbacks, and helps me discover the fun of devising state of the art solutions. In addition, he gave me great freedom as a PhD student and created a lively and accommodating atmosphere. I do feel extremely grateful and respectful to him.

I would like to thank Dan Ellis for being a wonderful mentor. He is keen in research and easygoing in life. Attending his courses and discussions is always fun and exciting. Most of all, he is always willing to give me a hand as I knock on his door.

I would like to thank my thesis committee members, Prof. Dragomir R. Radev and Prof. Dimitris Anastassiou for their kind work during my defense and the insightful comments they provided on my thesis.

It is a great luxury to work in the Digital Video and Multimedia group. I would like to thank the members for making my stay at Columbia fascinating and memorable – Lyndon Kennedy, Chih-Wei Huang, Lexing Xie, Dongqing Zhang, Shahram Ebadollahi, Eric Zavesky, Tian-Tsong Ng, Jessie Hsu, Akira Yanagawa, Wei Jiang, and Jun Wang. They always bring laughter and great resources to research discussions. During my early days in the group, the senior group members, Hari Sundaram, Ana Belen Benitez, and Alejandro Jaimes, gave me great research

advice. They smoothed my transition from industry to academia. I would like to express my appreciation to them as well.

Thanks to John Smith for giving me the opportunity to work with researchers at IBM T.J. Watson in many exciting projects. I would also like to thank other members at the research center for kindly providing useful discussions and joyful affiliations: Ching-Yung Lin, Milind Naphade, Apostol Natsev, Jelena Tesic, Giridharan Iyengar, Bell Tseng, and Martin Franz.

Working with Prof. Chang, I am lucky to interact with the visiting scholars who have provided me great sources of fun and research inspirations. I also like to thank Nevenka Dimitrova, Masaki Miura, Masami Mizutani, Ryoma Oami, and Jason Fang.

The most important appreciation is to my sweet family, Sunny, Cynthia, and Ethan – for without their unwavering support and warmest love, none of the research would have been possible.

Chapter 1

Introduction

This thesis presents a computational and statistical framework for video structuring, threading, and search in a large-scale video database.

1.1 Background

With recent advances in communications, computers, and storage capacities, digital media is becoming an increasingly important information format in a variety of domains such as personal video recordings, TV programs, instant messaging, video blogs, broadcast news videos, etc. For example, "Information flows through electronic channels – telephone, radio, TV, and the Internet – contained almost 18 exabytes of new information in 2002, three and a half times more than is recorded in storage media," Lyman *et al.* reported in [4]. "There are approximately 123 million hours of broadcasting videos and the World Wide Web (WWW) contains about 170 terabytes of digital media information."

To deal with such enormous amounts of information, several promising applications have emerged. For example, the search engine Google [5], though primarily using the text modality, has shown interesting progress in exploiting information

from web images by utilizing their surrounding text or captions. Online news web sites such as Yahoo! news [6], aggregate news (text) stories from different media providers and summarize news topics. Flickr [7] and YouTube [8] enable consumer photo and video sharing and provide preliminary search capability through user-annotated keywords (or tags).

Motivated by strong application needs and theoretical interests, image/video indexing and retrieval have become an active research topic in the multimedia research community. Several projects, such as Columbia CuVid Search System [9], Informedia Project [10], IBM Marvel [11], and other activities based on TRECVID video benchmarks [3, 12, 13], have shown promising results in image and video indexing and retrieval. There are numerous proposed approaches derived from such multimedia retrieval systems. For example, in certain applications (e.g., broadcast news video story segmentation and low-level image matching), heuristic rules are often used to determine what features to use for representing visual content and computing similarity. Such assumptions usually limit the scalability and performance over cross-domain and large-scale data.

In other applications (e.g, audio-visual concept classification, video retrieval, etc.), a new interesting direction is to introduce "mid-level" features that can help bridge the semantic gap between low-level features and semantic targets. Examples of such mid-level features include location (indoor, waterfront), people (one person vs. group), objects (car, building), production syntax (anchor), etc. Promising performance from such mid-level representations have been demonstrated. However, most of these mid-level features are manually selected, defined, annotated, and, thus require costly human intervention and become infeasible for large-scale applications.

Moreover, in video search, the additional use of multiple modalities such as image features, audio, face detection, and high-level concept detection has been shown to

improve upon the text-based video search systems [9, 12, 14, 15]. Much of the improvement is achieved by using multiple query example images, applying specific concept detectors, or incorporating highly-tuned retrieval models for specific types of queries. However, none of these approaches are practical for large-scale applications. It will be difficult for users to acquire example images for example-based queries. Retrieval by matching semantic concepts, though promising, strongly depends on the availability of robust classifiers and requires training data. Likewise, it will be difficult for the developers of the system to develop highly-tuned models for every class of query and apply the system to new domains. It is clear, then, that we need to develop automatic methods that systematically discover feasible solutions using information from available multimodal cues in the data set without depending too much on system users or developers.

Beyond the multimodal representations, there are a few important observations regarding the dissemination process of multimedia information. As a critical event occurs, it will spread right away from the original source to other locations and media channels in different forms. For example, Fig. 1.1 illustrates the spread of two news events across domains and news sources. The left one is "Philippine President Arroyo ceases coup attempt" and the right one is "Muslim sectarian riot in Iraq." Both are widely covered in broadcast news videos, newspapers, online news, and even blogs of different languages and countries. Among these sources, we can easily find recurrent visual duplicates (e.g., Arroyo's pictures and demonstration scenes) or common classes of content in the videos or images (e.g., desert scenes, Muslim temples, military persons, in the Muslim topic.) Through these recurrent visual elements, even a person who does not understand the language of a particular story can understand what topics it addresses. It is part of our objectives in this thesis to utilize these recurrent visual patterns to improve the search and linking

Figure 1.1: As a critical event occurs (e.g., tsunami or hurricanes), bursts of news stories emerge from diverse sources across domains (channels, the x-axis) and over time (y-axis) with the recurrence of visual duplicates and semantic high-level concepts.

functions of video databases.

In this thesis, we focus on investigation of several fundamental problems: (1) How to select and fuse a large number of heterogeneous multimodal features from image, speech, audio, and text? (2) How to automatically discover and model mid-level features for multimedia content? (3) How to model similarity between multi-modal documents such as news videos or multimedia web documents? (4) How to exploit unsupervised or pseudo-supervised situations in video search to boost performance in an automatic fashion?

We approach these problems by: (1) applying statistical feature fusion and selection from heterogeneous content modalities; (2) proposing novel information-theoretic approaches to automatically construct semantic features from multi-modal

sources; (3) discovering semantic threads across video sources; (4) improving video retrieval performance using syntax structures, contextual threads, and recurrent visual patterns, etc. The proposed techniques are unified under an information-theoretic framework by extending theories such as Maximum Entropy, the Information Bottleneck principle, and random walk theory, etc. Optimization in time and space complexity is also addressed in several cases. We evaluate and validate the proposed approaches using large-scale video benchmark data such as TRECVID [3].

The proposed techniques are not only applicable to video data but also other media types such as digital books, MP3 music, medical information, digital photos, online lectures, voice recordings, etc. An interesting example is the one addressed in the MyLifeBits Project [16] which aims at manipulating and storing every piece of information encountered in the everyday life of a person.

1.2 Scope of Work

We use the diagram shown in Figure 1.2 to define the scope of research and target applications in this thesis. To support the high-level applications such as video search and topic threading, we need a basis for the semantic units of the domain (i.e. stories in broadcast news videos or web pages in WWW). Feature selection plays a central role in finding the adequate representations of semantic units as well as effective measures for similarity between multimedia documents. Therefore a central theme of our research has been around innovative and effective ways of feature selection at both low- and mid-levels. We then test our theories and algorithms in solving problems in broadcast news video story segmentation, video search reranking, and video topic threading.

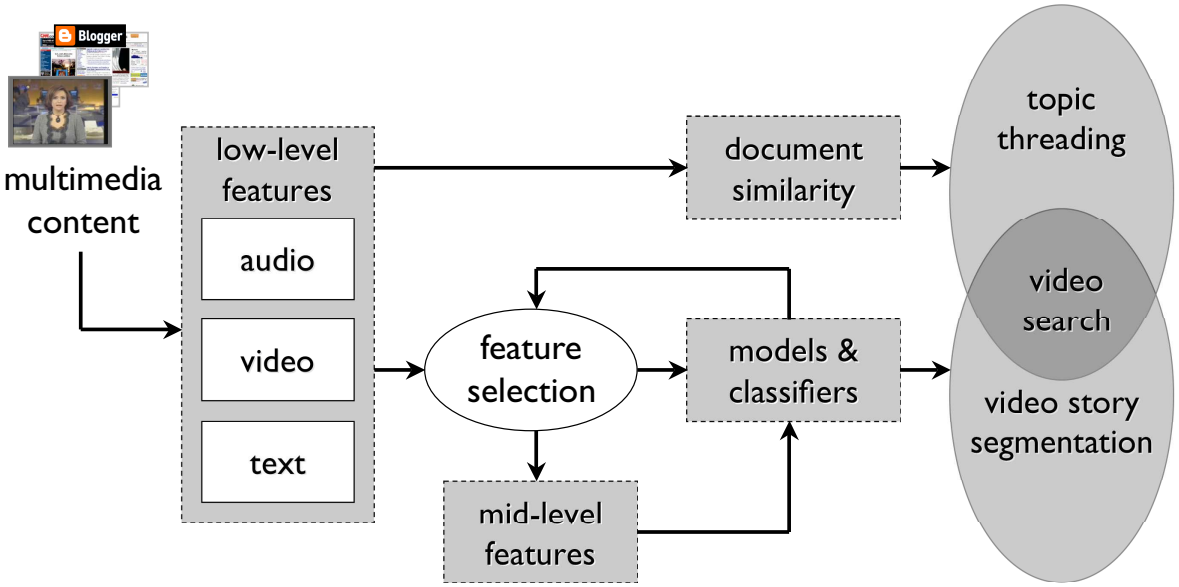


Figure 1.2: The thesis scope – an automatic framework for video indexing and retrieval on large-scale video databases – includes the end applications such as topic threading, video search (proposed in two novel perspectives), and broadcast news video story segmentation. Based on multimodal features (e.g., audio, video, and text), statistical and information-theoretical approaches are proposed for multimodal feature selection and fusion, mid-level feature representations, cross-domain document similarities, and classifications, etc.

1.2.1 Related Work and Motivation

1.2.1.1 Semantic Structure Segmentation

News story segmentation is an important technology for information exploitation in news video, which is a major information source in the new era. For example, the authors of [17] use the story as a basic unit to associate audio-visual concepts, obtained through unsupervised discovery, with text annotations. In [18], a spectral clustering algorithm for multi-source clustering problems based on segmented stories is presented. The authors of [19] confirmed the benefits of video search using the video story unit compared with other ad-hoc approaches.

There have been several works addressing story segmentation in broadcast news

videos. Authors of [20] adopted ad-hoc rules on combining image, audio, transcripts, and closed captions to locate story boundaries. However, for international news programs, closed captions, and accurate speech recognizers are usually unavailable. Besides, the production rules vary for different channels, countries, or time periods. Qi *et al.* [21] identified the story boundaries by a clustering-based algorithm that detects anchorpersons by performing an AND operation on visual and speech anchor segments. The image clustering method may not achieve required accuracy due to camera zooming effects in studio settings, multiple anchors, and superimposed captions and graphics. A similar method [22] was proposed to discover visual and acoustic cues to identify anchorpersons. Their face region detection process is applied to the key frame of each shot only and is sensitive to shot detection errors. The authors of [23] incorporated 4-state (Story Start, Story End, Advertisement, and Other) Hidden Markov Models (HMM) with discrete state-dependent emission probabilities to detect the story boundaries. Besides, several methods are based on the assumption that each story starts with an anchor segment. These heuristic algorithms lack generality in handling diverse video sources with different features and production rules.

In another HMM approach [24], Chaisorn *et al.* explored mid-level concept detection from each shot and models the temporal transitions among shots. The strategy of this approach is to categorize constituent shots within the news videos according to the production rules and observations; in TRECVID 2003, they determined 17 shot types specific to CNN and ABC news, namely Anchor, Two-Anchor, Top story, Lead-in/out, Sport logo, Play (of the day), Health, etc. A supervised decision tree classifier is applied to categorize each shot to one of the predefined 17 types. A 4-state HMM approach is then applied to detect the story boundaries based on the shot sequence. This work showed very promising results in TRECVID 2003.

Though promising, the mid-level categories used in [24] are manually chosen relying on expert knowledge of the application domain. Once the mid-level categories are chosen, extensive manual efforts are needed to annotate training data for learning the detector of each mid-level feature. A further goal of ours is to automate the selection process for the mid-level features given defined semantic class labels. Provided the collection of data, consisting of low-level features and associated semantic labels, we want to discover the mid-level features automatically. There is still a need for labeling the semantic label of each data sample, but the large cost associated with annotation of the training corpus for each manually chosen mid-level feature is no longer necessary. In addition, the dimensionality of the mid-level features will be much lower than that of the low-level features.

1.2.1.2 Video Threading

Due to the explosion of Internet bandwidth and broadcast channels, video streams are easily accessible in many forms such as news video broadcasts, blogs, and podcasts. As news of a critical event breaks, bursts of news stories of the same topic emerge either from professional news or amateur videos. Topic threading is an essential task for organizing video content from distributed sources into coherent topics for further manipulations such as browsing or search.

Current topic threading approaches primarily exploit text information from speech transcripts, closed captions, or web documents. A major research effort for topic threading based on text has been conducted under the NIST Topic Detection and Tracking (TDT) benchmark [25], which includes three tasks: (1) story link detection, determining whether two stories discuss the same topic; (2) topic tracking, associating incoming stories with topics that are known to the system; (3) topic detection, detecting and tracking topics that are not previously known to the

system.

In the thesis, we mainly focus on topic tracking across international broadcast news videos. One representative work of the text-based approach can be found in [26], where the authors represent documents as vectors of words, weighted by term-frequency inverse-document-frequency (TF-IDF). The cosine angle is used for measuring document-pair similarity. A modified k nearest neighbor (kNN) approach is then used for classification.

The use of multimodal information such as visual duplicates [1] or semantic visual concepts [3], though not explored before, are very helpful. There are usually recurrent visual patterns in video stories across sources that can help topic threading. For example, Fig. 6.1 has three example stories from Chinese, Arabic, and English news sources, which cover the same topic. The stories from different channels share a few near-duplicates such as those showing George Bush and Tony Blair, the press conference location, and the audience. Such duplicates, confirmed by our analysis later, are proven to be effective for news threading across languages (cf. Section 6.2.3).

Some recent works study new techniques using multimodal information for story topic tracking. Xie *et al.* [27] applied Hierarchical HMM models over the low-level audio-visual features to discover spatio-temporal patterns, latent semantic analysis to find text clusters, and then fused these multimodal tokens to discover potential story topics. In [28], the authors studied the correlation between manually annotated visual concepts (e.g., sites, people, and objects) and topic annotations, and used graph cut techniques in story clustering. In [29], the authors addressed the problem of linking news stories across two English channels on the same day, using global affine matching of key-frames as visual similarity. In all of these prior works, neither visual duplicates nor automatically detected visual concepts were used. In

addition, performance comparisons with text-based approaches were not clear.

1.2.1.3 Image and Video Search

Video and image retrieval has been an active and challenging research area thanks to the continuing growth of online video data, personal video recordings, digital photos, and 24-hour broadcast news. In order to successfully manage and use such enormous multimedia resources, users need to be able to conduct semantic searches over the multimodal corpora either by issuing text keyword queries or providing example video clips and images (or some combination of the two).

Current successful semantic video search approaches usually build upon text search against text associated with the video content, such as speech transcripts, closed captions, and video OCR text. The additional use of multiple modalities such as image features, audio, face detection, and high-level concept detection has been shown to improve upon the text-based video search systems [9, 12, 14, 15]. Much of the improvement is achieved by using multiple query example images, applying specific concept detectors, or incorporating highly-tuned retrieval models for specific types of queries. However, none of these approaches are practical for large-scale applications. It will be difficult for users to acquire example images for example-based queries.

Practically, it has been shown that users prefer "search by text" rather than "search by examples" [30]. Moreover, Battelle mentioned that "users are incredibly lazy" [31]. He even pointed out, "We [users] type in a few words at most, then expect the engine to bring back the perfect results. More than 95 percent of us never use the advanced search features most engines include." Hence, a search-by-text method automatically utilizing multimodal cues is essential for leveraging user satisfactions in image and video search applications.

Retrieval by matching semantic concepts, though promising, strongly depends on the availability of robust classifiers and requires training data. Similarly, it will be difficult for the developers of the system to develop highly-tuned models for every class of query and apply the system to new domains. It is clear, then, that we need to develop automatic methods that systematically discover feasible solutions using information from available multimodal cues in the data set without depending too much on system users or developers.

1.2.2 Recurrent Patterns

In the thesis, we also aim at automating the mid-level feature discovery in certain video applications in order to avoid costly human labor for labeling. Besides, in order to address the problems of example-based video search approaches and avoid the use of specialized models, we have developed a generic method that explores the use of recurrent visual patterns inherent in many contexts, such as the returned image sets from initial text queries, or video stories from distributed channels.

As illustrated in Fig. 1.1, across diverse video sources or domains, there are often recurrent patterns with close similarities either in visual or audio modalities. Such recurrent patterns may be used as (soft) "contextual links" between videos to improve the initial search results based on text retrieval. For example, in Fig. 8.1, an initial text query retrieves video stories with the keywords "tony blair;" however, there are still certain relevant stories not retrieved due to the lack of keyword annotations associated with such videos. The contextual links (e.g., visual duplicates, text, etc.) may be used to link such missing stories to the initial text queries.

Such recurrent images or videos are commonly observed in image search engines (e.g., Yahoo or Google) and photo sharing sites (e.g., Flickr). Interestingly, the

authors of [32] quantitatively analyzed the frequency of such recurrent patterns (in terms of visual duplicates) for cross-language topic tracking – a large percentage of international news videos share common video clips or near duplicates.

1.3 Contributions of the Thesis

The major focus of the thesis is to investigate a principled statistical framework to enhance video indexing and retrieval over large-scale video database of diverse sources and languages. The framework is adaptive and automatic and can be applied to different problems and domains. Experimentally, the framework is shown to achieve significant results in large-scale benchmarks such as TRECVID [3]. The major problems addressed in the thesis include the following:

- **How to select and fuse a large number of heterogeneous multimodal features from image, speech, audio, and text?**: Multimodal fusion and feature selection has been of great interest to the research community. We investigate issues pertaining to multimodal fusion and selection in broadcast news video story segmentation, which has been shown to be essential for video indexing, summarization, and intelligence exploitation. In this context, we applied and extended the Maximum Entropy statistical model to systematically select and effectively fuse diverse features from multiple levels and modalities, including visual, audio, and text, in international broadcast news videos. We have included various features such as motion, face, music/speech discrimination, speech rapidity, high-level text segmentation information, prosody, etc., and some novel features such as syllable cue terms (in Mandarin) and significant pauses for international broadcast news video segmentation. The statistical fusion model is used to automatically discover

relevant features that are most useful for the detection of story boundaries. We also introduced a novel feature wrapper to address heterogenous features – usually asynchronous, variant-timescale, and either discrete or continuous. We demonstrated excellent performance (F1 measures 0.76 in ABC news, 0.73 in CNN news, and 0.90 in Mandarin news). Our method was first applied on broadcast news video segmentation in TRECVID 2003 and achieved one of the best performances among the submissions.

- **How to discover and model mid-level features for multimedia content?**: Recent research in video analysis has shown a promising direction, in which mid-level features (e.g., people, locations, objects) are abstracted from low-level features (e.g., color, texture, motion, etc.) and then used for discriminative classification of semantic labels. However, in most systems, such mid-level features are selected manually. Instead, we propose an information-theoretic framework, to automatically discover adequate mid-level features – semantic-consistent clusters. The problem is posed as mutual information maximization, through which optimal clusters are discovered to preserve the highest information about the semantic labels. We extend the Information Bottleneck (IB) framework to high-dimensional continuous features to discover these semantic-consistent clusters as a new mid-level representation. The biggest advantage of the proposed approach is to remove the dependence on the manual process of choosing mid-level representations and the huge labor cost involved in annotating the training corpus for training classifiers for each mid-level feature. The proposed IB framework is general and effective, with the potential to solve many other problems in semantic video analysis.

  We apply these automatic semantically-consistent clusters as new mid-level

representations for the video classification problem in chapter 4. The proposed approach achieves promising performance gain over representations derived from conventional clustering techniques and even the mid-level features selected manually. It is the first work to remove the dependence on the manual process in choosing the mid-level visual features and the huge labor cost involved in annotating the training corpus. We have reported excellent performance in testing this method to story segmentation in TRECVID 2004 and 2005. It was the only method shown effective for processing the diverse videos from multiple international channels in 2005.

We also show that the discovered mid-level features can be used to significantly improve the video search quality. The IB-based clusters provide a novel way for denoising the raw search scores from text retrieval, and constraining the visual reranking step. The method improves upon the text-search baseline by up to 23%, in average performance, and is comparable with other sophisticated models, which use highly tuned models for different query types..

- **How to model similarity between multi-modal documents such as news videos or multimedia web documents?**: Adequate models and computational methods for video similarity at the story level are important for video topic tracking and video story search. Current solutions primarily rely on text features only and therefore encounter difficulty when text is noisy or unavailable. In this thesis, we extend the text modality and propose new representations and similarity measures for news videos based on low-level visual features, visual near-duplicates, and high-level semantic concepts automatically detected from videos.

  Contrary to similarity measures defined over fixed windows or shots, here

we consider feature representations at the semantic level – a (video) story or document, which usually has variable lengths. We need to address the many-to-many pair-wise multimodal similarity measures in such story-level feature representations. Fig. 6.5 illustrates the notion of similarity between a broadcast news video and three web news articles of different languages covering the same topic. The video duplicates provide a very strong similarity link between cross-domain documents.

- **How to exploit unsupervised or pseudo-supervised situations in video search to boost performance in an automatic fashion?**: Pseudo-relevance feedback (PRF) [14, 33, 34], is a promising technique that has been shown to improve upon basic text search results in both text and video retrieval. PRF is initially introduced in [35], where the top-ranking documents are used to rerank the retrieved documents assuming that a significant fraction of top-ranked documents will be relevant. This is in contrast to relevance feedback where users explicitly provide feedback by labeling the top results as positive or negative.

  The underlying assumption for PRF is that there are some hidden context or topics in the relevant documents. PRF approaches are used to derive other keywords relevant to the (hidden) topics associated with the target documents. The approach fails for video search since the hidden topics in the initial video search results are usually much more ambiguous and difficult to find [33]. Thus, most query expansion experiments based on PRF are restricted to the text modality only.

  Instead of inferring the hidden topics, we rely on the multimodal similarities between videos. Having observed recurrent visual patterns in image search

engines, photo sharing sites, and video databases, we propose to leverage this multimodal similarity between semantic units to improve the initial text query results. The approach is formulated as a random walk problem along a context graph where the graph nodes are video documents, connected by the edges weighted with pair-wise multimodal contextual similarities (including visual duplicates, high-level concepts, text tokens, etc.) discussed in chapter 6. The stationary probability of the random walk computed from the multimodal context graph is used to represent the refined ranking scores of the videos. The proposed random walk is biased towards stories with higher initial text search scores – a principled way to consider both initial text search results and their implicit contextual relationships.

We have shown experimentally that the proposed approach improves retrieval performance by an average gain of up to 32% relative to the baseline text search method in terms of story-level MAP. Furthermore, in the people-related queries, which usually have recurrent coverage across news sources, we can have up to 40% relative improvement in story-level MAP. Through parameter sensitivity tests, we have also found that the optimal text/visual weight ratio for reranking baseline text search results is .15:.85.

1.4 Organization of the Thesis

The rest of the thesis is organized as the following. In chapter 2, we extend the Maximum Entropy statistical model to fuse diverse perceptual features from multiple levels and modalities and conduct experiments of international broadcast news video segmentation. In chapter 3, to remove the dependency on the manual and labor-intensive processes in choosing, annotating, and training the mid-level features,

we propose an information-theoretic approach to automatically construct mid-level clusters and derive new semantic representations over the discovered clusters. In chapter 4, such semantic-consistent clusters are applied to the international broadcast news video story segmentation task. In chapter 5, we explore a new method for video search reranking. Our approach explores the automatically discovered clusters as new bases to de-noise initial text search scores and constrain the computation model for the recurrence frequency. The new method is generic, efficient, and complementary with alternative methods using special models.

In chapter 6, we introduce story-level multimodal similarities such as visual duplicates, cue word clusters, high-level concepts, etc., for novel feature representations of the semantic units. These story-level similarities are first used for cross-source topic tracking in chapter 7, where we found that visual duplicates are competitive features, compared with text or visual concepts. Building upon story-level similarities, we propose a random walk framework for reranking the text-based video search results in chapter 8. Significant performance improvement is demonstrated, in comparison with text-based methods. We then summarize the thesis and present conclusions and future work in chapter 9.

Chapter 2

Fusion and Selection of Multimodal Features

Multimodal fusion and feature selection has been great interest to the research community. We investigate issues pertaining to multimodal fusion and selection in broadcast news video story segmentation, which has been shown essential for video indexing, summarization, and intelligence exploitation. In this chapter, we applied and extended the Maximum Entropy statistical model to systematically induce and effectively fuse diverse features from multiple levels and modalities, including visual, audio, and text, in international broadcast news videos. We have included various features such as motion, face, music/speech discrimination, speech rapidity, high-level text segmentation information, prosody, etc., and some novel features such as syllable cue terms (in Mandarin) and significant pauses for international broadcast news video segmentation. The statistical fusion model is used to automatically discover relevant features contributing to the detection of story boundaries. We also introduced a novel feature wrapper to address heterogenous features – usually asynchronous, variant-timescale, and either discrete or continuous. We demonstrated encouraging performance (F1 measures up to 0.76 in ABC news, 0.73 in CNN news, and 0.90 in Mandarin news), presented how to leverage these multi-level multi-modal features in a probabilistic framework.

2.1 Introduction

News story segmentation is an important underlying technology for information exploitation in news video, which is a major information source in the modern era. There are some projects addressing news video engineering such as the Informedia Project [10], which tries to search, summarize, and visualize linear video in meaningful abstractions and takes the stories as the basic units. Currently, authors of [17] use the story as a basic unit to associate audio-visual concepts, obtained in unsupervised discovery, with text annotations. In [18], the authors present a spectral clustering algorithm for multi-source clustering problems based on segmented stories which possess consistent and concise semantic concepts. Authors of [19] have confirmed the effectiveness of video search in the video story unit comparing with other ad-hoc approaches.

There have been several works addressing story segmentation in broadcast news videos. Authors of [20] adopted ad-hoc rules on combining image, audio, transcripts, and closed captions to locate story boundaries. However, for international news programs, closed captions, and accurate speech recognizers are usually unavailable. Besides, the production rules vary for different channels, countries, or time periods. Qi *et al.* [21] identified the story boundaries by a clustering-based algorithm that detects anchorpersons by performing an AND operation on visual and speech anchor segments. The image clustering method may not achieve required accuracy due to camera zooming effects in studio settings, multiple anchors, and superimposed captions and graphics. A similar method [22] was proposed to discover visual and acoustic cues to identify anchorpersons. Their face region detection process is applied to the key frame of each shot only and is sensitive to shot detection errors. The authors of [23] incorporated 4-state (Story Start, Story End, Advertisement,

and Other) Hidden Markov Models (HMM) with discrete state-dependent emission probabilities to detect the story boundaries. Besides, several methods are based on the assumption that each story starts with an anchor segment. These heuristic algorithms lack generality in handling diverse video sources with different features and production rules.

In another HMM approach [24], Chaisorn *et al.* explored mid-level concept detection from each shot and models the temporal transitions among shots. The strategy of this approach is to categorize constituent shots within the news videos according to the production rules and observations; in TRECVID 2003, they determined 17 shot types specific to CNN and ABC news, namely Anchor, Two-Anchor, Top story, Lead-in/out, Sport logo, Play (of the day), Health, etc. A supervised decision tree classifier is applied to categorize each shot to one of the predefined 17 types. A 4-state HMM approach is then applied to detect the story boundaries based on the shot sequence. This work showed a very promising result in TRECVID 2003.

In this chapter, we take another perspective and argue that there exist consistent statistical characteristics within news videos from each channel, and with adequate learning, a general model with a generic pool of computable and perceptual features can be systematically optimized to construct effective segmentation tools for each news channel. We focus on the multi-modal perceptual features, conveying to the viewers the change of topics or stories, rather than the multi-channel production rules or types. A statistical framework with feature selection and fusion is mandatory to adaptively deploy the system. We gained competitive performance in experiments in both Mandarin and US news video programs under the same encouraging framework.

We had a pilot work in [36], where the Maximum Entropy (ME) approach is initially adopted by fusing dozens of features on hours of Mandarin news. We

further extend that approach by including novel perceptual features, solving multi-modal fusion issues with a novel feature wrapper, and evaluating on 218 half-hour ABC/CNN news programs in [37].

The Maximum Entropy principle [38] tries to estimate a model that satisfies the existent (equality or inequality) constraints and assumes those unknown as a uniform distribution. The solution to the problem is optimal and generally in the exponential form as illustrated in Equation 2.1. Based on this principle, authors of [39] derived the feature induction mechanism and applied it to a French-to-English machine translation problem. In [40], Beeferman *et al.* developed the ME algorithms for topic segmentation in text documents, where a family of weighted, exponential functions of binary features is used to account for the text boundary posterior probability at each sentence boundary. Such features may include the occurrences of certain key terms, such as *Does the word "MR" appear in the next sentence?*, *Does the word "SAID" appear in the previous five sentences but not in the next five sentences?*. The authors also use an induction procedure to automatically find the most salient features.

A news story is defined as a segment of a news broadcast with a coherent news focus which contains at least two independent declarative clauses. Other coherent segments are labeled as non-news. These non-news stories cover a mixture of footage: commercials, lead-ins, and reporter chit-chat. A story can be composed of multiple shots; e.g., an anchorperson introduces a reporter and the story is finished back in the studio-setting. On the other hand, a single shot can contain multiple story boundaries; e.g., an anchorperson switches to the next news topic.

The story boundaries along with the TRECVID benchmark [3] include those of normal news stories as well as boundaries of sports and weather. Figure 2.1 illustrates common types of stories that can be found in broadcast news videos such

Table 2.1: Counts and percentages of different types of video stories (defined in Figure 2.1) from 22 randomly selected CNN videos in TRECVID 2003 data set.

Types	a	b	c	d	e	f	g	h	i	all
Story Bdry. #	244	48	67	114	162	16	22	58	28	759
Percentage (%)	32.0	6.3	8.8	15.0	21.3	2.1	2.9	7.6	3.7	100

as CNN. The proportion of different types in the whole collection is listed in Table 2.1 (row 3: percentage). Note that there are a broad range of story types with significant percentage of data. The most popular types of news stories are type `a` (story starting with a visual anchor), type `e` (a series of stories without visual anchors), and type `d` (sports briefing series) in the descending order. With the combination of these three types accounting for 68.3% of data from 759 stories in 22 (randomly selected) CNN news programs. There are six different types of stories that have percentages of at least 5%.

To assess the baseline performance, we also conduct an experiment by evaluating story boundaries with visual anchor segments only and yield a baseline result, shown in Table 2.3, where boundary detection F1[1] measures in ABC is 0.67 and is 0.51 in CNN with only 0.38 recall and 0.80 precision rates. The definition of evaluation metrics is explained in section 2.5.3. We will present significant performance gain over the baseline by using statistical multi-modal fusion and demonstrate satisfactory performance with F1 measures up to 0.76 in ABC news and 0.73 in CNN news.

The contributions of our work comes out as four folds: (1) We proposed a principled statistical approach that could adaptively and systematically discover and fuse multi-modal features from a generic feature pool and avoid heuristic rules and strict assumptions; (2) we invented some novel features such as syllable cue terms

[1] $F1 = \frac{2 \cdot P \cdot R}{P+R}$, where P and R are precision and recall rates

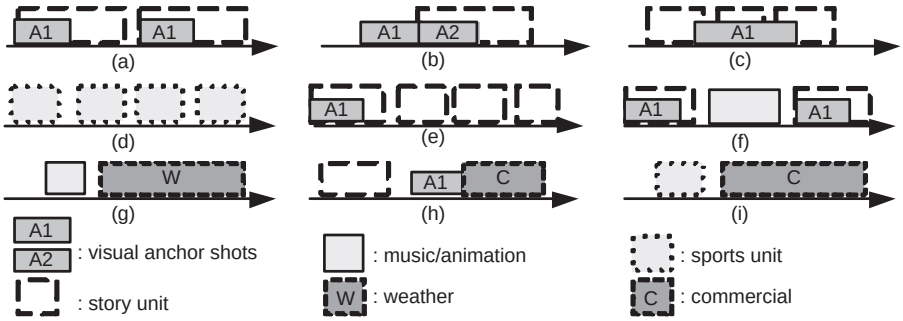


Figure 2.1: Common CNN story types seen in the TRECVID 2003 data set. $A1$ and $A2$ represent segments showing different visual anchor persons; (a) two stories both starting with the same visual anchor; (b) the second story starts with a different visual anchor; (c) multiple stories reported in a single visual anchor shot; (d) a sports section constitutes series of briefings; (e) series of stories that do not start with anchor shots; (f) two stories that are separated by long music or animation representing the station id; (g) weather report; (h) a non-story section consisting of an anchor lead-in followed by a long commercial; (i) a commercial comes right after a sports section. The percentages of theses story types are listed in Table 2.1.

and significant pauses, and deliver significant performance based on multi-modal, heterogenous, asynchronous, and perceptual features; (3) we applied the same framework on different languages, i.e. English and Mandarin, different channels, and still achieved competitive performance; (4) we evaluated the framework on TRECVID 2003 database with 120 hours of news videos and automatically induced salient perceptual features, some of which not only confirm the heuristic assumptions that most work rely on but also list the inadequacy of such assumptions.

The issues regarding multi-modal fusion are discussed in section 2.2. The probabilistic framework and the feature wrapper are addressed in section 2.3. Relevant features are presented in section 2.4. The experiment, evaluation metrics, and discussions are listed in Section 2.5, followed by the summary and discussions in section 2.6.

2.2 Issues with Multi-modal Fusion

There are generally two perspectives on story segmentation – one is boundary-based and the other is segment-based. The former models the characteristics of features at the boundary points as shown in Figure 2.2; the latter models the temporal dynamics within each story. We adopt the first approach in this paper. In such an approach, one basic issue is the determination of candidate points, each of which is tested and classified as either a story boundary or a non-boundary.

2.2.1 Candidate Points

A good candidate set should have a very high recall rate on the reference boundaries and indicate the places where salient and effective features occur. Shot boundaries are usually the candidate points used in most news segmentation projects such as those from [24] and [23]. In audio segmentation, for example, Shriberg *et al.* evaluated the candidate points at pause points [41]; in text segmentation, most work resides in non-speech segments or starts of sentences [40, 42] . After thorough investigations, we found that taking the shot boundaries alone is not complete. We evaluate the candidate completeness by detecting reference boundaries with 5-second fuzzy window (defined in section 2.5.3). Surprisingly, the recall rate for the shot boundaries on ABC/CNN is only 0.91. The reason is that some reference boundaries are not necessarily at (or close to) the shot boundaries. In this work, we take the union of shot boundaries T_{sht} and audio pauses T_{pas} as candidate points but remove duplications within a 2.5-second fuzzy window by simply taking an "OR" operation (Appendix A.) $T_{sht} \oplus_{2.5} T_{pas}$. The union candidates yield 100% recall rate on the training set.

2.2.2 Data Labeling

We adopt a supervised learning process with manually annotated reference boundaries. Since the features are usually asynchronous across modalities and the annotated data is not necessarily aligned well with the ground truth, each candidate point is labeled as "1" (boundary) if there is a reference boundary within the 2.5-second fuzzy window. However, some reference boundaries could not locate corresponding candidates within the fuzzy window. The phenomenon also happens in our ASR text segmentation (Section 2.4.5) and we just insert these reference boundaries as additional candidate points in the training set.

2.3 Probabilistic Framework

News videos from different channels usually have different production rules or dynamics. We choose to construct a model that adapts to each different channel. When dealing with videos from unknown sources, identification of the source channel can be done through logo detection or calculating model likelihood (fitness) with individual statistical station models.

We propose to model the diverse boundary transitions and content dynamics by using statistical frameworks. The assumption is that there exist consistent statistical characteristics within news video from each channel, and with adequate learning, a general model with a generic pool of computable features can be systematically optimized to construct effective segmentation tools for each news channel.

2.3.1 Maximum Entropy Model

The ME model [36, 40] constructs an exponential log-linear function that fuses multiple binary features to approximate the posterior probability of an event (i.e.

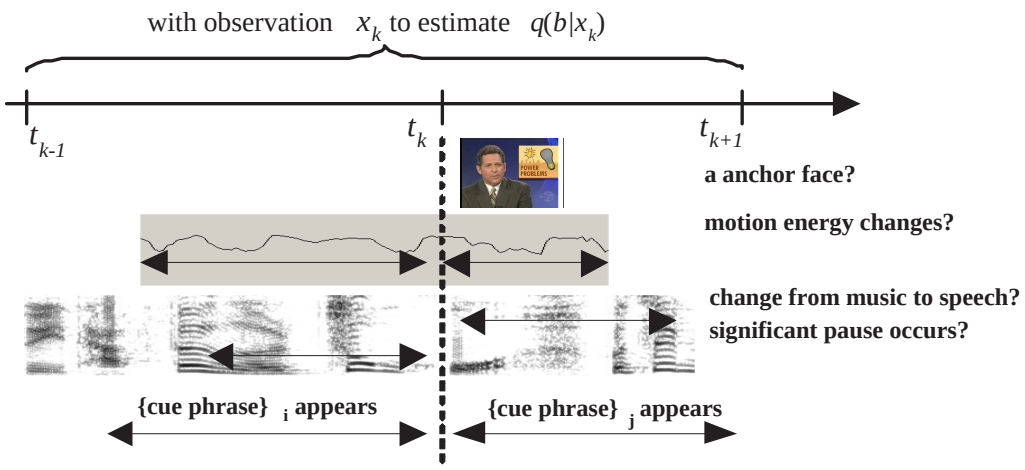


Figure 2.2: Estimation of the posterior probability $q(b|x_k)$, where $b \in \{0, 1\}$ is the random variable of boundary existence at the candidate point t_k, by fusing multiple mid-level perceptual features extracted from observation x_k.

story boundary) given the audio, visual or text data surrounding the point under examination, as shown in Equation 2.1. The construction process includes two main steps - parameter estimation and feature induction.

The estimated model, a posterior probability, is represented as $q_\lambda(b|x)$, where $b \in \{0, 1\}$ is a random variable corresponding to the presence or absence of a story boundary in the context x and λ is the estimated parameter set. Here x is the video and audio data surrounding a candidate point of story boundaries. From x we compute a set of binary features, $f_i(x, b) = 1_{\{g_i(x)=b\}} \in \{0, 1\}$. $1_{\{\cdot\}}$ is an indication function; g_i is a predictor of story boundary using the ith binary feature, generated from the feature wrapper (Section 2.3.2). f_i equals 1 if the prediction of predictor g_i equals b, and is 0 otherwise. The model is illustrated in Figure 2.2.

Given a labeled training set, we construct a linear exponential function as below,

$$q_\lambda(b|x) = \frac{1}{Z_\lambda(x)} \exp\{\sum_i \lambda_i f_i(x, b)\}, \tag{2.1}$$

where $\sum_i \lambda_i f_i(x, b)$ is a linear combination of binary features with real-valued parameters λ_i. $Z_\lambda(x)$ is a normalization factor to ensure Equation 2.1 is a valid conditional probability distribution. Basically, λ_i controls the weighting of the ith feature in estimating the posterior probability.

2.3.1.1 Parameter Estimation

The parameters $\{\lambda_i\}$ are estimated by minimizing the Kullback-Leibler divergence measure computed from the training set that has empirical distribution $\tilde{p}$. Since each video chunk is different, we just let $\tilde{p}$ be the inverse of the training sample size or uniform distribution. [40] also hold the sample assumption. The optimally estimated parameters are

$$\lambda^* = \underset{\lambda}{\operatorname{argmax}}\, D(\tilde{p} \parallel q_\lambda), \tag{2.2}$$

where $D(\cdot \parallel \cdot)$ is the Kullback-Leibler divergence defined as

$$D(\tilde{p} \parallel q_\lambda) = \sum_x \tilde{p}(x) \sum_{b \in \{0,1\}} \tilde{p}(b|x) \log \frac{\tilde{p}(b|x)}{q_\lambda(b|x)}. \tag{2.3}$$

Meanwhile, minimizing the divergence is equivalent to maximizing the log-likelihood defined as

$$L_{\tilde{p}}(q_\lambda) = \sum_x \sum_b \tilde{p}(x, b) \log q_\lambda(b|x). \tag{2.4}$$

The log-likelihood is used to measure the quality of the estimated model and $L_{\tilde{p}}(q_\lambda) \leq 0$ holds all the time. In the ideal case, $L_{\tilde{p}}(q_\lambda) = 0$ corresponds to a model q_λ, which is "perfect" with respect to $\tilde{p}$; that is, $q_\lambda(b|x) = 1$ if and only if $\tilde{p}(x, b) > 0$.

We use an iterative process to update λ_i until divergence is minimized. In each iteration, $\lambda_i^{'} = \lambda_i + \triangle\lambda_i$, where

$$\triangle \lambda_i = \frac{1}{M} \log \left\{ \frac{\sum_{x,b} \tilde{p}(x,b) f_i(x,b)}{\sum_{x,b} \tilde{p}(x) q_\lambda(b|x) f_i(x,b)} \right\} \tag{2.5}$$

and M is a constant to control the convergence speed. This formula updates the model such that the expectation values of features f_i with respect to the model are the same as their expectation values with respect to the empirical distribution from the training data. In other words, the expected fraction of events (x, b) for which f_i is "on" should be the same regardless if it is measured based on the empirical distribution $\tilde{p}(x, b)$ or the estimated model $\tilde{p}(x)q_\lambda(b|x)$. When the exponential model underestimates the expectation value of feature f_i, its weight λ_i is increased. Conversely, λ_i is decreased when overestimation occurs.

2.3.1.2 Feature Induction

Given a set of prospective binary features C and an initial maximum entropy model q, the model can be improved into $q_{\alpha,h}$ by adding a new feature $h \in C$ with a suitable weight α, represented as

$$q_{\alpha,h}(b|x) = \frac{\exp\{\alpha h(x,b)\} q(b|x)}{Z_\alpha(x)}, \tag{2.6}$$

where $Z_\alpha(x) = \sum_{b \in \{0,1\}} \exp\{\alpha h(x,b)\} q(b|x)$ is the normalization factor. A greedy induction process is used to select the feature that has the largest improvement in terms of gains, divergence reduction, or likelihood increase.

The gain $G_{qh}(\alpha) = L_{\tilde{p}}(q_{\alpha,h}) - L_{\tilde{p}}(q)$ represents how much improvement is yielded by introducing feature h with exponential weight α. The optimal α^* for each candidate feature h is derived under the assumption that the other parameters remain fixed during the computation. Since α^* occurs when $G'_{qh}(\alpha) = 0$, due to the con-

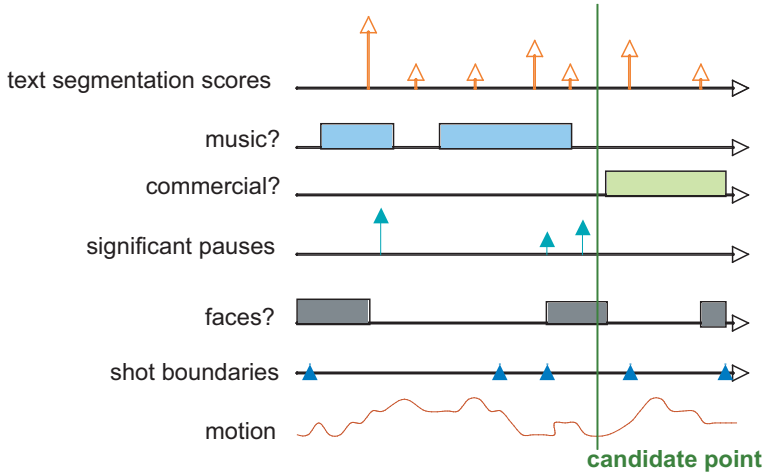


Figure 2.3: An example for multimodal fusion for boundary detection among candidates at abstract time points. These available (raw) multimodal features are mostly heterogeneous, asynchronous, variant time scale, and either real-valued or on impulse. A unification process is required to convert these raw features into homogeneous and effective representations for computational frameworks.

cavity of $G_{qh}(\alpha)$, we could apply Newton's method on $G'_{qh}(\alpha)$ to locate α^* through an iterative step such as $\alpha_{n+1} = \alpha_n - \frac{G'_{qh}(\alpha)}{G''_{qh}(\alpha)}$.

The selected feature h^* in each induction iteration is represented in Equation 2.7. h^* is then removed from the candidate pool C. The induction process iterates with the new candidate set $C - \{h^*\}$ until the stopping criterion is reached (e.g., upper bound of the number of features or lower bound of the gain).

$$
\begin{aligned}
h^* &= \underset{h \in C}{\operatorname{argmax}} \Big\{ \sup_{\alpha} \{ D(\tilde{p} \parallel q) - D(\tilde{p} \parallel q_{\alpha,h}) \} \Big\} \\
&= \underset{h \in C}{\operatorname{argmax}} \Big\{ \sup_{\alpha} \{ L_{\tilde{p}}(q_{\alpha,h}) - L_{\tilde{p}}(q) \} \Big\} \\
&= \underset{h \in C}{\operatorname{argmax}} \Big\{ G_{qh}(\alpha^*) \Big\}
\end{aligned}
\tag{2.7}
$$

2.3.2 Feature Wrapper

We have described the parameter estimation and feature induction processes from a pool of binary features in the previous sections. However, those available multi-modal multi-level features are usually asynchronous, continuous, or heterogeneous and require a systematic mechanism to integrate them. Meanwhile, the ME approach is based on binary features. For example in Figure 2.3, we have an example of multimodal fusion for boundary detection among candidates at abstract time points. These available (raw) multimodal features detected at different granularities or candidate points are mostly heterogeneous, asynchronous, and either real-valued or on impulse. A unification process is required to convert these raw features into homogeneous and effective representations for computational frameworks. For these purposes, we invent a feature wrapper to bridge raw multi-modal features and the ME model.

In Figure 2.4, we show the relation between the feature wrapper and the feature library which stores all raw multi-modal features. As the raw feature f_i^r is taken into the feature wrapper, it will be rendered into sets of binary features at each candidate point $\{t_k\}$ with the function $F_w(f_i^r, t_k, dt, v, B)$, which is used to take features from observation windows of various locations and lengths B, compute delta values of some features over time interval dt, and finally binarize the feature values against multiple possible thresholds, v.

Delta feature: The delta feature is quite important in human perception according to our experiment; for example, the motion intensity drops directly from high to low. Here we get the delta raw features by comparing the raw features with the time difference dt as $\Delta f_i^r(t) = f_i^r(t) - f_i^r(t - dt)$. Some computed delta features, in real values, will be further binarized in the binarization step with different thresholds.

Binarization: The story transitions are usually correlated with the changes in some dominant features near the boundary point. However, there is no prior knowledge about the quantitative threshold values for us to accurately detect "significant changes." For instance, what is the right threshold for the pitch jump intensity? How far would a commercial starting point affect the occurrence of a story boundary? Our strategy should be to find the effective binarization threshold level in terms of the fitness gain (i.e., divergence reduction defined in Equation 2.2) of the constructed model rather than the data distribution within the feature itself. Each raw or delta feature is binarized into binary features with different threshold levels v.

Observation windows: Different observation windows also impact human perception on temporal events. Here we take three observation windows $B = \{B_p, B_n, B_c\}$ around each candidate t_k. The first window B_p is the interval before the candidate point with window size T_w; another is the same time-span B_n after the candidate; the other is the window B_c surrounding the candidate, $[t_k - T_w/2, t_k + T_w/2]$. With different observation windows, we try to catch effective features occurring before, after, or surrounding the candidate points. This mechanism also tolerates time offset between different modalities. For example, the text segmentation boundaries or prosody features might imply likely occurrence of true story boundaries near a local neighborhood but not necessarily at the coincident location.

The dimension of binary features $\{g_j^i\}$ generated from raw feature f_i^r or delta feature Δf_i^r is the product of the number of threshold levels and number of observation windows (3, in our experiment). All the binary features generated at a candidate point are sequentially collected into $\{g_j\}$ and are further fed into the ME model; e.g., for pitch jump raw feature with 4 threshold levels, it would generate $3 \cdot 4 = 12$ binary features since we have to check if the feature is "on" in the 3

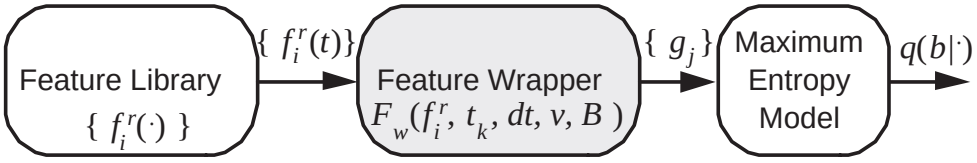


Figure 2.4: The raw multi-modal features f_i^r are collected in the feature library and indexed by raw feature id i and time t. The raw features are further wrapped in the feature wrapper to generate sets of binary features $\{g_j\}$, in terms of different observation windows, delta operations, and binarization threshold levels. The binary features are further fed into the ME model.

observation windows and each is binarized with 4 different levels.

2.4 Raw Multi-modal Multi-level Features

The raw multi-modal features in the feature library as shown in Figure 2.4, are from different feature detectors. We summarize some relevant and novel features in this section. Other features such as motion intensity and music/speech discrimination could be found in our prior work in [36].

In US news videos, the shot boundaries are directly from the common reference shot boundaries of TRECVID 2003. The shots have no durations of less than 2 seconds (or 60 frames); short shots have been merged with their neighbors. Therefore, many shots (roughly 20%) actually contain several sub-shots. When a shot contains several sub-shots, the corresponding key-frame is always chosen within the longest sub-shot. In Mandarin news, shots are detected by the approach of [43].

2.4.1 Anchor Face

A salient visual feature is the visual anchor segments reoccurring in the news video. Though the TRECVID 2003 videos are from TV news programs, the video quality varies a lot due to different recording resolution and lighting conditions. It is still

challenging to locate visual anchor segments from these 218 videos. Our prior work in [36] locates the visual anchor segments in three steps. (1) We first find those prospective face regions at each I-frame in each shot by an efficient face detector developed by [44]. It locates macro-blocks with possible skin-tone colors and further verifies vertical and size constraints both from DCT coefficients. The detector reports static face regions in each I-frame. (2) Within each shot, we take into account the temporal consistence by counting the face appearance frequency at each macro-block from the same shot to ensure that the faces should appear consistently in the same shot. (3) Regions of interest (ROI)–the upper body behind the face region, extended from detected face regions, of each shot are extracted and are further featured with HSV color histograms. A distance matrix between HSV histograms of regions of interest is later yielded and fed to an unsupervised agglomerative clustering algorithm [45] to locate the dominant cluster which implies the anchor segments in the entire video. We assume that the anchors are those dominant ROI among the candidates in the video.

To boost the performance, we add another face detection approach [46] that uses a GMM skin-tone model and geometric active contour to locate the possible set of face regions. From them, we repeat steps 1-3 to yield another set of possible anchor segments. Another unsupervised agglomerative clustering algorithm is applied on these two sets of anchor segments to distill more correct results.

2.4.2 Commercial

Frame matching based on image templates such as station logos and caption titles is used to discriminate commercial and non-commercial sections since we observe that in most news channels such as CNN and ABC the non-commercial portions are usually with certain logos or caption titles representing the station identification.

From the commercial detector, in the entire video, we label each frame as "1" if it is in the non-commercial portion (matched templates found in this frame) and "0" otherwise. The process yields the binary sequence $A \in \{0, 1\}$ of the entire video. However, the detection process could not avoid the noise due to the dynamic content within the commercial and the variances of production rules. Two morphological operators [47] are further applied and yield a smoothed result $A^{'}$ with temporal consideration as follows:

$$\begin{aligned} A^{'} &= (A \circ M_W) \bullet M_W, \\ M_W &= u[n] - u[n - W]. \end{aligned}$$

From $A^{'}$, we yield more correct commercial and non-commercial segments. Here, $\circ$ and $\bullet$ are morphological *OPEN* and *CLOSE* operators; M_W is a mask with W pulses represented by step function $u[n]$. We choose $W = 450$ and hypothesize that the caption titles or station logos in a non-commercial section might disappear but not longer than 450 frames or 15 seconds and there should be no caption logos lasting longer than this duration in the commercials.

2.4.3 Pitch Jump

Pitch contour has been shown to be a salient feature for the detection of syntactically meaningful phrase and topic boundaries [48, 49] and independent of language and gender [49]. A particularly useful behavior in pitch contour has been described as "pitch reset," mentioned in [41]. This behavior is characterized by the tendency of the speaker to lower his or her pitch towards the end of a topic and then to raise it, or reset it, at the beginning of the new topic. Past efforts have tried to characterize and detect this feature as a statistical combination of mean pitch, pitch variance

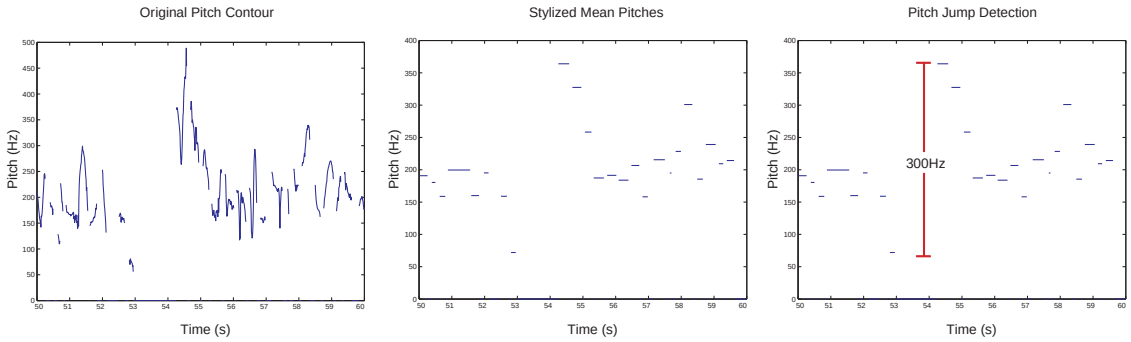


Figure 2.5: Visualization of original pitch contour, stylized chunk-level mean pitches, and pitch jump detection. For better illustration, the example is not normalized with the mean pitch.

[50] or stylized pitch contour [41], where slope of pitch change is relied upon heavily. We hypothesize that the mean pitch change will be sufficient for our task. Since the mean pitch and pitch variances vary between speakers, the pitch distributions used in our pitch jump detection scheme are normalized over same-speaker segments and represented in octaves.

In the pitch jump detection scheme, the pitch contour is extracted from audio streams automatically with the *Snack toolkit* [51]. The pitch contour is sampled at 100 estimates per second and is given as a pitch estimate (in Hz). The pitch estimates are converted into octaves by taking the base-2 logarithm of each estimate. To remove absolute pitch variances from different speakers, the octave estimates are then normalized in same-speaker segments by incorporating outputs from the ASR system (Section 2.4.5). The ASR system identifies segments in which there is a single active speaker. We iterate through all of these segments and normalize the octave pitch estimates according to the the mean in each coherent segment.

The pitch jump points were found by searching for points in the speech where the normalized magnitude of the inter-chunk pitch change was above a certain normalized threshold. We implemented the pitch jump detector by first segmenting the

pitch contour into chunks, which are groups of adjacent valid pitch estimates. We then find the mean normalized pitch of each chunk. After that, we find the change in pitch between each chunk by finding the difference or change between the mean normalized pitches of adjacent chunks. A threshold of the mean of the positive pitch change is then applied to all of the chunk boundaries. Chunk boundaries where the normalized pitch change is greater than the threshold are then selected as pitch jump points T_{pcj}. Figure 2.5 illustrates the processing applied to the pitch contour to find pitch jump points.

2.4.4 Significant Pause

Significant pause is another novel feature that we have developed for the news video segmentation task. It is somewhat inspired by the "pitch reset" behavior [41] that we discussed in Section 2.4.3 and the "significant phrase" feature developed by [50]. Significant pause is essentially an "AND" operation (Appendix A.) conducted on the pauses and the pitch jump points. We look for coincidences of pitch jump and pause in an attempt to capture the behavior where news anchors may simultaneously take a pause and reset their pitch contour between news stories.

In the significant pause detection scheme, we use the starting point of the pause (in seconds) and the duration of the pause (also in seconds), the pitch jump time (in seconds) and the normalized pitch jump magnitude. The significant pauses T_{sgps} are located by performing an *AND* operation, $T_{pcj} \odot_{0.1} T_{pas}$, on the pitch jump points T_{pcj} and pauses T_{pas} with 0.1-second fuzzy window. The normalized magnitudes and the pause durations associated with significant pauses are continuous raw features and are later binarized with different thresholds in the feature wrapper.

To gauge the potential significance of this feature before fusing into the framework, we test it on the reference boundaries and show the performance on ABC/CNN

Table 2.2: Boundary evaluation with significant pauses. The "uniform" column is to generate points uniformly with the same mean interval between significant pause points; it is 40.76 seconds in ABC and 41.80 seconds in CNN. The performance is evaluated with two fuzzy windows, 2.5-second and 5.0-second.

		Significant Pause			Uniform		
Set	ϵ	P	R	$F1$	P	R	$F1$
ABC	5.0	0.20	0.38	0.26	0.10	0.22	0.14
	2.5	0.16	0.34	0.22	0.10	0.22	0.14
CNN	5.0	0.40	0.45	0.42	0.20	0.24	0.22
	2.5	0.37	0.43	0.39	0.20	0.24	0.22

in Table 2.2. Taking the feature in CNN videos, we found its F1 measure to be 0.42, which is quite impressive compared with other features. As a comparison, another salient feature, anchor face on CNN, contributes to the story boundaries with a F1 measure of 0.51. In Table 2.2, we also compare the feature significance with uniformly generated points of the same mean interval between significant pause points. Apparently, significant pauses deliver more significance than uniformly sampled points both on ABC and CNN news. However, the feature performance on ABC is less significant than that on CNN.

According to our observation, the feature points usually match the boundaries where the anchors tend to raise a new topic. Upon further inspection, we find that the non-significance of significant pause on ABC and the success of it on CNN is due not to an inconsistency in the feature detection, but to a difference in anchor behaviors and production styles on the two stations. The CNN data is from CNN Headline News, which is a newscast dedicated to bringing all of the top stories quickly in short formats. On CNN, the anchors often switch between short stories without any visual cues being conveyed to the viewers. They compensate for this lack of visual cue by emphasizing the change of story with their voice by injecting a pause and resetting their pitch. This pause/reset behavior is very salient and is

the behavior that we were trying to capture with our features. Closer examination of the ABC data, however, revealed that the anchor behavior on story changes is rather different from CNN. The ABC news is a nightly newscast that presents fewer stories in 30 minutes than CNN news and dedicates more time to each story. On ABC, the stories rarely change without some sort of visual cue or without cutting back to the anchor from some field reports. This visual obviousness of story change causes the anchors to have less of a need for inserting story-change cues in their speech. We therefore attribute the weakness of the audio features on ABC, when compared with CNN, to a fundamental difference in anchor speech patterns on the two channels.

2.4.5 Speech Segments and Rapidity

We extract the speech segment, a continuous segment of the same speaker, from the ASR outputs. Two adjacent speech segments might belong to the same speaker, separated by a long non-speech segment, a pause or music. The segment boundaries, starting or ending, might imply a story boundary. However, there are still some boundaries that are within the speech segment; e.g., an anchor person briefs several story units continuously without a break.

In the ASR outputs, along with the TRECVID 2003 data, as illustrated in the following paragraph, the SGML tags are used to describe the recognized transcripts and associated structures. A speech segment is defined by the tag "<SpeechSegment>" and the associated attributes "*stime*" and "*etime*," which represent the starting and ending points of the segment. Transcribed words are described by the tag "<Word>." In this example, the speech region starts at 11.80 second and ends at 20.73 second in the video "19980204_CNN.mpg."

```
<Audiofile filename="19980204_CNN">
    <SpeechSegment lang="us" spkr="FS4" stime="11.80" etime="20.73">
        <Word stime="12.35" dur="0.34" conf="0.953"> THREE </Word>
        <Word stime="12.69" dur="0.49" conf="0.965"> EUROPEAN </Word>
        <Word stime="13.18" dur="0.64" conf="0.975"> BALLOONISTS </Word>
        ...
```

We further measure the speech rapidity by counting words per second in each segment. Usually the speaker tends to speak faster at the start of a story. For those speech segments with speech rapidity larger than the mean plus a stand deviation measured within the same video are fast speech segments S_{fsp}.

2.4.6 ASR-based Story Segmentation

The ASR-based story segmentation scheme in this work is a combination of decision tree and maximum entropy models. It is based on the IBM story segmentation system used in TDT-1 evaluation [42]. It takes a variety of lexical, semantic and structural features as inputs. These features are calculated from the ASR transcript. The performance in the validation set is shown in Table 2.3.

Prior to feature extraction, the ASR transcript is converted into "sentence" chunks. The preprocessing converts the transcript into chunks consisting of strings of words delimited by non-speech events such as a silence or pause. These chunks are then tagged by an HMM part of speech (POS) tagger and then stemmed by a morphological analyzer which uses the POS information to reduce stemming ambiguities. The feature extractor extracts a variety of features from the input stream, for example, the number of novel nouns in a short lookahead window [42].

The decision tree model is similar to what [52] proposed. It uses three principal groups of input features to model the probability of a story boundary. Our

observation has been that the most important group of features related to questions about the duration of non-speech events in the input. The second group of features depends on the presence of key bigrams and key words indicating the presence of story beginning (such as "good morning", "in new-york"). The last group of features compares the distribution of nouns on the two sides of the proposed boundary. The ME models uses three categories of features as well. The first group of features encodes the same information used by the decision tree model. The next group of features looks for n-gram ($n \leq 3$) extracted from windows to the left and right of the current point. The last category of features is structural and detect large-scale regularities in the broadcast news programming such as commercial breaks [42].

In the specific ASR transcript provided with the TRECVID 2003 evaluation data, we observe that a manually provided story boundary in the development data mostly does not correspond to non-speech events in the transcript. This is observed in 15-20% of the boundaries. One possibility is that the ASR system used to generate the transcript was not tuned to report on short pauses. Since our system relies on non-speech events to chunk the text and compute features, in the instances when the manual boundary did not correspond to a non-speech event in the transcript, we inserted a 0-second non-speech event to facilitate feature extraction at such points in the development data. We note that this is a limitation of our system as this pre-processing cannot be achieved for the test data. However, with tuning of the ASR engine to provide information about short-pauses, this limitation becomes less relevant.

2.4.7 Syllable Cue Terms (in Mandarin)

Cue terms are essential to provide lexical and semantic information for story boundaries. Instead of using cue terms in word or character levels, in Mandarin news, we

derive the syllable cue terms directly from the Chinese speech recognizer developed by [53]. There are several reasons for choosing syllable cue terms rather than Chinese characters or words. First, the numbers of commonly used Chinese words and characters are around 90,000 and 10,000 respectively. However, there are only 1345 tonal syllables and un-tonal ones are even as few as 408. Also, Mandarin Chinese has very free word order in sentences without word boundaries. New words are very easily generated by combining a few characters or syllables. Meanwhile, the construction of Chinese words is usually quite flexible. Such cue phrases are closer to the semantic level than the above audio-visual features, and are useful indicators of story boundaries.

We check whether any cue phrases exist in the speech surrounding candidate points. The bag of cue terms are selected automatically with term frequency and inverse document frequency (TFIDF) approach from the end and beginning of story boundaries in the training set. For Mandarin speech, we specifically adopt 5 types of syllable-level combinations suggested by [53] to define the cue phrases. The syllable-level combinations have been shown effective in broadcast speech information retrieval by [53]. These 5 types include 3 from syllable segments $S(n)$ and 2 from syllable pairs $P(n)$. Assuming there is a ten-syllable sequence $s_1 s_2 s_3 ... s_9 s_{10}$, the combinations of syllable terms are as the following:

- Syllable segments $S(n)$ (n-gram syllables):

  $S(1)$ e.g.: $(s_1)(s_2)(s_3)...$

  $S(2)$ e.g.: $(s_1 s_2)(s_2 s_3)...$

  $S(3)$ e.g.: $(s_1 s_2 s_3)(s_2 s_3 s_4)...$

- Syllable pairs $P(n)$ separated by n syllables:

$P(1)$ e.g.: $(s_1 s_3)(s_2 s_4)(s_3 s_5)...$

$P(2)$ e.g.: $(s_1 s_4)(s_2 s_5)(s_3 s_6)...$

These syllable terms are designed to cover all possible character combinations of Chinese syllables. Further feature selection based on TFIDF is conducted on these combination terms.

Some of the selected cue terms contain the location where the field report was recorded usually appearing at end of a footage; i.e. /tai-bei-bao/ (part of "report in Taipei") and /tao-yuan-bao/ (part of "report in Taoyuan") selected near the end of the story boundaries in the training set. Some terms just lead to another new topic; i.e. /zai-guo-ji/ (part of "in international news"). More interestingly, some combine the lexical and acoustic features at the topic transition points and are composed of a pause duration and the topic changing terms, frequently used in Mandarin, such as /__-da-yue/ (pause + "about") and /_-er-lin/ (pause + "another"), where _ means a pause or silence also reported by the Chinese speech recognizer.

The existence of such cue syllable terms, selected by TFIDF, can be formulated as binary perceptual features in the feature library for international broadcast news video story boundary detection. Though with hundreds of syllable segments of cue pairs or segments, the feature induction process can select the salient ones in a rigorous approach – likelihood reduction modeled by the ME approach.

2.4.8 Combinatorial Features

We observe that some of the features in a specific state or the combination of some features would provide more support toward the story boundaries; e.g., the significant pauses are composed of pitch jump and pause points. With the help of these features, we are able to boost the challenging parts or rate events useful

for story segmentation. Meanwhile, the combinatorial features are useful to embed some domain knowledge regarding the video dynamics. Some of these features also present significance in the feature induction process as shown in Table 2.4. We further generate some combinatorial binary features based on previous features, temporal operators ($\odot_\epsilon$) and filters (Ψ_ϵ) (Appendix A.) such as:

- *Pitch-jump near the start of the speech segments*: we tend to use this feature to catch the instance when a speech segment starts with a pitch reset. The feature is composed of the pitch jump points T_{pcj} and the start of speech segments T_{ssps} and are computed from $T_{pcj} \odot_\epsilon T_{ssps}$.

- *Significant pauses near shot boundaries*: we hope to use this feature to catch the short briefings without leading visual anchor segments at the start of stories but with shot changes such as types `d` or `e` in Figure 2.1. The feature is calculated with significant pauses T_{sgps} and shot boundaries T_{sht} by taking $T_{sgps} \odot_\epsilon T_{sht}$.

- *Fast speech segments within the non-commercial sections*: we hope to catch the starting points of fast speech segments T_{sfsp} within the non-commercial segments S_{ncom} by taking $\Psi_\epsilon(T_{sfsp}, S_{ncom})$.

- *Significant pauses within the fast speech segments or non-commercial segments*: we design these features to boost the detection in short news briefings. Usually the anchor tends to speak faster or has a pitch reset when changing the topic. The features are yielded by simply filtering significant pauses T_{sgps} within fast speech segments S_{fsp} or non-commercial segments S_{ncom} by filters $\Psi_\epsilon(T_{sgps}, S_{fsp})$ or $\Psi_\epsilon(T_{sgps}, S_{ncom})$.

2.5 Experiments

We conduct feature fusion and induction on Mandrin and US news videos. The overall performance and induced features are introduced in the following sections.

2.5.1 Data Set

In US broadcast news videos, we use 218 half-hour ABC World News Tonight and CNN Headline News broadcasts recorded by the Linguistic Data Consortium from late January 1998 through June 1998. The video is in MPEG-1 format and is packaged with associated files including automatic speech recognition (ASR) transcripts (provided by [54]) and annotated story boundaries, called *reference boundaries*. The data are prepared for TRECVID 2003 [3, 55] with the goal of promoting progress in content-based video retrieval via open metric-based evaluation.

From the 111 videos in the development set, we found that the story length ranges from 4.05 to 223.95 seconds on CNN and from 7.22 to 429.00 seconds on ABC. The average story length on CNN is 42.59 second and is 71.47 on ABC. Apparently, CNN tends to have shorter and more dynamic stories.

As for Mandarin news, due to data availability, we collected a data set of 3.5-hour news programs with totally 100 stories, mixing from 3 different channels and different recording times. The corpora are with 8 unique anchorpersons in 10 video clips. The average story length on Mandarin news is 89.24 seconds; the longest is 259.40 seconds and the shortest is 16.02 second. Apparently, these news broadcasts possess different product rules in terms of story length and dynamics.

2.5.2 Boundary Detection on Mandarin News

To understand the behavior in international news, we deliberately select foreign news (Mandarin news in Taiwan) to test our model. Foreign news videos usually do not have special markers in closed captions to indicate the story boundaries. The Mandarin news video is 3.5-hour and with totally 100 stories, mixing from 3 different channels and different recording times. There are 8 unique anchorpersons among 10 video clips. We use a subset (38 stories) for training and the rest for testing.

2.5.2.1 Boundary performance on Mandarin news

The raw multi-modal features used in the Mandarin test set are grouped into 3 categories including, A $\equiv$ {music/speech discrimination, speaker identification}, V $\equiv$ {motion intensity, occurrences of superimposed caption, anchor face, static face}, and S $\equiv$ {syllable cue terms}. With all the features, S+A+V, the F1 measure is 0.90, 0.82 in A+V, and 0.47 in S alone. With anchor face alone, the F1 measure is 0.65. The overall boundary detection performance is higher than those in the US news videos, either on CNN or ABC (Section 2.5.3). It might be due to that the Mandarin news we had has simpler production rules and the original annotation on the data set made the assumption that each story boundary coincides with the shot boundaries, which is mostly not the case in US news videos, which contain many types of story boundaries, such as type c illustrated in Figure 2.1.

2.5.2.2 Significant features on Mandarin news

As described in section 2.3.1.2, the optimal features are selected based on divergence reduction or log-likelihood increase. Log-likelihood $L_{\tilde{p}}(q_\lambda)$ in Equation 2.4 is used

to measure the quality of the estimated model . Given the training data, the ideal model should have a log-likelihood of 0 and a random-guess model's is -0.6931. The log-likelihood value after each feature induction is shown and compared in Figure 2.6. It is not surprising that with all features S+A+V reached a high log-likelihood. A notable degradation in log-likelihood is observed when the syllable cue phrases S are excluded. However, when the syllable cue terms are used alone, the performance is degraded significantly. This indicates that the S features are helpful when combined with the audio and visual features. But when used alone, the audio-visual combinations are better than the cue phrases features. The comparison, in terms of F1 measures, between ASR text and audio-visual features in CNN/ABC news is similar and can be seen in Table 2.3.

We further inspect the first best feature selected in each feature set. In S+A+V or A+V, the first feature is an A/V combination feature, i.e. co-occurrence of a static face and an anchor speech segment. The finding confirms the popular assumption that story boundaries usually start with the anchorperson shot. However, according to our experiment, with anchor face only, we gained 0.61 F1 measure only in Mandarin news. The second selected feature in S+A+V is the occurrence of the syllable cue terms $S(3)$ selected at the end of the story boundary (Section 2.4.7); i.e. /tai-bei-bao/ (part of "report in Taipei") and /tao-yuan-bao/ (part of "report in Taoyuan"). These cue terms match our observations: On the Mandarin news data set, at the end of a report, the field reporters usually end the footage with the city name where they reported.

2.5.3 Boundary Detection on US News

In US news videos, we use 111 half-hour video programs for development, 66 of which are used for detector training and threshold determination. The remaining

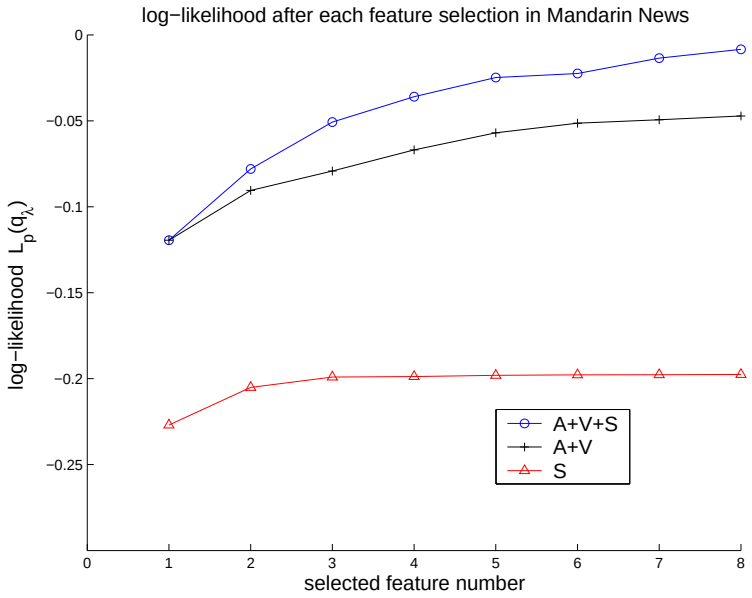


Figure 2.6: Log-likelihood $L_{\tilde{p}}(q_\lambda)$ after each feature induction iteration in Mandarin news. The ideal value is 0 and the random-guess log-likelihood on the training data is -0.6931. The legends are defined in section 2.5.2.1.

45 video programs are further separated for fusion training and model validation. In TRECVID 2003, we have to submit the performance of a separate test set, composed of 107 ABC/CNN videos.

In the experiment, we try to ensure that we have adequate training sample size. For example, to train a CNN boundary detection model with A+V modalities, we use 34 CNN videos (∼17 hours) with 1142 reference boundaries and 11705 candidate points. Each candidate is with 195 binary features, among which the feature induction process selects 30 of them.

The segmentation measure metrics are precision P_{seg} and recall R_{seg} and are defined in the following. According to the TRECVID metrics, each reference boundary is expanded with a fuzzy window of 5 seconds in each direction, resulting in an evaluation interval of 10 seconds. A reference boundary is *detected* when one or more computed story boundaries lie within its evaluation period. If a computed boundary

does not fall in the evaluation interval of a reference boundary, it is considered a *false alarm*. The precision P_{seg} and recall R_{seg} are defined in Equations 2.8 and 2.9; $|\cdot|$ means the number of boundaries; B_{cpt} and B_{ref} are computed and reference boundaries (given boundary ground truth) and formal evaluation fuzzy window ϵ is 5 second.

$$P_{seg} = \frac{|\text{computed boundaries}| - |\text{false alarms}|}{|\text{computed boundaries}|} = \frac{|B_{cpt} \odot_{\epsilon} B_{ref}|}{|B_{cpt}|} \quad (2.8)$$

$$R_{seg} = \frac{|\text{detected reference boundaries}|}{|\text{reference boundaries}|} = \frac{|B_{ref} \odot_{\epsilon} B_{cpt}|}{|B_{ref}|} \quad (2.9)$$

2.5.3.1 Boundary performance on US news

The performance in CNN/ABC news videos is shown in Table 2.3, where A means audio cues, V is visual cues and T is text. At A+V, F1 measure on ABC is better than CNN. It is probably due to ABC stories being dominated by anchor segments or type `a` in Figure 2.1; while in CNN, there are some short briefings and tiny dynamic sports sections which are very challenging and thus cause a lower recall rate and these short stories are types `d` and `e` in Figure 2.1. In CNN news, the A+V boosts the F1 measure of anchor face from 0.51 to 0.69. The main contributions come from significant pauses and speech segments since they compensate CNN's lack of strong visual cues.

Since we are estimating posterior probability $q(b|\cdot)$ accounting for the existence of a story boundary, a straightforward boundary decision is just to select those candidate points with $q(1|\cdot) > 0.5$. We found that the story segmentation problem with the ME model also suffers from imbalanced-data learning presented by [56] since the boundary samples are much fewer than non-boundary ones. The best F1 measure of boundary detection does not come from the decision threshold 0.5 but requires

Table 2.3: Boundary detection performance in ABC/CNN news with the best decision thresholds, 0.25 for CNN and 0.35 for ABC, on posterior probability $q(b = 1|\cdot)$.

	ABC			CNN		
Modalities	P	R	$F1$	P	R	$F1$
Anchor Face	0.67	0.67	0.67	0.80	0.38	0.51
T	0.65	0.55	0.59	0.50	0.70	0.59
A+V	0.75	0.67	0.71	0.70	0.68	0.69
A+V+T	0.85	0.70	0.76	0.72	0.75	0.73

a boundary movement [56] meaning that we have to move the posterior probability threshold from 0.5 to a certain shifted value, determined in a sperate small development set to maximize the boundary detection F1 measure. Here we take the decision values 0.25 for CNN and 0.35 for ABC. Intuitively, a smaller value trades a lower precision for a higher recall rate. For a more complete performance comparison, we plot the precision vs. recall curves of story segmentation performance with modalities A+V and A+V+T on ABC/CNN news in Figure 2.7.

As for fusing modality features such as fusing text segmentation into A+V, the precision and recall are both improved even though the text feature is with real-valued scores and computed at non-speech points only, which may not coincide with those used for the audio-visual features. It is apparent that the fusion framework successfully integrates these heterogeneous features which compensate for each other.

An interesting phenomenon is that the introduction of text segmentation features in ABC news improves the precision significantly at the low or mid-recall area (Figure 2.7). The enhancement on CNN news is less and probably due to that most of difficult cases in CNN are those with short briefings and sports dynamics.

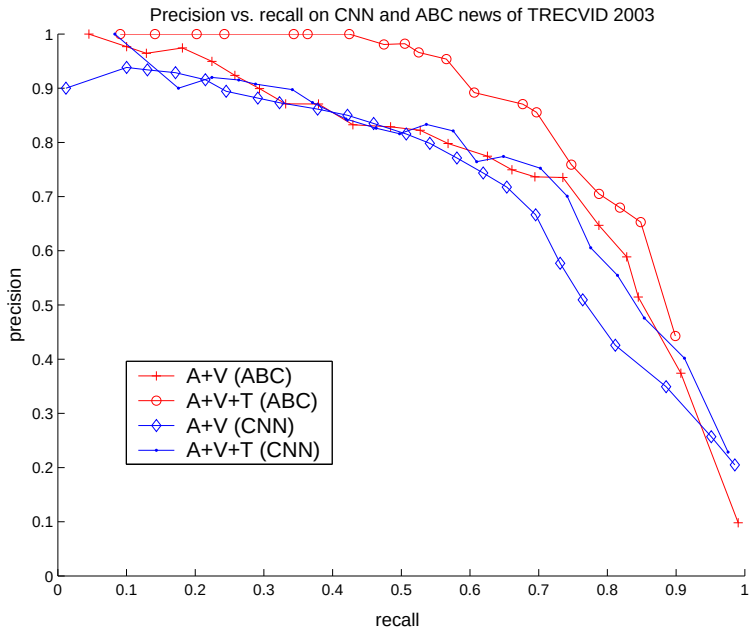


Figure 2.7: Precision vs. recall curves of story segmentation performance with modalities A+V and A+V+T on ABC/CNN news.

2.5.3.2 Significant features on US news

To illustrate the binary features induced in the feature selection process of the ME model, we list the first 12 induced features from the CNN A+V model in Table 2.4. Clearly, the anchor face feature is the most relevant feature according to the training set but the relevance also depends on the location of observation windows. The next induced binary feature is the significant pause within the non-commercial section from combinatorial features. The audio pauses and speech segments are also relevant to the story boundaries since they usually imply a topic change. Interestingly, those features further filtered with non-commercial sections deliver more significance than the original features since the video contains transitional dynamics or genres and features in different genres have different significance; for example, significant pauses in non-commercial segments are much more relevant to boundaries than those in commercials; also, the rapid speech segments in commercials are less relevant to

story boundaries. Another interesting feature is the 7th induced binary feature that a commercial starting in 15 to 20 seconds after the candidate point would imply a story boundary at the candidate point. After our inspection on CNN news, it matches the dynamics of the news source since before turning to commercials the video is finished back to the anchors who would take seconds to shortly introduce coming news. The binarization threshold for the commercial feature is selected by the feature induction process in the training set rather than by heuristic rules.

2.6 Summary

Multimodal fusion and feature selection has been of great interest to the research community. We investigate such issues in the broadcast news video story segmentation, which remains a challenging issue even after years of research. We believe multi-modality fusion through effective statistical modelling and feature induction are keys to the solutions. In this chapter, we have proposed a systematic framework for fusing multi-modal features at different levels. We demonstrated significant performance improvement over single modality solutions and illustrated the ease in adding new features through the use of a novel feature wrapper and ME model on CNN, ABC, and Mandarin news programs.

These mid-level perceptual features are typically manually decided relying on expert knowledge of the application domain. Once the mid-level features are chosen, additional manual efforts are needed to annotate training data for learning the detector of each mid-level feature. A further goal is to automate the selection process of the mid-level features given defined semantic class labels. In addition, dimensionality of the mid-level features will be much lower than that of the low-level features. We will further introduce such a framework in chapter 3 and its empirical

experiments in chapter 4.

Table 2.4: The first 12 induced features from the CNN A+V model. λ is the estimated exponential weight for the selected feature; Gain is the reduction of the divergence as the feature added to the previously constructed model. $\{B_p, B_n, B_c\}$ are three observation windows; B_p is before the candidate point; B_n is after the candidate point; B_c is surrounding the candidate.

#	Raw Feature Set	λ	Gain	Descriptions
1	Anchor face	0.4771	0.3879	An anchor face segment just starts in B_n.
2	Significant pause + Combinatorial	0.7471	0.0160	A significant pause within the non-commercial section appears in B_c.
3	Pause	0.2434	0.0058	An audio pause with the duration larger than 2.0 second appears in B_c.
4	Significant pause	0.7947	0.0024	B_c has a significant pause with the pitch jump intensity larger than the normalized pitch threshold v^0_{pcj} and the pause duration larger than 0.5 second.
5	Speech segment	-0.3566	0.0019	A speech segment starts in B_p.
6	Speech segment	0.3734	0.0015	A speech segment starts in B_c.
7	Commercial	1.0782	0.0015	A commercial starts in 15 to 20 seconds after the candidate point.
8	Speech segment	-0.4127	0.0022	A speech segment ends in B_n.
9	Anchor face	0.7251	0.0016	An anchor face segment occupies at least 10% of B_n.
10	Pause	0.0939	0.0008	B_c has a pause with the duration larger than 0.25 second.
11	Speech rapidity + Combinatorial	0.6196	0.0006	A fast speech segment within the non-commercial section starts in B_c.
12	Significant pause	-0.5161	0.0004	B_n has a significant pause with pitch jump intensity larger than the normalized pitch threshold v^0_{pcj} and pause duration larger than 0.5 second.

Chapter 3

Automatic Discovery of Semantic-Consistent Clusters

Recent research in video analysis has shown a promising direction, in which mid-level features (e.g., people, anchor, indoor) are abstracted from low-level features (e.g., color, texture, motion, etc.) and used for discriminative classification of semantic labels. However, in most systems, such mid-level features are selected manually. In this chapter, we propose an information-theoretic framework, to automatically discover adequate mid-level features – semantic-consistent clusters. The problem is posed as mutual information maximization, through which optimal cue clusters are discovered to preserve the highest information about the semantic labels. We extend the Information Bottleneck framework to high-dimensional continuous features to locate these semantic-consistent clusters as a new mid-level representation. The biggest advantage of the proposed approach is to remove the dependence on the manual process of choosing mid-level representations and the huge labor cost involved in annotating the training corpus for training the detector of each mid-level feature. The proposed IB framework is general and effective, with the potential to solve other problems in semantic video analysis.

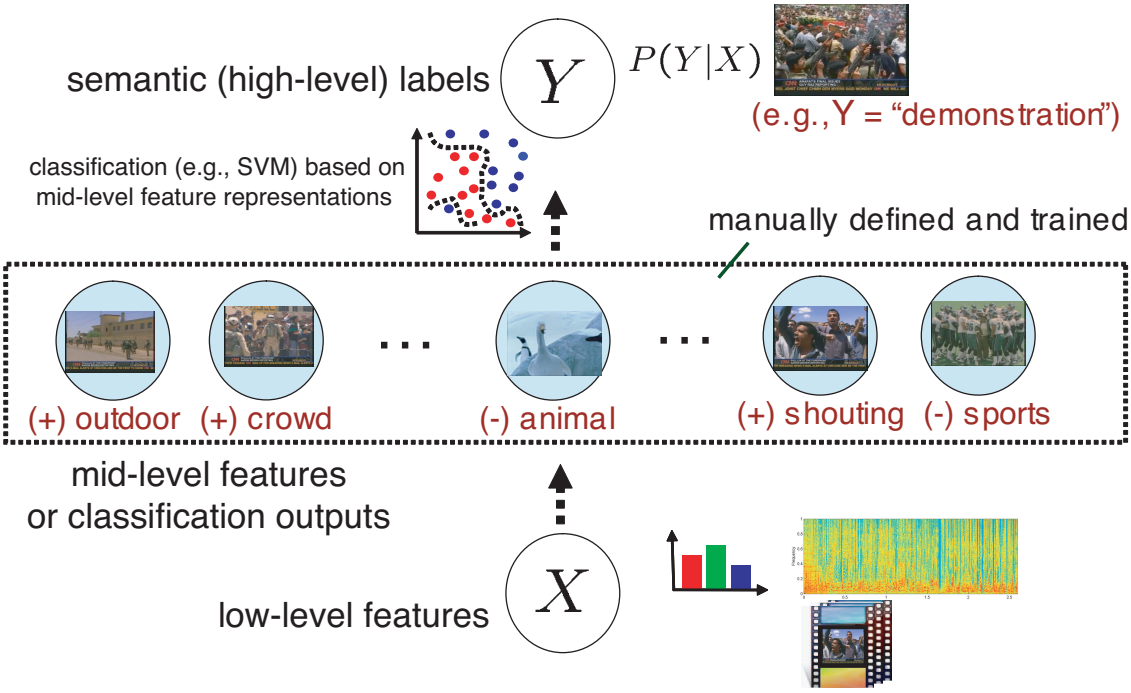


Figure 3.1: An example of using mid-level feature representations to support generic video classification by classifying low-level features X to semantic label Y. The mid-level features (e.g., "outdoor," "crowd," etc.) are mostly manually defined, annotated, and trained. Meanwhile, they are selected because of being positively "+" or negatively "-" related to the semantic label (e.g., "demonstration").

We apply these automatic semantic-consistent clusters as new mid-level representations for the video classification problem in chapter 4. More excitingly, these automatic semantic -consistent clusters can be extended to de-noise video search posterior probability and localize feature density estimation for video search reranking detailed in chapter 5.

3.1 Introduction

In the research of video retrieval and analysis, a new interesting direction is to introduce "mid-level" features that can help bridge the gap between low-level features and semantic concepts. Examples of such mid-level features include location (in-

door), people (male), production (anchor), etc., and promising performance due to such mid-level representations have been shown in recent work in news segmentation and retrieval [55, 57]. It is conjectured that mid-level features are able to abstract the cues from the raw features, typically with much higher dimensions, and provide improved power in discriminating video content of different semantic classes. However, selection of the mid-level features is typically manually done relying on expert knowledge of the application domain. Once the mid-level features are chosen, additional extensive manual efforts are needed to annotate training data for learning the detector of each mid-level feature.

For example, Figure 3.1 illustrates a generic framework to classify low-level observations based on mid-level feature representations. The low-level features X are first classified or quantized into real-valued representations in the dimensions of mid-level features, which are, in most cases, decided manually. For example, if $Y =$"demonstration," there might be several concepts are related, either positively ("+") or negatively ("-") to the semantic label. Intuitively, in this example, we can have "outdoor," "crowd," "shouting," "sports," etc. Based on these mid-level representations, a classifier, either generative (e.g., Hidden Markov Model) or discriminative (e.g., Support Vector Machines), can be applied to classify them into semantic labels Y.

Such approaches are intuitive and widely adopted in prior work but incur severe problems by requiring huge amounts of labor for annotation and classifier training. Assuming that it takes 15 minutes for each mid-level feature annotation and 30 videos (15 for training and 15 for validation) to adequately train a mid-level feature detectors, then assessing a new data or video source over the 10 mid-level features would require at least $10 \cdot (15+15) \cdot (15/60) \approx 75$ hours. Note that the extra manual effort to determine which 10 concepts to use are not counted here. Moreover, the

other cost for training these classifiers is mandatory. Such a process is not feasible for a large-scale video database or in the real applications where new video sources emerge very frequently [4]. Meanwhile, another drawback of such manual efforts is that there is no quantitative measure to describe if those selected mid-level features are adequate, too few, or too many.

To address such severe problems in lage-scale video databases, our goal is to automate the selection process of the mid-level features given defined semantic class labels Y. Given a collection of data, each consisting of low-level features and associated semantic (high-level) labels, we want to discover the mid-level features automatically. There is still a need for labeling the semantic label of each data sample, but the large cost associated with annotating the training corpus for each manually chosen mid-level feature is no longer necessary. In addition, the dimensionality of the mid-level features will be much lower than that of the low-level features. As illustrated in Figure 3.1, these mid-level features are selected because they are either positively or negatively related to the semantic labels. To discover these mid-level features automatically, such observations motivate our proposed work which utilizes the mutual information between the low-level features and the (target) semantic labels.

Discovery of compact representations of low-level features can be achieved by conventional clustering methods, such as K-means and its variants. However, conventional methods usually derive clusters that have high similarities in the low-level feature space but often do not have strong correlation with the semantic labels. Some clustering techniques, such as LVQ [58], take into account the available class labels to influence the construction of the clusters and the associated cluster centers. However, the objective of preserving the maximum information about the semantic class labels was not optimized.

Recently, a promising theoretic framework, called Information Bottleneck (IB), has been developed and applied to show significant performance gain in text categorization [59, 60]. The idea is to use an information-theoretic optimization methodology to discover "cue word clusters" which can be used to represent each document at a mid level, from which each document can be classified to distinct categories. The cue clusters are the optimal mid-level clusters that preserve the most mutual information between the clusters and the class labels.

In this chapter, we propose new algorithms to extend the IB framework to the visual domain, specifically video. Starting with raw features such as the color, texture, and motion of each shot, our goal is to discover the cue clusters that have the highest mutual information about the final class labels, such as video story boundary, semantic concepts, or video search relevance. Our work addresses several unique challenges. First, the raw visual features are mainly continuous (unlike the word counts in the text domain) and of high dimensions. We propose a method to approximate the joint probability of features and labels using kernel density estimation. Second, we propose an efficient sequential method to construct the optimal clusters and a merging method to determine the adequate number of clusters. Third, a feature selection criteria based on mutual information can be utilized to quantitatively select informative mid-level features or reject meaningless ones. Finally, we develop a rigorous analytic framework to project new video data to the visual cue clusters. The probabilities of such projections over the cue clusters are then used for the final discriminative classification of the semantic labels.

Our work is significantly different from [61] which uses the IB principle for image clustering. In [61], 3 CIE-Lab colors and 2 horizontal and vertical positions are used as the raw input features. The dimension is much lower than that in this chapter. The distribution in the raw feature space was first fit by a Gaussian Mixture

Model (GMM), whose estimated parameters were then used for the IB clustering. In contrast, we do not assume specific parametric models in our approach, making our results more generalizable. Most importantly, preservation of mutual information about the semantic labels was not addressed in [61].

We introduce the IB framework over continuous high-dimentional features in section 3.2. The methods to conduct feature projection and feature selection is described in section 3.3, followed by section 3.4, where we have a quick summary and lead in to promising applications based on the proposed framework.

3.2 The IB Principle

The variable X represents features and Y is the variable of interest or auxiliary labels associated with X. X might be documents or low-level feature vectors; Y might be document types in document categorization or sematic class labels, or search relevance. In this context, we want the mapping from $x \in X$ to cluster $c \in C$ to preserve as much information about Y as possible. As in the compression model, the framework passes the information that X provides about Y through a "bottleneck" formed by the compact summaries in C. On the other hand, C should catch the consistent semantics of X. The semantics are defined by the conditional distribution over the label Y (i.e., $p(y|x)$).

The above goal can be formulated by the IB principle, which states that among all the possible clusterings of the objects into a fixed number of clusters, the optimal clustering is the one that minimizes the loss of mutual information (MI) between the features X and the auxiliary labels Y. Assume that we have joint probability $p(x, y)$ between these two random variables. According to the IB principle, we seek a clustering representation C such that, given a constraint on the clustering quality

$I(X;C)$, the information loss $I(X;Y) - I(C;Y)$ is minimized.

Note that in this work we emphasize discovering the visual cue clusters of images or videos represetned in high-dimensional visual features. To understand more regarding the basic ideas of IB principle in text categorization research, we can refer to [59, 60]. We also have a few examples of "cue word clusters" – clusters of semantic-consistent words – of the same implementation depicted in section 6.2.1.

3.2.1 Mutual Information

For discrete-valued random variables X and Y, the MI [38] between them is

$$I(X;Y) = \sum_{y}\sum_{x} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}.$$

We usually use MI to measure the dependence between variables. In the IB framework, we represent the continuous D-dimensional features with random variable $X \in R^D$; the auxiliary label is a discrete-valued random variable Y representing the target or relevance labels. We have feature observations with corresponding labels in the training set $S = \{x_i, y_i\}_{i=1..N}$. Since X is continuous, the MI is

$$I(X;Y) = \sum_{y}\int_{x} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dx.$$

However, based on S, the practical estimation of MI from the previous equation is difficult. To address this problem, the histogram approach is frequently used but only works for scalar-valued variables. An alternative approach is to model X through Gaussian Mixture Model (GMM) which is limited to low-dimensional features due to the sparsity of data in high-dimensional spaces [61].

We approximate the continuous MI with Eq. 3.1 for efficiency. The summariza-

tion is only over the observed data x_i assuming that $p(x, y) = 0$ if $x \notin S$. Similar assumptions are used in other work (e.g., the approximation of Kullback-Leibler divergence in [61]). According to our experiments, the approximation is satisfactory in measuring the MI between the continuous feature variable X and the discrete auxiliary variable Y.

$$I(X;Y) \cong \sum_{x_i \in S} \sum_{y} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)} \tag{3.1}$$

3.2.2 Kernel Density Estimation

To approximate the joint probability $p(x, y)$ based on the limited observations S, we adopt the kernel density estimation [62]. The method does not impose any assumption on the data and is a good method to provide statistical modeling among sparse or high-dimensional data.

The joint probability $p(x, y)$ between the feature space X and the auxiliary label Y is calculated as follows:

$$p(x, y) = \frac{1}{Z(x, y)} \sum_{x_i \in S} K_\sigma(x - x_i) \cdot \bar{p}(y|x_i), \tag{3.2}$$

where $Z(x, y)$ is a normalization factor to ensure $\sum_{x,y} p(x, y) = 1$, K_σ (Eq. 3.3) is the kernel function over the continuous random variable X. $\bar{p}(y|x_i)$ is an unsmoothed conditional probability of the auxiliary labels as observing feature vector x_i. We assume that Y is binary in this experiment and $\bar{p}(y|x_i)$ can be assigned by considering different strategies. For example, in section 4.2, we simply used the manual annotation (i.e., broadcast news video story boundaries); whereas, in section 5.3.2, different approximation strategies are tested. Note that Y can extend to multinomial cases in other applications.

From our observations, $\bar{p}(y|x_i)$ is usually sparse. Eq. 3.2 approximates the joint probability $p(x, y)$ by taking into account the labels of the observed features but weighted and smoothed with the Gaussian kernel, which measures the non-linear kernel distance from the feature x to each observation x_i. Intuitively, nearby features in the kernel space will contribute more to Eq. 3.2.

Gaussian kernel K_σ for D-dimensional features is defined as:

$$K_\sigma(\mathbf{x}_r - \mathbf{x}_i) = \prod_{j=1}^{D} \exp \frac{-||x_r^{(j)} - x_i^{(j)}||}{\sigma_j}, \tag{3.3}$$

where $\sigma = [\sigma_1, .., \sigma_j, .., \sigma_D]$ is the bandwidth for kernel density estimation. We can control the bandwidth to embed prior knowledge about the adopted features; for example, we might emphasize more on color features and less on the texture features by changing the corresponding σ_j.

3.2.3 Sequential IB Clustering

We adopt the sequential IB (sIB) [59] clustering algorithm to find optimal clusters under the IB principle. It is observed that sIB converges faster and is less sensitive to local optima when compared to other IB clustering approaches [59].

The algorithm starts from an initial partition C of the objects in X. The cluster cardinality $|C|$ and the joint probability $p(x, y)$ are required in advance. We will discuss the selection of $|C|$ in section 3.2.5. At each step of the algorithm, one object $x \in X$ is drawn out of its current cluster $c(x)$ into a new singleton cluster. Using a greedy merging criterion, x is assigned or merged into c^* so that $c^* = \text{argmin}_c\, d_F(\{x\}, c)$. The merging cost, the information loss due to merging of the

two clusters, represented as $d_F(c_i, c_j)$, is defined as (cf. [60] for more details):

$$d_F(c_i, c_j) = (p(c_i) + p(c_j)) \cdot D_{JS}[p(y|c_i), p(y|c_j)], \tag{3.4}$$

where D_{JS} is Jensen-Shannon (JS) divergence and $p(c_i)$ and $p(c_j)$ are cluster prior probabilities. JS divergence is non-negative and equals zero if and only if both its arguments are the same and usually relates to the likelihood measure that two samples, independently drawn from two unknown distributions, are actually from the same distribution.

The sIB algorithm stops as ϵ, the ratio of new assignments among all objects X to new clusters, is less than a threshold, which means that the clustering results are "stable" and no further reassignments are needed. Decreasing the threshold will cause the algorithm to take more time and more iterations to find "stable" clusters and increase $I(C;Y)$. The impact of ϵ on time complexity and application performance is discussed in section 5.3.3.

Practically, multiple random initializations are used to run sIB multiple times and select the result that has the highest cluster MI $I(C;Y)$, namely the least information loss $I(X;Y) - I(C;Y)$.

3.2.4 Cluster Conditional Probability

During each cluster merging or splitting process, as described in Section 3.2.3, for each cluster c, we should update its cluster conditional probability $p(y|c)$, which is also an input of JS divergence of Eqn. 3.4.

$$p(y|c) = \frac{1}{p(c)} \sum_{x \in c} p(x, y), \tag{3.5}$$

where $p(x) = \sum_y p(x, y)$ and $p(c) = \sum_{x \in c} p(x)$ is the cluster prior. See more explanations in [60].

3.2.5 Number of Clusters

Exactly how we should learn the optimal number of clusters in the clustering algorithm is still an open issue. G-means [63] is one of the options but limited to low-dimensional data due to its Gaussian assumption. IB proposes a natural way to determine the number of clusters by discovering the break point of MI loss along the agglomerative IB (aIB) clustering algorithm [60, 61]. The algorithm is a hard clustering approach and starts with the trivial clustering where each cluster consists of a single item. To minimize the overall information loss, a greedy bottom-up process is applied to merge clusters that minimize the criterion in Eq. 3.4, which states the information loss after merging clusters c_i and c_j. The algorithm ends with a single cluster with all items. Along the merging steps, there is a gradual increase in information loss. We can determine the "adequate" number of the clusters by inspecting the point where a significant information loss occurs.

3.3 Visual Cue Clusters – Semantic Mid-level Represetnations

In the previous section, we described the approach to discover visual cue clusters which are essential to the semantic labels by minimizing the mutual information loss. Images in the same clusters have the same semantics, though not necessarily of the same visual appearance. However, in the generic video classification task, as shown in Fig. 3.1, we have to address two problems: (1) how to select salient visual cue clusters or reject those meaningless visual patterns – in video applications,

there are generally many non-informative or noisy images not relevant to the target application; (2) once the visual clusters are discovered, we need to invent an effective way to turn images into such representations. Conventionally, these two tasks are determined manually and require cumbersome labor in mid-level feature annotation and classifier training. By contrast, we propose a rigorous and automatic framework to address these two problems in the following subsections.

3.3.1 Feature Selection

After sIB clustering, the cluster MI between the induced feature clusters C and auxiliary label Y is measured with $I(C;Y) = \sum_c I(c)$ and can be decomposed into summation of the MI contribution of each cluster c, defined in Eq. 3.6. We further utilize this property to select the most significant clusters with the highest $I(c)$, and remove less significant or unstable clusters.

$$I(c) \equiv p(c) \sum_y p(y|c) \log \frac{p(c,y)}{p(c)p(y)} \tag{3.6}$$

3.3.2 Feature Projection

We use the discovered visual cue clusters to provide a new representation of discriminative features by transforming the raw visual features into the (soft) membership probabilities over those induced cue clusters which have different conditional probability $p(y|c)$ over the auxiliary label Y.

Each key frame with raw feature vector $\mathbf{x_r}$ is projected to the induced clusters and represented in visual cue feature $\mathbf{x_c}$, the vector of membership probabilities over

those K induced visual cue clusters;

$$\mathbf{x_c} = [x_c^1, ..., x_c^j, ..., x_c^K], \tag{3.7}$$

$$x_c^j = \hat{p}(c_j|\mathbf{x_r}) = \frac{J(c_j|\mathbf{x_r})}{\sum_{k=1}^{K} J(c_k|\mathbf{x_r})}, \text{and} \tag{3.8}$$

$$J(c_j|\mathbf{x_r}) = p(c_j) \cdot \hat{p}(\mathbf{x_r}|c_j) = p(c_j) \cdot \frac{1}{|c_j|} \sum_{\mathbf{x_i} \in c_j} K_\sigma(\mathbf{x_r} - \mathbf{x_i}). \tag{3.9}$$

$J(c_j|\mathbf{x_r})$ is proportional to the (soft) posterior probability $\hat{p}(c_j|\mathbf{x_r})$ depicting the possibility that the raw feature $\mathbf{x_r}$ belongs to cluster c_j, hence, can be represented by the product of the cluster prior $p(c_j)$ and the cluster likelihood $\hat{p}(\mathbf{x_r}|c_j)$; the latter is also estimated with KDE based on the visual features within the cluster c_j. The visual cue features $\mathbf{x_c}$ is later used as the input feature for discriminative classification. With this feature projection, we represent the raw feature $\mathbf{x_r}$ with the membership probabilities towards those visual cue clusters. Each cluster has its own semantics defined by the auxiliary label Y since all the visual features clustered into the same cluster have similar condition probability over Y.

3.4 Summary

We have proposed an information-theoretic IB framework, based on the Information Bottleneck principle, to associate continuous high-dimensional visual features with discrete target labels. Such a framework entails several promising applications and eases several problems in the large-scale video database. We can then utilize the IB framework to provide new representation for discriminative classification, feature selection, and prune "non-informative" visual feature clusters. In chapter 4, we will show that the proposed techniques actually general and effective and achieve close to the best performance when applied to TRECVID story segmentation tasks. Most

importantly, the framework avoids the manual procedures to select features and greatly reduces the amount of annotation in the training data. Such problems are commonly seen in video classification problems based on mid-level representations. In chapter 5, we will use the same framework to discover semantic clusters of the same (estimated) search relevance and utilize them to conduct (image) video search reranking.

Chapter 4

Application of Semantic Cluster – Automatic Feature Discovery in Story Segmentation

Recent research in video analysis has shown a promising direction, in which mid-level features (e.g., people, anchor, indoor) are abstracted from low-level features (e.g., color, texture, motion, etc.) and used for discriminative classification of semantic labels. However, in most systems, such mid-level features are selected manually. We proposed a new framework to automate the mid-level feature construction and discovery in chapter 3, which can be used to map any video into probabilistic memberships over the discovered cue clusters. The biggest advantage of the proposed approach is to remove the dependence on the manual effort involved in choosing mid-level features and the huge labor cost involved in annotating a corpus for training the detector of each mid-level feature. The proposed framework is general and effective, with a lot of potential for solving other problems in semantic video analysis. When tested on news video story segmentation, the proposed approach demonstrates promising performance gain over methods based on representations derived from conventional clustering techniques and even manually selected mid-level features.

4.1 Introduction

News story segmentation is an important underlying technology for information exploitation in news video, which is a major information source in the modern era. In chapter 2, we demonstrated statistical approaches for fusing diverse multimodal features in terms of feature selection (induction) and classification among sets of mid-level perceptual features. Though promising, these mid-level perceptual features are typically manually selected relying on expert knowledge of the application domain.

News channels are diverse and usually have different visual production styles, across channels and over time, which are statistically relevant to story boundaries. For years, researchers have tried different ways to manually enumerate all the possible production styles by inspection and then train the specific classifiers to classify them. For example, in [57], 17 domain-specific detectors are trained as mid-level features (e.g., *Intro/Highlight, Live reporting, Text-scene, Special, Finance, etc.*). The work requires intensive annotation and classifier training, which are usually limited to a single channel. The same processes are required to expand to new video sources. Another work [64] uses specific anchor, commercial detectors, and clustering algorithms to determine classes of "visual delimiters", where human intervention is required. Overall, researchers try to derive mid-level representations, *which deliver semantic meanings or are statistically relevant to the target event – story boundary*, and then apply a discriminative or Bayesian approach to classify the story boundaries.

These prior approaches are intuitive but not feasible for extension to multiple channels. For example, if we hope to deploy the approach worldwide on 100 channels, which on average require 5 mid-level classifiers, the cost is overwhelming since we

have to completely annotate and train these mid-level classifiers about 500 times. Even worse, human inspection is required to identify the mid-level representations. Some empirical estimations for such annotation efforts are also mentioned in section 3.1.

To address these problems in large-scale video databases, we automated the selection process of the mid-level features given defined semantic class labels in chapter 3. Given a collection of documents, where each document consists of low-level features and associated semantic labels, we want to discover the mid-level features automatically. There is still a need for labeling the semantic label of each data sample, but the large cost associated with annotating the training corpus for each manually chosen mid-level feature is no longer necessary. In addition, the dimensionality of the mid-level features will be much lower than that of the low-level features.

In this chapter, we propose new algorithms to extend the IB framework to the visual domain, specifically video. Starting with the raw features such as color, texture, and motion of each shot, our goal is to discover the cue clusters that have the highest mutual information about the final class labels, such as video story boundary or semantic concepts. Our work addresses several unique challenges. First, the raw visual features are continuous (unlike the word counts in the text domain) and of high dimensionality. We propose a method to approximate the joint probability of features and labels using kernel density estimation. Second, we propose an efficient sequential method to construct the optimal clusters and a merging method to determine the adequate number of clusters. Finally, we develop a rigorous analytic framework to project new video data to the visual cue clusters. The probabilities of such projections over the cue clusters are then used for the final discriminative classification of the semantic labels.

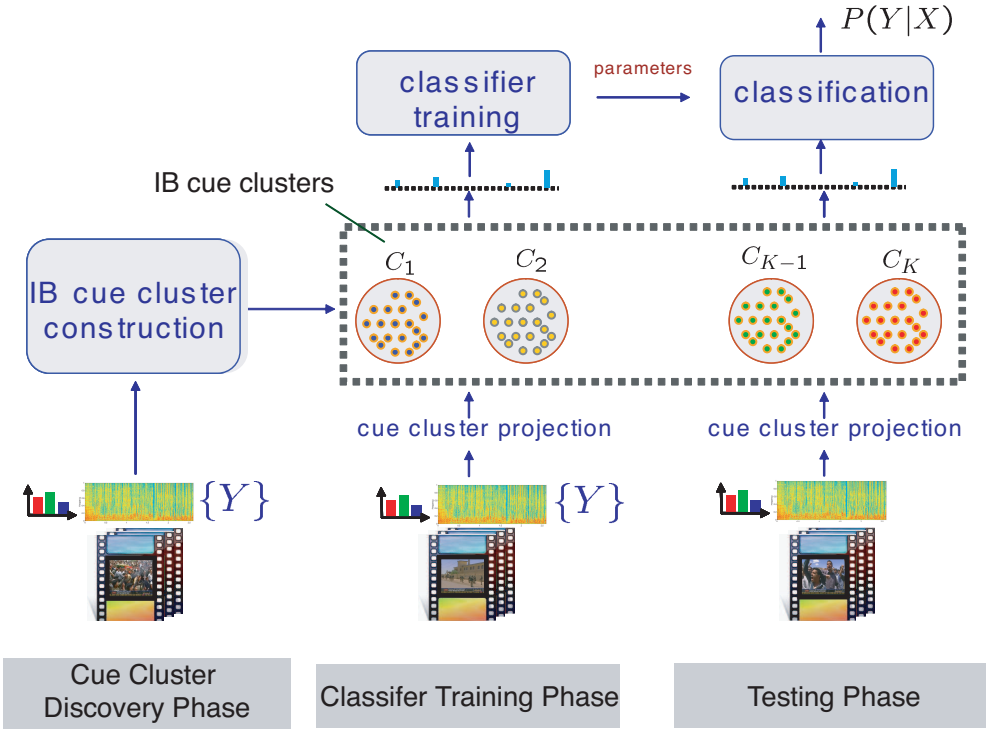


Figure 4.1: The illustration of generic video classification task based on automatically discovered visual cue clusters. It includes three major phases, "cue cluster discovery," "classifier training," and "testing."

We test the proposed framework and methods in story segmentation of news video using the corpus from TRECVID 2004 [65]. The results demonstrate that when combined with SVM, projecting videos to probabilistic memberships over the visual cue clusters is more effective than other representations such as K-means or even the manually selected mid-level features. An earlier un-optimized implementation was submitted to TRECVID 2004 story segmentation evaluation and achieved a performance very close to the top.

Based on the IB framework proposed in chapter 3, we apply a discriminative model and the feature selection process on the induced visual cue clusters in Section 4.2. In section 4.3, evaluation of the proposed techniques in news video story segmentation is described. We present conclusions and summary in section 4.4.

4.2 Discriminative Model

To utilize the automatically discovered visual clusters for generic video classification, we illustrate the framework in Figure 4.1. The framework includes three major steps. (1) *Cue cluster discovery phase*: Visual cue clusters are first discovered by the approaches mentioned in chapter 3 with the story boundary annotations and low-level visual features along with the videos. All the videos are then projected to the dimensions of visual cue clusters either in the training or test phase with the method introduced in section 3.3.2. (2) *Classifier training phase*: Classifier parameters are optimized in this phase based on the visual cue feature representations and provided story boundary labels. (3) *Testing phase*: With the same cue feature representations and optimized classifier parameters, the classification task is applied on the test data.

In this work, we adopted the discriminative approach, Support Vector Machines (SVM) [66], to classify the automatic mid-level representations – dimensions of visual cue clusters, since we had demonstrated that SVM performs the best among perceptual feature fusion in [67].

4.2.1 Support Vector Machines

SVM has been shown to be a powerful technique for discriminative learning [66]. It focuses on structural risk minimization by maximizing the decision margin. We applied SVM using the Radial Basis Function (RBF) as the kernel, $K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2), \gamma > 0$.

In the training process, it is crucial to find the right parameters C (tradeoff on non-separable samples) and γ in RBF. We apply five fold cross validation with a grid search by varying (C, γ) on the training set to find the best parameters achieving the highest accuracy.

4.3 Experiments

4.3.1 Broadcast News Story Segmentation

We tested the proposed IB framework on the story segmentation task in TRECVID [65]. A news story is defined as a segment of news broadcast with a coherent news focus which contains at least two independent declarative clauses. Story boundary detection is an interesting and challenging problem since there are no simple fixed rules of productions or features [67].

To solve this problem, researchers try different ways to manually enumerate the important production cues, and then train the specific classifiers to classify them. For example, in [57], 17 domain-specific detectors are manually selected and trained. In [37], a large set of manually picked features are fused using statistical methods like maximum entropy.

4.3.2 Approach: Discriminative Classification

We train a SVM classifier to classify a candidate point as a story boundary or non-boundary. The major features for the discriminative model is the visual cue features represented in the membership probabilities (Section 3.3.2) towards the induced visual cue clusters. Applying the IB framework, the continuous random variable X now represents the concatenated raw visual features of 144-dimensional color autocorrelogram, 9-dimensional color moments, and 48-dimensional Gabor textures for each keyframe (See explanations in [65]). The label Y is binary, "story" and "non-story."

The number of visual cue clusters is determined by observing the break point of accumulated MI loss as described in section 3.2.5 and is 60 both for ABC and CNN videos. To induce the visual cue clusters, 15 videos for each channel are used; 30

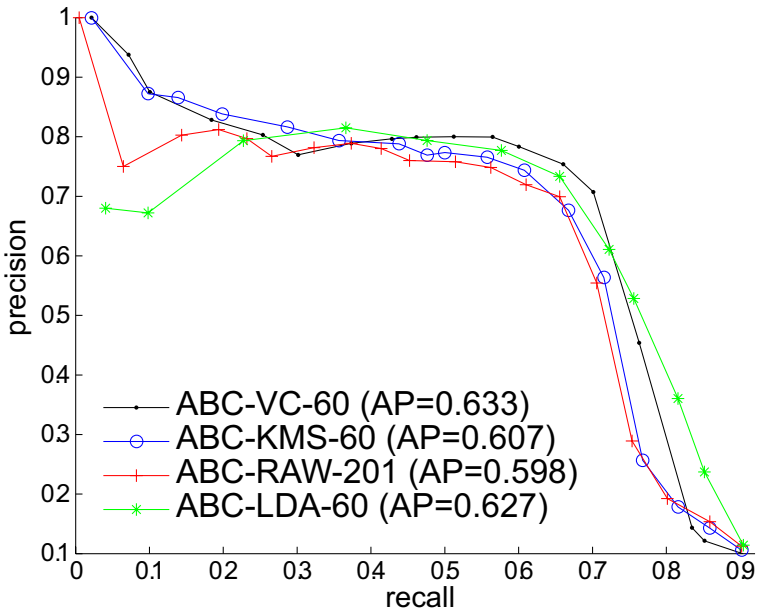


Figure 4.2: PR curves of story boundary detection on ABC videos with feature configurations via IB framework (ABC-VC-60), K-means (ABC-KMS-60), raw visual features (ABC-RAW-201), and LDA (ABC-LDA-60).

videos, with keyframes represented in the cue cluster features, for each channel are reserved for SVM training; the validation set is composed of 22 CNN and 22 ABC videos. They are all from TRECVID 2004 development set.

4.3.3 Results and Discussions

We present the boundary detection performance in terms of the precision and recall (PR) curve and Average Precision (AP) which averages the (interpolated) precisions at certain recalls. For a $M+1$ point AP, $AP = \frac{1}{M+1}\sum_{i=0}^{M} P(r_i)$; $r_i = i/M$ indicates the designated recall sample; $P(r_i) = \max_{r_i \le r} P(r)$ is the interpolated precision, where $\{r, P(r)\}$ are the available recall-precision pairs from the classification results. Intuitively, AP can characterize the PR curve in a scalar. A better classifier, with a PR curve staying upper-right corner of the PR plane, will have higher AP, and vice versa. In this experiment, we set $M = 20$.

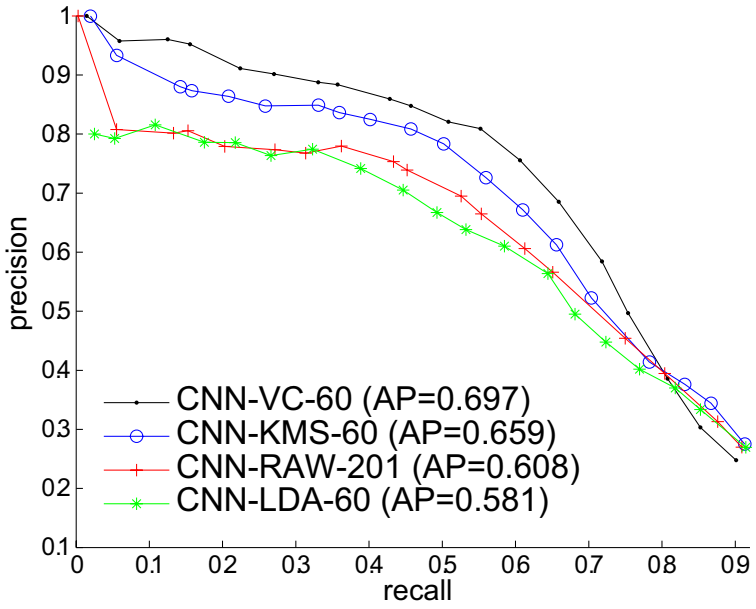


Figure 4.3: PR curves of story boundary detection on CNN videos with feature configurations via IB framework (CNN-VC-60), K-means (CNN-KMS-60), raw visual features (CNN-RAW-201), and LDA (CNN-LDA-60).

Figures 4.2 and 4.3 show the discriminative classification of story boundaries on ABC and CNN videos in PR curves and APs. All boundary detection methods use SVM but on different feature configurations. The IB approach on both video sets (ABC/CNN-VC-60) performs better that those with raw visual features (ABC/CNN-RAW-201). The performance gap is significant in CNN due to the diversity of the channel's production effects. The semantic representations from mid-level cue clusters benefit the boundary classification results.

With IB visual cue clusters, we transform 201-dimensional raw visual features into 60-dimensional semantic representations. To show the effectiveness of this approach, we compare that with the feature reduction via Linear Discriminative Analysis (LDA), which usually refers to a discriminative feature transform that is optimal when the data within each class is Gaussian [68]. LDA features (ABC-LDA-60) and performs almost the same with IB approach in ABC and even better than those

raw visual features, but not in CNN videos. It is understandable because diversity of CNN breaks the Gaussian assumption of LDA.

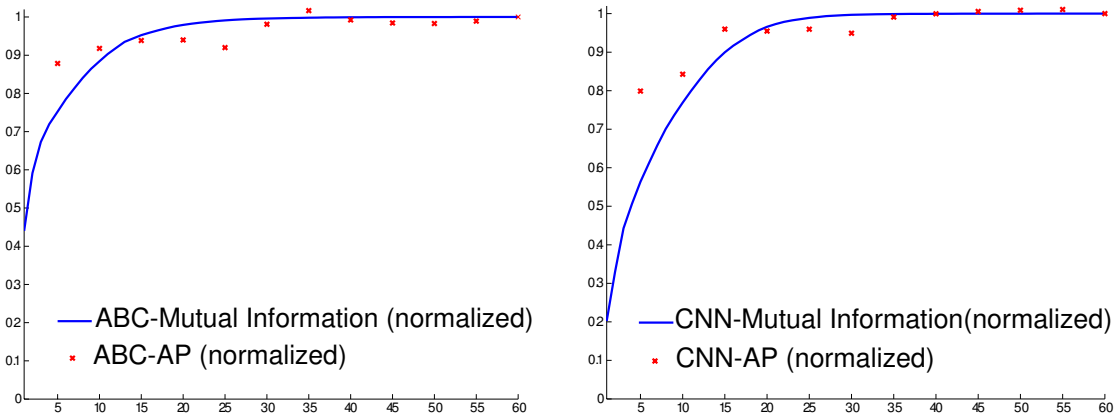
Comparing with the K-means[1] approach (ABC/CNN-KMS-60), which clusters features considering only Euclidean distance in the feature space, the IB discriminative features (ABC/CNN-VC3-60) perform better in both channels. The reason is that IB clustering takes into account the auxiliary (target) label rather than by feature similarity only, which is what K-means is restricted to.

Even with the same cluster number, the cluster MI $I(C;Y)$ through IB approach is larger than that through K-means; e.g., $I(C;Y)$ is 0.0212 for IB approach and 0.0193 for K-means in ABC, and 0.0108 and 0.0084 respectively in CNN. The difference between K-means and IB approach MI in CNN videos is more significant than that in ABC videos. It might explain why CNN IB approach has more performance gain over the K-means approach. In ABC, since positive data mostly form compact clusters in the feature space (e.g., boundaries are highly correlated with anchors, etc.), the IB framework does not differ a lot from other approaches.

4.3.4 Feature Selection

In feature selection among those induced visual cue clusters, the accumulated MI between the top N visual cue clusters (x-axis), $\sum_{i=1}^{N} I(c_i)$, and detection AP, are shown in Fig. 4.4. The MI curves and classification performance (AP) are all normalized by dividing the corresponding values with all (60) cue clusters. The results show that preserved MI of the selected cue clusters is a good indicator of the classification performance. It also allows us to determine the required number of clusters by applying a lower threshold to the cumulative MI. As seen in Fig. 4.4(b),

[1]For fair comparison, "soft" membership probability of Eq. 3.8 is used to derive features towards those K-means clusters and significantly outperforms the common "hard" membership.



(a) (b)

Figure 4.4: Relation of preserved MI and AP of top N visual clusters; (a) normalized AP vs. MI of the top N selected visual cue clusters in ABC. (b) AP vs. MI in CNN.

CNN videos need more cue clusters to reach the same AP.

4.3.5 IB Framework vs. Prior Work

Evaluated on the same CNN validation set, the IB approach described in this paper, with automatically induced visual features **only**, has AP=0.697. When augmented with speech prosody features, the performance improves to 0.805 AP and outperforms our previous work [67], which fuses detectors of anchors, commercials, and prosody-related features through SVM (AP=0.740) on the same data set. More discussions regarding multi-modality fusion and their performance breakdowns in different (visual) story types can be seen in [65].

4.4 Summary

We have proposed an information-theoretic IB framework, based on the Information Bottleneck principle for associating continuous high-dimensional visual fea-

tures with discrete target labels. We utilize the IB framework to provide a new representation for discriminative classification and feature selection, and to prune "non-informative" visual feature clusters. The proposed techniques are general and effective, achieving close to the best performance in TRECVID 2004 story segmentation. Most importantly, the framework avoids the manual labor needed to select features and greatly reduces the amount of annotation in the training data.

Chapter 5

Application of Semantic Cluster – Video Search Reranking

In this chapter, we propose a novel and generic video/image reranking algorithm, *IB reranking*, which reorders results from text-only searches by discovering the salient visual patterns of relevant and irrelevant shots from the approximate relevance provided by text results. The IB reranking method, based on a rigorous Information Bottleneck (IB) principle, as discussed in chapter 3, finds the optimal clustering of images that preserves the maximal mutual information between the search relevance and the high-dimensional low-level visual features of the images in the text search results. Evaluating the approach on the TRECVID 2003-2005 data sets shows significant improvement upon the text search baseline, with relative increases in average performance of up to 23%. The method requires no image search examples from the user, but is competitive with other state-of-the-art example-based approaches. The method is also highly generic and performs comparably with sophisticated models which are highly tuned for specific classes of queries, such as named-persons. Our experimental analysis has also confirmed the proposed reranking method works well when there exist sufficient recurrent visual patterns in the search results, as often

the case in multi-source news videos.

5.1 Introduction

Video and image retrieval has been an active and challenging research area thanks to the continuing growth of online video data, personal video recordings, digital photos, and 24-hour broadcast news. In order to successfully manage and use such enormous multimedia resources, users need to be able to conduct semantic searches over the multimodal corpora either by issuing text keyword queries or providing example video clips and images (or some combination of the two). Current successful semantic video search approaches usually build upon the text search against text associated with the video content, such as speech transcripts, close captions, and video OCR text. The additional use of other available modalities such as image content, audio, face detection, and high-level concept detection has been shown to improve upon the text-based video search systems [9, 12, 14, 15]. However, such multimodal systems tend to get the most improvement through leveraging multiple query example images, applying specific semantic concept detectors, or by developing highly-tuned retrieval models for specific types of queries, such as using face detection and speaker recognition for the retrieval of named persons. In the end, though, it will be quite difficult for the users of multimodal search systems to acquire example images for their queries. Retrieval by matching semantic concepts, though promising, strongly depends on the availability of robust detectors and required training data. Likewise, it will be difficult for the developers of the system to develop highly-tuned models for every class of query and apply the system to new domains or data sets. It is clear, then, that we need to develop and explore approaches for leveraging the available multimodal cues in the search set without complicating

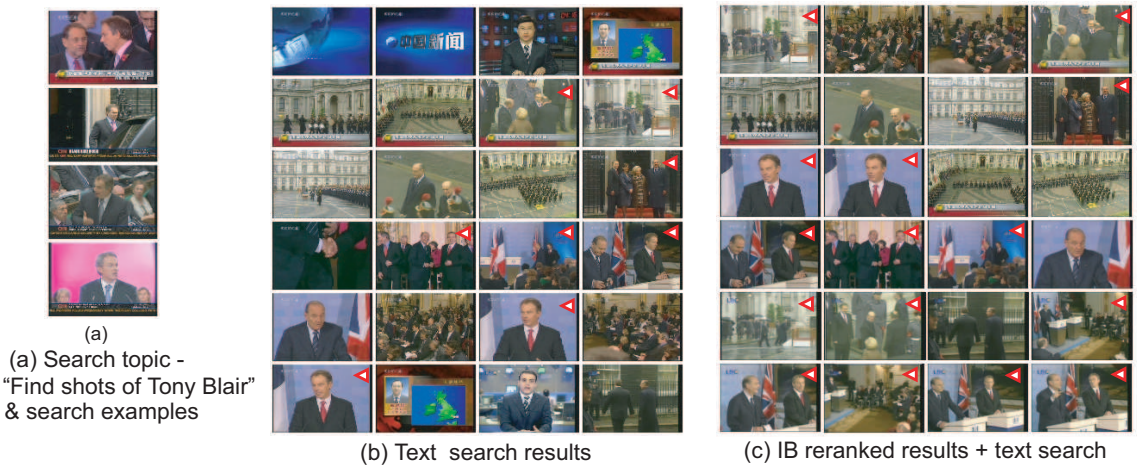


Figure 5.1: (a) TRECVID 2005 search topic 153, "Find shots of Tony Blair." (b) Top 24 returned shots from story-level text search (0.358 AP) with query terms "tony blair." (c) Top 24 shots of IB reranked results (0.472 AP) with low-level color and texture features. The red triangles mark the true positives.

things too much for the system users or developers.

Pseudo-relevance feedback (PRF) [14, 33, 34], is one such tool which has been shown to improve upon simple text search results in both text and video retrieval. PRF is initially introduced in [35], where the top-ranking documents are used to rerank the retrieved documents assuming that a significant fraction of top-ranked documents will be relevant. This is in contrast to relevance feedback where users explicitly provide feedback by labeling the top results as positive or negative. The same concept has been implemented in video retrieval. In [14], authors used the textual information in the top-ranking shots to obtain additional keywords to perform retrieval and rerank the baseline shot lists. The experiment was shown to improve MAP[1] from 0.120 to 0.124 (3.3% improvement) in the TRECVID 2004 video search task [3]. In [33], authors sampled the pseudo-negative images from the lowest rank of the initial query results; taking the query videos and images as the positive exam-

[1]MAP: mean average precision, a search performance metric used in TRECVID. See more details in Section 5.5.1.

ples, the retrieval is then formulated as a classification problem which improves the search performance from MAP 0.105 to 0.112 (7.5% improvement) in TRECVID 2003. In [15], the authors made the assumption that very few images in the data set are actually relevant to the query and sampled pseudo-negative examples randomly from the data set. The pseudo-negative examples were used with the provided positive query examples to train multiple discriminative classifiers. Besides, authors of [69] used the Google image search return set as (pseudo-)positives and utilized a parts-based approach to learn the object model and then used that to rerank the search baseline images. The object model was selected, through a scoring function, among a large number ($\sim$100) of hypothesized parts-based models, which are very time consuming. Furthermore, their experiment was limited to image queries of simple objects such as bottles, cars, etc., instead of natural language queries as those in TRECVID.

The deficiency of current PRF approaches, however, is that the "visual" pseudo-positives are not utilized. Until now, only the textual information from the top-ranking or the visual pseudo-negatives from the lowest rank are considered. The reason is due to the poor accuracy of current video retrieval systems and the scarceness of true positive images in the top-ranking results (See examples in Figure 5.1). Because of this problem, existing systems avoid choosing the top-ranking video shots as pseudo-positives and rely on external visual search examples only [33]. However, the availability of video or image examples is another problem. For example, image search on Yahoo or Google allows textual queries only: users can not upload visual examples. In fact, many users would be reluctant or unable to provide such examples, anyway, as they can be difficult and time-consuming to find. When image examples are actually provided, they are usually quite sparse. The examples are few in number and fail to capture the relevant regions of high-dimensional feature

space that might contain the true positive video shots. For example in Fig. 5.1, the query examples are not visually similar to many of the true positives and in some cases even some look like false positives.

In this work, to ease the problems of example-based approaches and avoid highly-tuned specific models, our goal is to utilize both the pseudo-positive and pseudo-negative examples and learn the recurrent relevant visual patterns from the estimated "soft" pseudo-labels. Instead of using "hard" pseudo-labels, the probabilistic relevance score of each shot is smoothed over the entire raw search results through kernel density estimation (KDE) [62]. An information-theoretic approach is then used to cluster visual patterns of similar semantics. We then reorder the clusters by the cluster relevance and then the images within the same cluster are ranked by feature density.

The semantic clustering approach is based on Information Bottleneck (IB) principle, shown to achieve significant performance gain in text clustering and categorization [59, 60]. The idea, as applied to text clustering, has been to use the information-theoretic optimization methodology to discover "cue word clusters," words of the same semantics, which can be used to represent each document at a mid level. The clusters are optimal in preserving the maximal mutual information (MI) between the clusters and the class labels.

Based on the automatically discovered semantic clusters and their information-theoretic property introduced in chapter 3, we propose a novel reranking method to find the best smoothed clusters which preserve the highest MI between (high-dimensional continuous) visual features and (hidden) search relevance labels. Meanwhile, we investigate multiple strategies to estimate the probabilistic relevance scores from initial search results. To balance reranking performance and efficiency, we experiment with different parameters used in IB ranking.

We tested the approach on the automatic video search tasks of TRECVID 2003-2005 and demonstrated its robust capability in boosting the baseline text search. Even without video/image search examples, the relative improvements, in terms of MAP, from the text baseline results are 20.8% in 2003, 17.7% in 2004, and 23.0% in 2005. In the TRECVID 2005 automatic search task, our text search plus IB reranking approach boosted the baseline story-level text retrieval to 0.107 MAP. This is a significant result, with comparable average performance to the state of the art using external example images (MAP 0.106, [15, 34]). By analyzing the experiment results over TRECVID 2005 data, we observed that the proposed IB reranking methods worked best for named people query topics. This is intuitive, as there are usually recurrent videos of named subjects in the news across multiple broadcast channels. The IB reranking method takes advantage of such patterns in improving the rankings. With the same rationale, the method expectably suffers performance loss for a small number of topics when such repeated patterns are lacking. But in general the average performance over all search topics has shown significant improvement. Furthermore, by applying the class dependent query method [70], we may apply the proposed reranking method adaptively to the query class (e.g., named people) that are predicted to benefit most from reranking.

Moreover, IB reranking is highly generic and requires no training in advance but is comparable to the top automatic or manual video search systems, many of which are highly tuned to handle named-person queries with sophisticated models such as face detection and speaker identification (cf. Section 5.5.5).

We provide the rationale for the approach in Section 5.2. Based on the IB framework in Section 3.2, the IB reranking algorithm is proposed in Section 5.3. We describe the feature representations and text search in Section 5.4. Evaluation of the proposed techniques on TRECVID video search tasks is detailed in Section

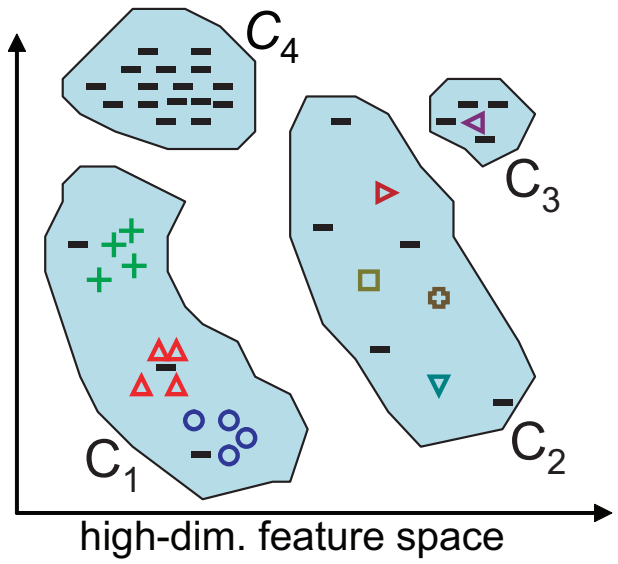


Figure 5.2: An example of 4 search relevance-consistent clusters C_1 to C_4, where C_1 has the highest denoised posterior probability $p(y = 1|c)$ and C_4 has the lowest. "–" is a pseudo-negative and the others are pseudo-positives. Pseudo-positives in the same shape are assumed having similar appearances. Note that the "hard" pseudo-labels are for illustration only; instead we use "soft" labels in this work.

5.5. We present summary and conclusions in Section 5.6.

5.2 Motivation: Video Reranking

In video search systems, we are given a query or a statement of information need and we then need to estimate the relevance $R(x)$ of each video of the shots in the search set, $x \in X$, and order them by their relevance score. Many approaches have been tested in recent years, ranging from plainly associating video shots with text search scores to sophisticated fusion of multiple modalities. Some approaches rely on user-provided query images as positive examples to train a supervised classifier to approximate the posterior probability $P(Y|X)$, where Y is a random variable representing search relevance [34]. The posterior probability is then used for $R(x)$ in video ranking.

There are certain problems and limitations for example-based video search approaches as discussed in Section 5.1. To overcome these problems, we observe that text search results can generally bring up certain relevant videos near the top of the return set (i.e., Fig. 5.1-(b)). In a large image or video database, in particular, some positive images may share great visual similarity, but receive quite different initial text search scores. For example, Fig. 5.1-(b) are the top 24 shots from story-level text search for TRECVID 2005 topic 153, "Find shots of Tony Blair." We can see recurrent relevant shots, of various appearances, dispersed among the return set but mixed with some irrelevant ones (e.g., anchor or graphic shots).

Such recurrent images or videos are commonly observed in image search engines (e.g., Yahoo or Google) and photo sharing sites (e.g., Flickr). Interestingly, authors of [32] had quantitatively analyzed the frequency of such recurrent patterns (in terms of visual duplicates) for cross-language topic tracking – a large percentage of international news videos share re-used video clips or near duplicates. Such visual patterns were used in [32] for tracking topics and will be the basis for search result reranking in this chapter.

According to the above observation, instead of using user-provided search examples, we argue to consider the search posterior probability, estimated from initial text-based results in an unsupervised fashion, and the recurrent patterns (or feature density) among the image features and use them to rerank the search results. A straightforward implementation of the idea is to fuse both measures, search posterior probability and feature density, in a linear fashion. This fusion approach is commonly used in multimodal video search systems [71]. It can be formulated as follows:

$$R(x) = \alpha p(y|x) + \beta p(x), \tag{5.1}$$

where $p(y|x)$ is the search posterior probability and $p(x)$ is the feature density of the retrieved images[2] and α and β are scalars for linear fusion.

The above equation incurs two main problems, which are confirmed in our experiments (cf. Section 5.5.2 or Table 5.1). The posterior probability $p(y|x)$, estimated from the text query results and (soft) pseudo-labeling strategies, is noisy; a "denoised" representation for the posterior probability is required. Furthermore, the feature density estimation $p(x)$ in Eqn. 5.1 may be problematic since there are usually frequent recurrent patterns that are irrelevant to the search (e.g., anchors, commercials, crowd scenes, etc.). Instead, we should consider only those recurrent patterns within buckets (or clusters) of higher relevance.

To exploit both search relevance and recurrent patterns, we propose to represent the search relevance score $R(x)$ as the following:

$$R(x) = \alpha p(y|c) + \beta p(x|c), \tag{5.2}$$

where $p(y|c)$ is a "denoised" posterior probability smoothed over a relevance-consistent cluster c, which covers image x, and $p(x|c)$ is the local feature density estimated at feature x. The cluster denoising process has been shown effective in text search [72]. Meanwhile, the local feature density $p(x|c)$ is used to favor images that occur multiple times with high visual similarity. Choices of parameters α and β will affect the reranked results. In the preliminary experiments, we observed that the denoised posterior probability $p(y|c)$ is more effective and plays the main role for search relevance when compared to the pattern recurrence within the same relevant clusters. Accordingly, an intuitive approach is to let α be significantly larger than β so that the reranking process first orders clusters at a coarse level and then refines the order

[2] In this work, visual features are extract from key-frames of each video shot.

of images in each cluster according to local feature density. The effectiveness of such an approach will be verified in the experiment section.

Two main issues arise in the above proposed approach: (1) how to find the relevance-consistent clusters, in an unsupervised fashion, from noisy text search results and high-dimensional visual features; (2) how to utilize the recurrent patterns across video sources. To address the first problem, we adopt the IB clustering approach mentioned in chapter 3, which finds the optimal clustering of the images that preserves the maximal mutual information about the search relevance. The denoised posterior probabilities $p(y|c)$ are iteratively updated during the clustering process. The feature densities $p(x|c)$ are then estimated from each cluster c accordingly.

The idea is exemplified in Fig. 5.2, where 4 relevance-consistent clusters are discovered automatically. Images of the same cluster (i.e., C_1) have the same denoised posterior probability $p(y|c)$, but might have recurrent patterns of different appearances. For example, C_1 has 3 different regions which have high density in the feature space. We first rank the image clusters by $p(y|c)$ and then order within-cluster images by the local feature density $p(x|c)$. In short, those visually consistent images which occur the most frequently within higher-relevance clusters will be given higher ranks.

5.3 IB Reranking Approach

The essence of the IB reranking approach is to smooth the noisy text search results in the high-dimensional visual feature space and to find relevance-consistent clusters, which are discovered automatically based on the IB framework described chapter 3. We will further extend the framework to video search problems.

As in any clustering problem, determining the number of clusters is non-trivial.

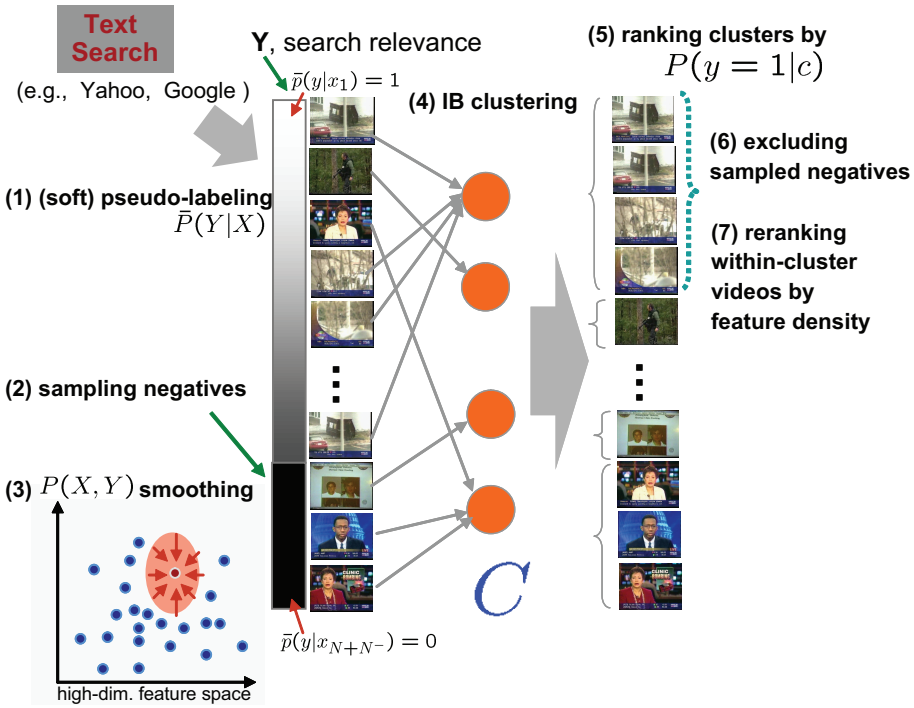


Figure 5.3: The IB reranking steps from baseline text search. See details in Section 5.3.1.

In chapter 3, we utilized the MI distortion to locate the most significant cluster number for given items. The approach is rigorous but requires extensive computation. Instead, for this experiment, we set the cluster number $K = \lceil \frac{N}{N_c} \rceil$, where $N_c = 25$ is a preset expected number of items in each cluster and is empirically determined through validation experiments. More details about determining the cluster number are in Section 3.2.5.

5.3.1 Reranking Steps

IB reranking reorders the search results derived from other search mechanisms (e.g., text search). The inputs are lists of shot or image indices J_i and their corresponding text search relevance scores s_i (if available) and can be represented as $\{(J_i, s_i)\}$. Note that the output score s_i might be unavailable in some cases such as the image return

sets from Google or Yahoo searches. For IB reranking, the hidden variable Y is the search relevance random variable we are interested in. The features $x_i \in X$ for each item J_i can be derived accordingly. The major steps of IB reranking are as follows (also illustrated in Fig. 5.3):

1. Estimate (soft) pseudo-labels and treat them as un-smoothed conditional probability $\bar{p}(y|x_i)$ (cf. Section 5.3.2) from the top N images $\{(J_i, s_i)\}$ of text search output.

2. Sample N^- negative examples from the database and set $\bar{p}(y = 1|J) = 0$ for these negative examples.

3. Calculate smoothed joint probability $p(x, y)$ through Eqn. 3.2 for these $N+N^-$ items.

4. Conduct sIB clustering (cf. Section 3.2.3) on $N + N^-$ images into K clusters with convergence threshold ϵ.

5. Rerank clusters by search relevance, cluster conditional probability $p(y|c)$ (See Eqn. 3.5), in descending order.

6. Exclude those N^- sampled negative examples.

7. Rerank images within each cluster c in descending order by feature density estimated as follows[3]:

$$p(x_j|c) = \frac{1}{Z_c} \sum_{x_k \in c, k \neq j} K_\sigma(x_k - x_j). \tag{5.3}$$

8. Output the ordered list $J_{i'}$.

[3] Z_c in Eqn. 5.3 is a normalization factor to ensure $\sum_j p(x_j|c) = 1$.

The aim of the IB reranking approach is to statistically learn the recurrent relevant patterns and reorder them according to the denoised cluster conditional probability $p(y|c)$. We first estimate the un-smoothed conditional probability $\bar{p}(y|x_i)$ from the initial text search output. Then we smooth it through the whole feature space approximated by the top N images and N^- sampled negatives. This is reasonable since the pseudo-labeling approach is just approximating the correlation between the semantic (search relevance, Y) and feature space X. Intuitively, those low-ranked images or sampled negative examples are helpful [33, 34] to push down the false positives in the top of text output. Vice versa, salient top-ranked positive patterns can pull up other relevant but low-ranked images from the preliminary text search. Such examples are seen in Fig. 5.1-(b) and (c). Even though the anchor/graphics images appear frequently in the top of the text results, they also appear equally as frequently throughout the bottom of the results or sampled negative images. The smoothing process tends to push them down the search list.

The sIB clustering further groups images of similar relevance into consistent clusters and computes its corresponding denoised cluster conditional probability $p(y|c)$, which is another smoothing process by using the cluster posterior $p(y|c)$ to replace the individual posterior $p(y|x)$.

The within-cluster images are further ordered by their local feature density $p(x|c)$ estimated from the items of the same cluster, where we assume that images with more frequently recurrent patterns are more relevant than isolated dissimilar images in the same cluster. Note that in all the reranking steps we need not have "hard" positive or negative samples, which are required in prior example-based approaches, but only their soft un-smoothed conditional probability $\bar{p}(y|x)$ available from initial text-based search.

5.3.2 Pseudo-labeling Strategies

The IB reranking method estimates the un-smoothed search relevance probability, $\bar{p}(y|x_i)$, by using the initial text search score s_i. x_i is the image feature of image J_i and $y \in Y$ is the relevance random variable. We experimented with three different strategies for such estimation of "soft" pseudo-labeling.

5.3.2.1 Binary

In the "binary" approach, we estimate the un-smoothed search relevance probability of image J_i with text search score s_i in a binary form.

$$\bar{p}(y = 1|x_i) = 1_{\{s_i \geq e_s\}},$$

where $1_{\{.\}}$ is an indication function and e_s is the search score threshold. Empirically, one can use the mean plus one standard deviation of the entire text search scores. Or one could use cross-validation experiments to set a suitable e_s value.

5.3.2.2 Normalized Rank

For certain cases when text search scores are unavailable and only ranking orders are given, we adopt the normalized rank [33] to estimate $\bar{p}(y|x_i)$ of image J_i, which is the i'th ranked image:

$$\bar{p}(y = 1|x_i) = 1 - \frac{i}{N},$$

where N is the number of return set to be reranked from the initial text search output. Naturally, the first-ranked image will be $\bar{p}(y|x) = 1$ and the last will be 0.

5.3.2.3 Score Stretching

In the "score stretching" approach, $\bar{p}(y=1|x_i)$ is estimated by setting the middle point (0.5) at the text search score threshold e_s and linearly stretching those scores above or under e_s to be within $[0,1]$.

$$\bar{p}(y=1|x_i) = \frac{1}{2} + 1_{\{s_i \geq e_s\}} \cdot \frac{s_i - e_s}{2(max_s - e_s)} - 1_{\{s_i < e_s\}} \cdot \frac{e_s - s_i}{2(e_s - min_s)},$$

where max_s and min_s is the maximum and minimum text search scores.

5.3.3 Reranking Complexity vs. Thresholds

The sIB convergence threshold ϵ affects the time complexity for sIB clustering (cf. Section 3.2.3) and the clustering quality, in terms of MI distortion, and ultimately will influence the reranking performance. A lower threshold ϵ will force the algorithm to take a longer time to converge. This usually means a more "stable" clustering result. However, we are curious about the trade-off between time complexity and reranked performance. We tested the reranking process at different thresholds ranging from 0.01 to 0.5 on both TRECVID 2003 and 2004 queries. Results of time complexity vs. reranking performance, in a normalized scale, are shown in Fig. 5.4. The experiments reranked 1250 items, 1000 (N) video shots plus 250 (N^-) sampled negative examples. Increasing the threshold sharply reduces the clustering time but only degrades the performance slightly in both TRECVID data sets. A setting of $\epsilon = 0.20$ reduces the computation time by 10 folds while keeping almost unchanged performance (MAP). It suggests that most of the relevant or irrelevant data are in the right order after just a few sIB iterations. In the following experiments, we

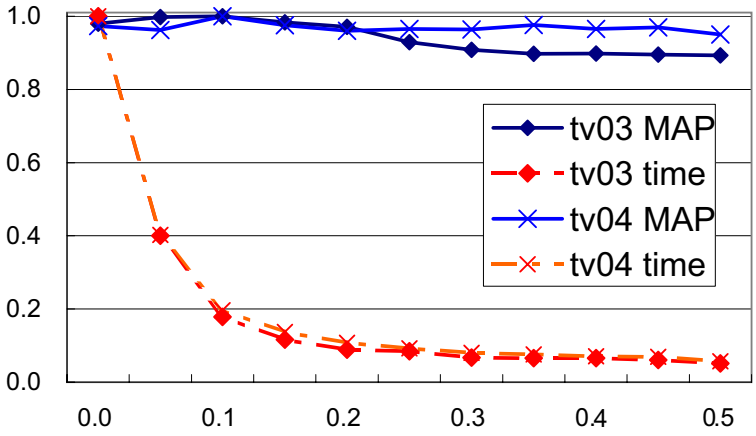


Figure 5.4: Time complexity vs. reranking performance, in a normalized scale, on TRECVID 2003 (tv03) and 2004 (tv04) data sets. Note that "score stretching" pseudo-labeling approach is used. See explanations in Section 5.3.3

fix the threshold $\epsilon = 0.20$. Implemented in MATLAB on a regular Intel Pentium server, it takes around 18 seconds to rerank a query of 1250 images.

5.4 Feature Representations

5.4.1 Low-level Features

For low-level features X, we represent each key-frame with a 273 dimensional feature vector composed of two low-level global visual features. The first is color moments over 5x5 fixed grid partitions [15], where the first 3 moments of 3 channels of CIE Luv color space are extracted; it results in a 225-dimensional feature vector per key-frame. The second is Gabor texture [15] where we take 4 scales and 6 orientations of Gabor transformation and further use their means and standard deviations to represent the whole key-frame and result in a 48-dimensional feature. Though they are basic image features, prior work such as [9, 15] has shown their excellent performance in image retrieval and semantic concept detection. To analyze the contribution of each feature, we have also evaluated the reranking performance by

using grid color moments only, which leads to a relative performance drop of 8% compared to using both color and texture features described above.

5.4.2 Text Search

IB reranking is built on top of text search against the text (e.g., ASR or machine translation) transcripts of the videos in the database. We specially choose the best possible text search approach to provide the best baseline for improvement with IB. The text searches are conducted automatically using the natural language statements of information need, provided by NIST [3]. The natural language queries (such as "Find shots of Iyad Allawi, the former prime minister of Iraq" or "Find shots of a ship or boat") are parsed through part-of-speech tagging [73] and named entity tagging [74]. Keywords (like "ayad allawi iraq" and "ship boat") are extracted. If named entities exist, then those are chosen as the keywords. If not, then the nouns are chosen. Finally, if no nouns or named entities are present, then the main verb is chosen. A standard list of stop words is also removed. The extracted keywords are then used to issue queries against the speech recognition transcripts using the Okapi BM-25 formula [75] and all terms are stemmed using Porter's algorithm.

Since the retrieval unit in video search is the shot (a single continuous camera take), there arises the problem of how to associate chunks of text with each video shot. A simple approach might be to take the text document for each shot to be the text that is contained temporally within the boundaries of the shot. Or, we could compensate for the asynchrony between the speech transcripts and the video stream by including some buffer text from a fixed window before and after the shot as well. Our experiments, however, have shown that the best approach is to use the text from the entire story within which the shot is contained. This makes sense since the true semantic relationships between images and the text transcript exist

at the story level: if a concept is mentioned in the text it is likely to appear in the video stream somewhere within the same story, but it is unlikely to appear in the next story or the previous one. Story boundaries can be extracted (imperfectly, but with reasonable reliability) automatically through the visual characteristics of the video stream and the speech behavior of the anchorperson. Different story boundary detectors are trained separately for each language – English, Chinese, and Arabic. The performance, evaluated with TRECVID metrics (F1[4]), is 0.52 in English, 0.87 in Arabic, and 0.84 in Chinese [76]. Our experiments have shown that choosing text within automatically detected story boundaries to associate with shot documents outperforms the fixed-window based approach consistently with approximately 15% improvement in terms of MAP across the TRECVID 2003, 2004, and 2005 data sets. Using manually annotated story boundaries offers an additional 5-10% increase in performance. Note that automatically detected boundaries are used in this work.

The use of entire stories for retrieval gives an increase in recall (more of the true relevant shots are found), but gives a decrease in precision (more noise also turns up). This provides an excellent motivation for the application of IB reranking, since, if text search is working well, then many of the relevant documents are found and ranked highly and we can exploit a method to mine the salient visual patterns from those shots and push down the noisy irrelevant shots out of the top-ranked spots.

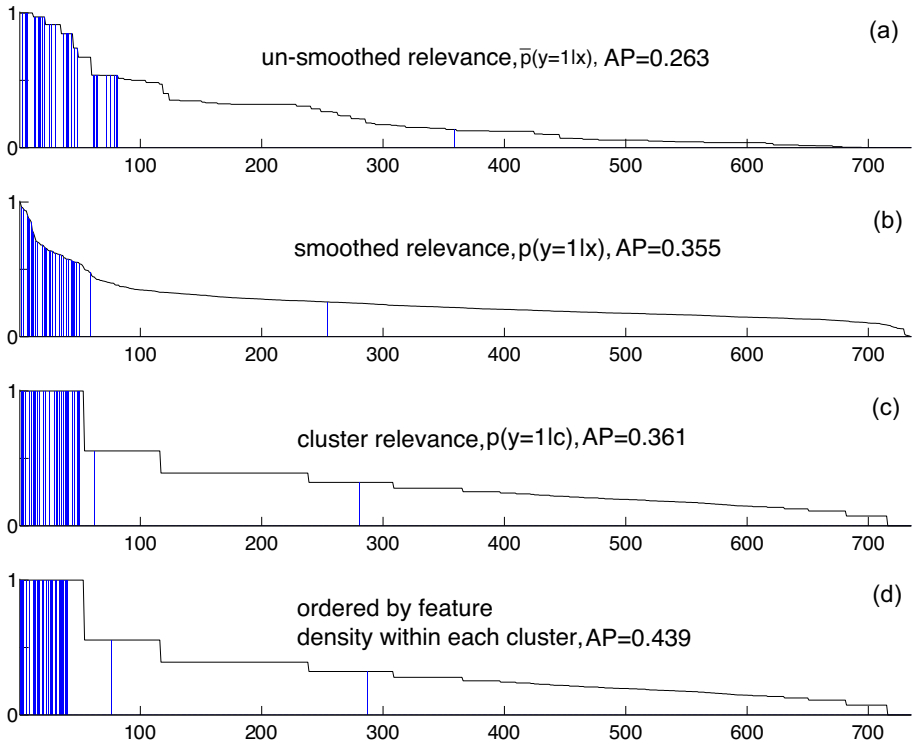


Figure 5.5: Reranking steps and their corresponding AP of topic 171, "Find shots of a goal being made in a soccer match"; (a)-(c) are normalized scores of video shots, ordered by their corresponding measures $\bar{p}(y=1|x)$, $p(y=1|x)$, and $p(y=1|c)$; In (d), shots in the cluster (with the same $p(y=1|c)$) are ordered by feature density. The blue lines mark the true positives.

5.5 Experiments

5.5.1 Data Set

We conduct the experiments on TRECVID 2003-2005 data sets [3]. In TRECVID 2003-2004, there are 133 and 177 hours of videos respectively with English ASR transcripts both from CNN and ABC channels in 1998. The TRECVID 2005 data set contains 277 international broadcast news videos which include 171 hours of videos from 6 channels in 3 languages (Arabic, English, and Chinese). The time

[4] $F1 = \frac{2 \cdot P \cdot R}{P+R}$, where P and R are precision and recall rates defined in TRECVID [3] story boundary detection task.

span is from October 30 to December 1, 2004. The ASR and machine translation (MT) transcripts are provided by NIST [3].

For the performance metric we adopted non-interpolated average precision (AP), which corresponds to the area under a (non-interpolated) recall/precision curve. Since AP only shows the performance of a single query, we use mean average precision (MAP), which is the mean of APs for multiple queries, to measure average performance over sets of different queries in a test set. See more explanations in [3].

5.5.2 Breakdowns in Reranking Steps

Before we present the average performance over all queries, let's analyze in depth an example query to understand the contributions from each step of the IB reranking process. For search topic 171, "Find shots of a goal being made into a soccer match," terms "goal soccer match" are automatically determined and used to derive the text search results. Fig. 5.5-(a) are scores of video shots ordered by the estimated unsmoothed conditional probability $\bar{p}(y|x)$ through "score stretching" pseudo-labeling strategy (cf. Section 5.3.2.3). The AP is the same as the original text search score. Fig. 5.5-(b) are those by the smoothed conditional probability $p(y|x) = \frac{p(x,y)}{p(x)}$, estimated from the top N^+ text return set and N^- sampled pseudo-negatives. The smoothing process can bring up some positives or push down negatives, and hence improve AP accordingly. This can be confirmed by the fact that most true positives (as blue lines in Fig. 5.5-(b)) are moved to higher ranks. Interestingly, through the sIB clustering approach, almost all of the recurrent relevant scenes, though of diverse appearances, are clustered in the same and the most relevant cluster, as shown in Fig. 5.5-(c), where the plateau represents the same $p(y|c)$ within the same cluster. Further ordered by feature density $p(x|c)$, the relevant shots are further pushed to the top, as shown in Fig. 5.5-(d).

#	steps/measures, $R(x)$	MAP	improvement
1	$\bar{p}(y\|x)$	.0858	0.0%
2	$p(y\|x)$	.0974	13.5%
3	$\alpha p(y\|x) + \beta p(x)$	.0930	8.3%
4	IB reranking	.1031	20.1%
5	IB reranking+text	.1050	22.4%

Table 5.1: The performance breakdowns in major reranking steps evaluated in TRECVID 2005 queries. The absolute MAPs and relative improvements from the text baseline are both shown. "IB reranking+text" means that the reranking results are fused with the original text scores via ranking order fusion method. Row 3 lists the best MAP among sets of (α, β) in the implementation of Eqn. 5.1. $\bar{p}(y|x)$ is the initial search relevance from text search scores. $p(y|x)$ is a smoothed version of $\bar{p}(y|x)$. See more explanations in Section 5.5.2.

As shown in Fig. 5.5, exceptions can also be found. Not all true positives are included in the first cluster (See 2 sparse blue lines in Fig. 5.5-(c) and (d)). Such cases may actually become worse after reranking. However, as supported by the overall performance comparisons (Table 5.1), the majority of results for most queries benefit from the reranking process. It is also important to note that the quality of top results are most important for search tasks, as intuitively captured by the definitions of AP performance metric.

The contributions from the major reranking steps averaging across all official queries of TRECVID 2005 are listed in Table 5.1. Consistent improvements by IB reranking are confirmed[5]. Besides, relevance probability smoothing has a large impact on performance (13.5% gain over initial text search). Fusion with prior (as in Eqn. 5.1) without clustering actually hurts (performance gain dropped to 8.3%). The proposed IB reranking achieves a significant gain at 20.1%, which is increased to 22.4% if IB reranking results are further fused with initial text search results.

[5]Note that the scores in Table 5.1 and Fig. 5.5 are slightly lower than those in Table 5.2 since less than 1000 shots are used for the evaluation. The reranking process is applied to 1000 sub-shots which need to be merged into shots. It is a requirement for TRECVID [3]. The number of shots is less than 1000 after the merging.

5.5.3 Performance on All TRECVID Queries

We conduct IB reranking on all queries of TRECVID 2003-2005. We first compared the three pseudo-labeling strategies on both TRECVID 2003 and 2004 data sets. As shown in Table 5.2, "score stretching" approach is the most effective and is later solely used in the TRECVID 2005 test, since it naturally utilizes the continuous relevance output from the text modality; IB reranking improves the performance (MAP) of text search baseline and up to 23%.

The performance (story text vs. story text + IB reranking) in AP across all queries is listed in Fig. 5.8, where IB reranking improves or retain the same performance of text search baseline for a large majority of queries. IB reranking benefits the most for queries with salient recurrent patterns; i.e., "Sam Donaldson" (135), "Omar Karami" (151), "Blair" (153), "Mahmoud Abbas" (154), "baseball" (102), "Spinx" (116), "Down Jones" (120), and "soccer" (171). This makes sense since the approach, though requiring no example images, tries to infer the recurrent patterns highly relevant to the search relevance based on the initial search scores and visual density estimation. Specifically, the visual patterns present in the search results will help boost the posterior probabilities of relevant data through denoising (Eqn. 3.5) and local density based reranking in each cluster (Eqn. 5.3).

Fig. 5.8 also shows several query topics where performance is degraded after IB reranking. The queries include "building-fire" (147), "Pope" (123), and "Boris Yelstin" (134). Upon further examination, we found the relevant videos for such queries are either of a small number or lack consistent visual patterns. For example, scenes of the Pope are actually of different events and thus do not form consistent visual appearances. This explains why IB reranking does not provide benefits for such queries.

Exps./Stratgies	text baseline	binary	normalized rank	score stretching
TRECVID 2003	0.169	0.187 (10.6%)	0.177 (5.0%)	0.204 (20.8%)
TRECVID 2004	0.087	0.089 (1.7%)	0.098 (12.9%)	0.102 (17.7%)
TRECVID 2005	0.087	–	–	0.107 (23.0%)

Table 5.2: IB reranking performance (top 1000 MAP) and comparison with the baseline (story) text search results of TRECVID 2003, 2004, and 2005. Each column uses a different pseudo-labeling strategy. Percentages shown in parentheses are improvement over the text baseline.

IB reranking requires no external image examples but just reranks images from the text search output directly. This is an important advancement in video search since users do not have or are reluctant to provide image examples. Surprisingly, the novel approach is competitive with and actually complementary to those state-of-the-art example-based search approaches (cf. Section 5.5.6 and 5.5.5). Nevertheless, visual examples significantly outperform text search in certain queries such as "tennis" (156), "basketball" (165), etc. This offers a promising direction for further expanding the proposed approach – as external example images are available, we may consider embedding these "true" positives in the IB framework for reranking; i.e., setting the un-smoothed conditional probability $\bar{p}(y|x) = 1$ for example images.

5.5.4 Number of Clusters

The number of clusters is an important parameter for clustering algorithms. Our prior work [76] used an information-theoretic measure to determine the optimal cluster number of the "visual cues," which are later used as bases for a new feature representation. To avoid the large computational cost, we empirically select the best cluster number threshold N_c through multiple experiment runs. We experimented with different cluster thresholds over TRECVID 2003 and 2004 data sets and choose the one that resulted in the best performance. Then for the new data

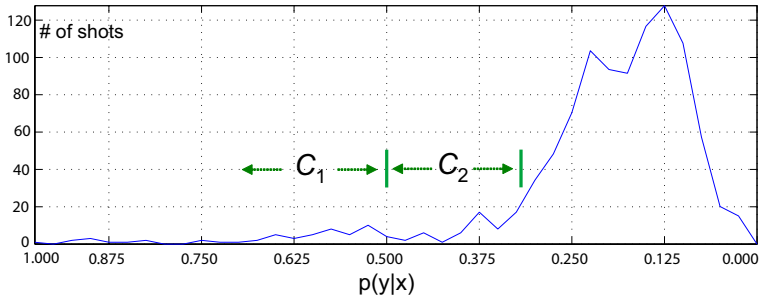


Figure 5.6: Histogram of normalized search posterior probability $p(y = 1|x)$ from top 1000 shots of topic 171. The top clusters (C_1 and C_2) correspond to natural cut points in terms of (estimated) search relevance scores.

set in TRECVID 2005, we applied the same cluster threshold without readjustment to assure the generality of the empirical choice. We have found a cluster number corresponding to an average cluster size $N_c = 25$ would provide satisfactory results.

In addition, we have found the reranking performance is not sensitive to the choice of the cluster numbers. In most queries, the most critical results (in terms of AP calculation and user satisfaction) are in the first few clusters or pages. Thus, slight changes of the number of clusters will not significantly affect the distribution of such top results.

In Fig. 5.6., we show a histogram, the number of video shots, over normalized search posterior probability $p(y = 1|x)$ (cf., Fig. 5.5-(b)) from top 1000 images of topic 171. The first two most relevant clusters C_1 and C_2 are also labeled. The results indicate that the IB clusters are effective and intuitive – the top clusters coincide with the intuitive cutoff points (0.5 and kneeling point).

5.5.5 Performance on Named-Person Queries

IB reranking works best when initial text search results contain a reasonable number of positive shots and the positive shots are somewhat visually related. Such conditions often occur when the video sources come from multiple concurrent channels

reporting related events. This is more the case for TRECVID 2005 than TRECVID 2003 and 2004.

In the TRECVID 2005 benchmark data, we have observed that IB reranking works particularly well on queries for named persons. This is surprising since the visual content of example images and shots has not previously been shown to be very helpful for retrieval of named persons. For example, fusing simple story-based text search with a high-performing content-based image retrieval (CBIR) system [15, 34] provides only modest gains over story-based text search on the six named person queries in the TRECVID 2005 set. Story-based text search results in a MAP of 0.231 over these six queries, while fusing with the CBIR system [15, 34] provides a very minor improvement in MAP to 0.241, a relative improvement of 4%. On the other hand, if we apply IB reranking after story-based text search, we get an improvement in MAP to 0.285, a big improvement of over 23%. So, IB reranking is able to capture the salient visual aspects of news events contained in the search set in which particular named people appear, which is very difficult to do with example images which come from sources other than the search set or from a different time span. When compared to the performance of all official automatic and manual submissions on the six named person queries, illustrated in Fig. 5.7, IB reranking outperforms all manual runs and is second (but comparable) to only one automatic run (MAP: .299), which is highly tuned to handle named person queries with face detection and other models requiring external resources. However, the IB approach is highly generic and requires no training (specific to the named person search) in advance.

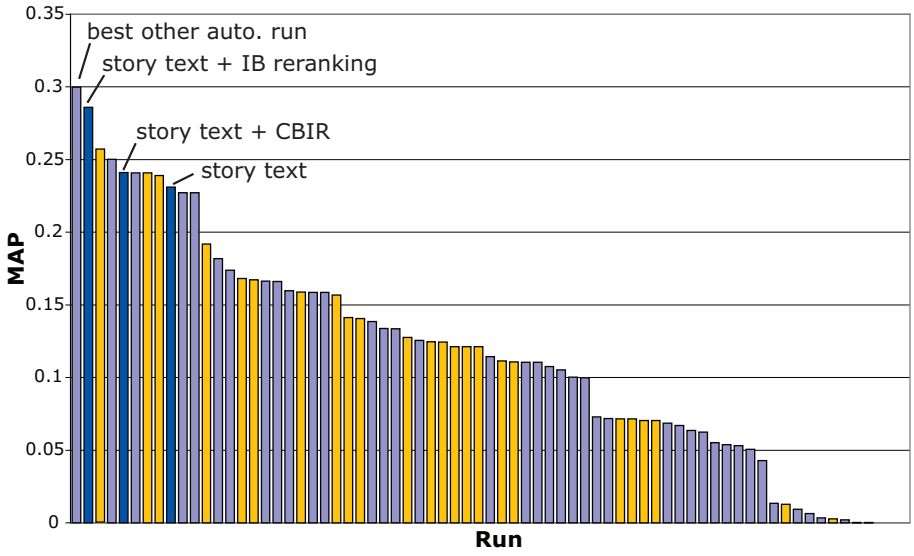


Figure 5.7: MAP on the six named-person queries of all automatic (blue) and manual (orange) runs in TRECVID 2005.

5.5.6 Class-Dependent Fusions

We have seen that the benefit of IB reranking is significant on named person queries. Similar to prior work in class-dependent queries [70], we also tested a query-class-dependent model, where different approaches can be used for different classes of queries. One simple example of such a model is the use of IB reranking when a named person is detected[6] in the query and a high-performing content-based image retrieval (CBIR) when no named entity is detected. We have experimented with this approach on the 24 queries in the TRECVID 2005 data set, using IB for the six named person queries and CBIR for the remaining 18 queries. Each approach has a MAP of about .11 across all 24 topics; however, using such a simple class-dependent fusion results in a jump in MAP to over 0.176, which outperforms the top manual submission (MAP of .168) as well as the top automatic run (MAP of .125).

[6] Automatic classification of query topics have been shown highly feasible for classes like named persons [70]

5.6 Summary

We proposed a novel reranking process for video search, which requires no image search examples and is based on a rigorous IB principle. The approach is based on the assumption that most search targets (images or videos) generally have recurrent patterns across sources. We can utilize such recurrences within each relevance-consistent cluster to rerank the text search results. Evaluated on TRECVID 2003-2005 data set, the approach boosts the text search baseline over different topics in terms of average performance by up to 23%.

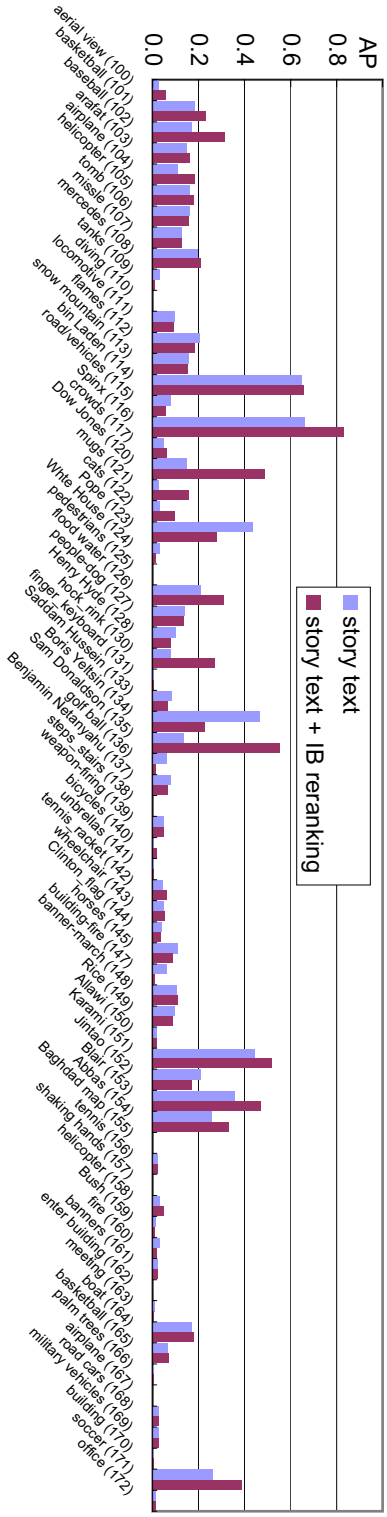


Figure 5.8: Performance of IB reranking and the baseline text search across all queries of TRECVID 2003-2005.

Chapter 6

Video Similarity Based on Multimodal Features

Videos from distributed sources (e.g., broadcasts, podcasts, blogs, etc.) have grown exponentially. Advanced video indexing and retrieval techniques such as topic threading or video search are very essential for organizing such large-volume information sources. Current solutions primarily rely on text features only and therefore encounter difficulty when text is noisy or unavailable. In this chapter, we extend beyond the text modality and propose new representations and similarity measures for news videos based on low-level visual features, visual near-duplicates, and high-level semantic concepts automatically detected from videos. We will further demonstrate novel and effective representations to exploit rich multimodal cues, beyond just text, for topic threading and video search in chapter 7 and 8 respectively.

In contrast to feature representations and similarity measures in fixed windows surrounding certain time points (cf. chapter 2) or in a shot level (cf. chapter 5), here we consider feature representations at a semantic level – a (video) story or document, which usually has variant-length videos (shots), text tokens, and detectable objects (e.g., cars, explosions, etc). We need to address the many-to-many pair-wise multimodal similarity measures in such story-level feature representations.

6.1 Introduction

Due to the explosion of Internet bandwidth and broadcast channels, video streams are easily accessible in many forms such as news video broadcasts, blogs, and podcasts. As a critical event breaks out (e.g., tsunami or hurricanes), bursts of news stories on the same topic emerge either from professional news or amateur videos. Some examples are shown in Fig. 1.1. There is a fundamental need for approaches for automatically organizing video content from distributed sources into coherent topics (by topic threading) or to enable indexing for video search or further manipulation.

Current topic threading or search approaches primarily exploit text information from speech transcripts, closed captions, or web documents. The use of multimodal information such as visual duplicates [1] or semantic visual concepts [3], though not explored before, are very helpful. There are usually recurrent visual patterns in video stories across sources that can help topic threading. For example, Fig. 6.1 has three example stories[1] from Chinese, Arabic, and English news sources, which cover the same topic. The stories from different channels share a few near-duplicates such as those showing George Bush and Tony Blair, the press conference location, and the audience. Such duplicates, confirmed by our analysis later, are actually effective for news threading across languages (cf. Section 6.2.3). Another example is shown in Fig. 1.1 where recurrent stories share a few near-duplicates (e.g., Philippine President Arroyo) or certain dimensions of high-level concepts (e.g., soldiers, deserts, demonstration, temples, etc.).

In this chapter, we investigate novel representations and similarities for pairwise stories using multimodal information, including text, visual duplicates, and

[1] More example stories at "http://www.ee.columbia.edu/∼winston"

semantic visual concepts. In chapter 7, we will further utilize a general fusion framework for combining diverse cues and analyze the performance impact by each component to foster cross-domain topic tracking. Evaluating on the TRECVID 2005 data set [3], fusion of visual duplicates improves the state-of-the-art text-based approach consistently by up to 25%. For certain topics, visual duplicates, alone, can even outperform the text-based approach. In addition, we propose an information-theoretic method for selecting subsets of semantic visual concepts that are most relevant to topic tracking. Meanwhile, we will demonstrate the benefits of utilizing the recurrent patterns at a semantic story level for video search through a random walk perspective. Evaluating over the same data set for a video search benchmark, we can get an improvement in Mean Average Precision (MAP) of up to 32%. The boost in performance especially benefits people-related queries, which usually have cross-domain recurrent stories, and show a significant 40% relative improvement through the proposed random walk approach over a multimodal story-level similarity graph.

We describe the need for story-level feature and similarity representations in Section 6.2. The novel representations from cue word clusters are in Section 6.2.1, low-level visual features in Section 6.2.2, visual duplicates in Section 6.2.3, and semantic concepts in 6.2.4. We discuss the multimodal fusion in Section 6.3, followed by the summary in section 6.4.

6.2 Story-Level Feature Representations

Most of the work in video indexing and retrieval is conducted using shots – or single continuous camera takes. However, especially in the news video domain, a semantic unit – the video story, is more natural since it is a basic element presented from the

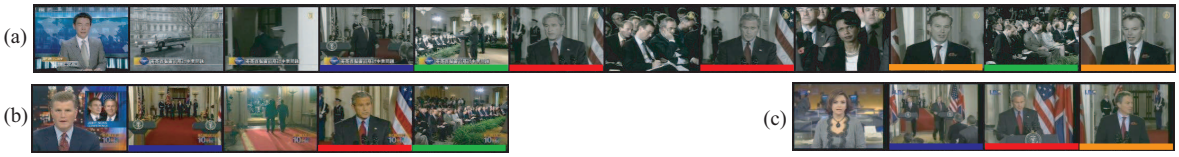


Figure 6.1: Key-frames from 3 example stories of topic "Bush and Blair meet to discuss Mideast peace;" (a) in Chinese from NTDTV channel, (b) in English from MSNBC channel, and (c) in Arabic from LBC channel. Different near-duplicate groups are indicated in different colors.

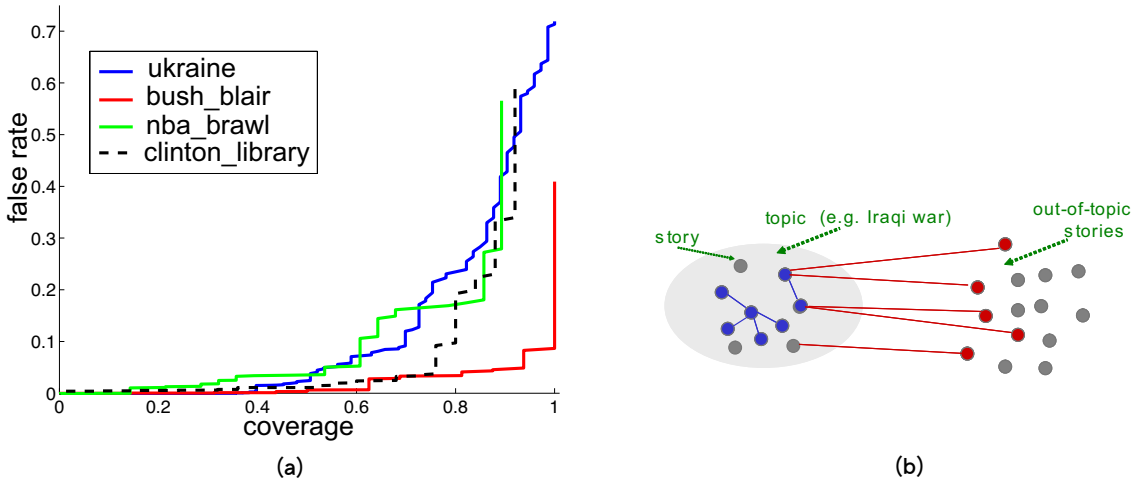


Figure 6.2: (a) The coverage and false alarm rate of visual duplicates among 4 topics by varying duplicate thresholds (cf. Section 6.2.3). (b) An illustration of coverage (blue) and false alarm rate (red) explained in Section 6.2.3.

media sources and it fits the users' behavior in video search and browsing. Moreover, in several applications, researchers have shown the significance of the story unit. For example, authors of [9] found that conducting video search in detected story boundaries associated with video shots outperforms video search with a fixed-window based approach, alone, consistently with approximately 15% improvement across the TRECVID 2003, 2004, and 2005 data sets. Similar observations were found in [77]. Furthermore, the authors of [78] associated names and faces within story boundaries, providing a natural boundary to preserve local semantic proximity and remove ambiguity against irrelevant names and faces which are common

mistakes in the fixed-window approaches.

Regarding automatic story boundary detection in the large-scale news video database, we have introduced principled and statistical approaches in chapter 2 and 5, where more state-of-the-art approaches are surveyed as well. However, to support further applications in video indexing and retrieval (i.e., topic threading and video search, etc.), two fundamental issues need to be addressed – (1) representation of each story and (2) measurement of similarity between story pairs. Note that since we extend the feature representation beyond the text modality, there are problems with how to discover salient features from video or audio streams and how to aggregate variant-length videos into a uniform representations. In this chapter, we first describe the text processing techniques to extract cue word clusters, and then present visual features at multiple levels, ranging from the low level visual features (color and texture), through parts-based near-duplicate similarity, to the high-level visual concepts. Furthermore, the application of these story-level feature and similarity representations for tracking novel topics and improving text-based video search will be discussed in chapter 7 and 8.

6.2.1 Cue Word Clusters

We represent the text modality of each story by compact "cue word clusters" or "pseudo-words" [79]. The text transcripts for each story are from the automatic speech recognition (ASR) and machine translation (MT) transcripts included in the data set (e.g., TRECVID 2005 [3]). The first text processing step involves stemming the ASR and MT tokens and removing the stop words, resulting in a total of 11562 unique words. Then a mutual information (MI) [38] approach is used to select the top 4000 informative words based on the MI between the stories and words. Specifically, given a set of stories D over the word set O, the MI between stories

and words can be computed as $I(O;D) = \sum_{o \in O} I(o)$, where $I(o)$ represents the contribution of word o to the total MI $I(O;D)$.

$$I(o) \equiv p(o) \sum_{d \in D} p(d|o) \log \frac{p(d|o)}{p(d)}, \tag{6.1}$$

where the probability terms needed above can be easily computed from the co-occurrence table between words and stories.

These words are further grouped into 120 cue word clusters (or pseudo-words) using the Information Bottleneck (IB) principle [79]. Words in the same cue word cluster are generally associated with the same semantic topic – for example, {*insurgent*, *insurgents*, *iraq*, *iraqis*, *iraqi*, *marine*, *marines*, *troops*} or {*budget*, *capitol*, *politics*, *lawmakers*, *legislation*, *legislative*, *reform*}. Later each story is represented as a 120-dimensional pseudo-word frequency vector. More examples are illustrated in Table 6.1.

Apparently, the common words co-occurring within a topic are clustered into the same cluster. These cue word clusters can also be measured quantitatively by MI in Eq. 6.1. They generally correspond to certain topics or semantics. For example, in Table 6.1, cluster 1 corresponds to "Iraqi war," cluster 2 to "Arafat's death," cluster 6 to "presidential elections," and cluster 7 to "US president." All the tokens in the stories are then projected to these cue word clusters and the representative dimension is reduced from thousands of words to only 120 cue word clusters. This process is not only for feature reduction but also provides a semantic mid-level representation. For example, two stories regarding "Arafat's death" might mention *Arafat* and *Yasser* exclusively; however, by mapping to these cue word clusters, they would have the same representative cue word cluster index "2" (See Table 6.1). So do *Bush*, *George*, and *President* in cluster 4 of Table 6.1.

The story-level similarity $\psi_t(s_i, s_j)$ is computed using the cosine similarity between the pseudo-word vectors of story s_i and s_j. For cosine similarity $cos(v_i, v_j)$ between arbitrary vectors v_i and v_j, it is to normalize inner product between to vector by diving that by the Euclidean distances of v_i and v_j. This ratio defines the cosine angle between the vectors, with values between 0 and 1 (See Eqn. 6.2).

$$cos(v_i, v_j) = \frac{v_i \cdot v_j}{||v_i|| \times ||v_j||} \tag{6.2}$$

Note that these pseudo-word vectors are also weighted by TF-IDF [26] and normalized into unit vectors, as shown in the following:

$$w(o, d) = \frac{(1 + \log_2 tf(o, d) \times \log_2(N/n_o)}{||d||}, \tag{6.3}$$

where $w(o, d)$ is the weight of token o in story d; $tf(o, d)$ is the within-story term frequency (TF); N is the total number of stories; n_o is the total number of stories where token o occurs; $\log_2(N/n_o)$ is the Inverted Document Frequency (IDF); $||d||$ is the 2-norm of vector d and to normalize $w(o, d)$ into a unit-length vector.

Note that word selection and grouping processes introduced in this section are all conducted in an unsupervised manner. The topic labels are shown for illustration and are not needed (or used) by the algorithm.

6.2.2 Low-level Visual Features

For low-level features, we represent each key-frame by a 273 dimensional feature vector x, which consists of two low-level global visual features. The first is 225-dimensional color moments over 5x5 fixed grid partitions of the image frame. The second is 48-dimensional Gabor texture [80]. More explanation can be found in [9].

The low-level visual feature similarity between two stories s_i and s_j is $\psi_l(s_i, s_j)$

order	words
1	army, artillery, baghdad, carl, casualties, civilian, civilians, coalition, commanders, complement, explosives, fighters, foliage, forces, insurgency, insurgent, insurgents, iraq, iraqi, iraqis, marines, military, offensive, operation, pockets, pollution, rebel, resistance, rocky, searches, soldiers, stations, stronghold, surgeons, triangle, troops
2	*arafat*, assets, authority, bank, buried, egypt, france, french, funeral, gaza, intensive, islamic, israel, israeli, israelis, jerusalem, leader, leadership, mysterious, palestinian, palestinians, paris, peace, plo, ramallah, *yasser*
6	campaign, claim, declared, democracy, election, elections, electoral, fraud, independent, january, ohio, opposition, parliament, planned, political, polls, presidential, protesters, results, selection, supporters, thirtieth, ukraine, vote, voted, voter, voting, winner
7	administration, agenda, *bush*, departure, education, *george*, mandate, policy, present, presidency, *president*, reagan, speculation, successor, term, terrorism, trip, vice

Table 6.1: Four randomly-selected examples of cue word clusters constructed by IB principle. Words of the same cluster are of similar semantics in the original data set. The left column is the cue word cluster sequence number measured by MI (Eq. 6.1) in the descending order.

defined in the following,

$$\psi_l(s_i, s_j) = \max_{i' \in s_i, j' \in s_j} \{\exp - \frac{|x_{i'} - x_{j'}|}{\sigma}\}, \tag{6.4}$$

where $\sigma = 2^5$ is empirically determined through cross-validation. Since most stories have multiple keyframes, the story-level similarity measure takes the maximum pairwise low-level similarity between any two pairs of keyframes within the stories. (cf. Fig. 6.1-(a)). Based on this measure two stories are considered likely to be related to each other if they contain at least one pair of visually similar shots. In our implementation, all the feature elements are normalized by dividing each dimension with its standard deviation.

Figure 6.3: Examples of visual duplicates and image models with parts-based representations from [1]. (a) Parts-based representations of salient parts and their spatial relationships; (2) Two duplicate candidates are compared in parts-based representations (i.e., ARG).

The multimodal similarity we are to investigate is between semantic units (i.e., news stories) which mostly have multiple video shots or representative keyframes. Comparing the shots between stories is a many-to-many matching problem. The most intuitive solution for such many-to-many similarity can be based on bipartite graph matching algorithms [81] such that finding the maximum matching score between story shot pairs [82]; the other is through Earth Mover's Distance (EMD) [83] which formulates the many-to-many matching as a transportation problem to find a least expensive flow of goods from the suppliers to the consumers that satisfiers the consumers' demand. Here we take the highest similarity scores between story keyframes for implementation efficiency. There are many other many-to-many matching algorithms are promising and need further investigation.

6.2.3 Visual Duplicates

As shown in Fig. 6.1, near-duplicates often occur in stories of the same topic. Detection of near-duplicates provides great potential for story linking. In our TRECVID 2005 evaluation work [9], we have found near-duplicate linking to be a highly effective tool in the interactive video search task.

To know the effectiveness and the discriminative capability of visual duplicates, we first conduct a pilot experiment in topic tracking (See more in chapter 7). For topic tracking, one interesting question arises: how many stories from the same topic actually have near-duplicates. To answer this, we address the following two issues: (1) *coverage* – the percentage of stories that share near-duplicates with other stories in the same topic; (2) *false rate* – the percentage of out-of-topic stories which have duplicates with within-topic stories. In the ideal case, the coverage is 1 and the false rate is 0. Fig. 6.2-(b) exemplifies coverage and false rate with a topic (e.g., Iraqi war) and some out-of-topic stories. A duplicate link between two stories is connected if their story-level duplicate score is larger than a threshold. In this example, the coverage is 7/19 and the false rate is 5/14.

Evaluating over the 4 topics in the data set (cf. Section 7.3.1), we plot the coverage vs. false rate curves by varying the near-duplicate detection thresholds in Fig. 6.2-(a). When a higher threshold value is used, we will get a lower coverage and at the same time a lower false rate. It is very impressive to see that we can achieve a moderate coverage (40%-65%) even at almost zero false rate. It strongly supports that story-level duplicate similarity is effective for topic threading.

For automatic detection of near-duplicates, we adopted the parts-based statistical model developed in our prior work [1]. First, salient parts are extracted from an image to form an attributed relational graph (ARG) as shown in Fig. 6.3-(a). Given two candidate images, detection of near-duplicate pairs is formulated as a hypothesis testing problem and solved by modeling the parts association between the corresponding ARGs, and computing the posterior probability (See Fig. 6.3-(b)). The detection score can then be used to derive the near-duplicate similarity between two images, or thresholded to make a binary decision. More details can be seen in [1].

Figure 6.4: Examples of certain concepts annotated in LSCOM-Light [2] for the TRECVID [3] task.

The parts-based duplicate scores are defined between key-frame pairs. Like section 6.2.2 , we represent the story-level similarity in visual duplicates as $\psi_d(s_i, s_j)$, which takes the highest duplicate scores between key-frames of story s_i and s_j respectively. Note that the duplicate similarity is normalized to $[0, 1]$ by a sigmoid function.

6.2.4 Semantic Concepts

Besides low-level visual features, detection of high-level semantic concepts has gained significant interest from researchers. The NIST TRECVID video retrieval evaluation has included high-level feature detection in the last few years [3]. Research has shown the power of using such concepts in improving video search [9, 15].

Various high-level semantic concepts such as "Indoor/Outdoor", "People walking/running", "Explosion or fire" etc., occur frequently in video databases. They also enhance the semantic aspects for video indexing and retrieval. The concepts are annotated with several tools such as that discussed in [84] . An ontology, LSCOM-Light [2] was defined in advance by researchers and experts. A larger set is defined

and formulated in the task of "Large-Scale Concept Ontology for Multimedia [85]." For example, "People walking/running" is a video segment containing video of more than one person walking or running; "Explosion or fire" is one of an explosion or fire. Some other concepts are exemplified in Fig. 6.4.

For concept detection, we adopted the SVM-based method [66, 86] over two low-level visual features mentioned in Section 6.2.2. A grid search with 5-fold cross-validation is first used to find the best parameters (e.g., C and RBF kernel bandwidth γ) for each concept [87]. An optimized SVM model was induced based the parameters and then applied to conduct concept detection, outputting the distance from the SVM decision boundary as a continuous score. Such detection methods have been shown to be general and effective [9, 88]. We apply the same detection framework on the whole set of 39 semantic concepts included in the TRECVID 2005 annotations.

A concept is present in a key-frame if its detection confidence score is larger than a threshold. Counting the present concepts across key-frames of the story results in a representing concept vector. Once we have the frequency vector of the visual concepts, we apply the same set of tools for text (in Section 6.2.1) to derive TF-IDF weighting and unit vector normalization.

The story-level semantic concept similarity $\psi_c(s_i, s_j)$ is defined as the cosine similarity between the concept vectors of two stories. Authors of [28] proposed to use mid-frequency concepts for better story representation. In our work, we have observed that TF-IDF weighting on semantic concepts is able to achieve the same effect since such weighting typically suppresses frequent concepts across stories. From our experiment, the cosine similarity on TF-IDF weighted semantic concepts shows $\sim$30% improvement over the "dice" measure, which is used in [28] to measure the (manually annotated) concept similarity. The dice measure is defined as follows:

order	semantic concepts
1-10	office, weather, computer_tv-screen, person, military, face, car, studio, urban, government-leader,
11-20	building, outdoor, crowd, sky, meeting, entertainment, vegetation, walking_running, road, sports,
21-30	maps, people-marching, explosion_fire, corporate-leader, flag-us, waterscape_waterfront, charts, desert, airplane, police_security,
31-39	truck, mountain, natural-disaster, boat_ship, court, snow, animal, bus, prisoner

Table 6.2: The 39 TRECVID concepts ordered by MI (Eq. 6.1).

$$dice(s_i, s_j) = \frac{2a}{2a + b + c}, \tag{6.5}$$

where a is the number of present concepts in common to both stories s_i and s_j, and b and c represent the number of concepts present in the story s_i but not in s_j, and the reverse, respectively.

The superiority of TF-IDF weighted cosine similarity over dice is natural. The former not only considers the direction of the concept feature vector but also its individual significance in each concept dimension; whereas, the latter considers only the ratio of overlap of two token set only.

In order to assess the influence of individual concepts on topic tracking, we applied the same information-theoretic approach, as described in Eq. 6.1, to measure the relative MI between each detected concept and stories, and to select the most informative concepts. Tab. 6.2 shows the ranking of 39 TRECVID 2005 concepts based on this criteria. Later, in Section 7.3.2, we will analyze the effect of concept selection on the accuracy of topic racking. Note that as opposed to [28] the concept tokens used here are from the automatic detectors rather than manual annotations.

(a) Arabic web news (BBC) (b) English web news (Associated Press) (c) Chinese web news (Chinatimes)

Pope Benedict XVI expressed regret for the reaction his remarks on Islam caused, but did not issue an outright apology for making them...

(d) English broadcast news video (CBS) with the transcripts

Figure 6.5: Examples of a broadcast news video (d) and three web news (a-c) of different languages covering the same topic "Pope sorry for his remarks on Islam," collected on September 17, 2006. The images of Pope Benedict XVI (e.g., those two in the red rectangle) are widely used (in near-duplciates) over all the news sources of the same topic. Aside from the text transcripts or web text tokens, the visual duplicates provide another similarity link between broadcast news videos or web news and help cross-domain topic threading (cf. chapter 7), or even boost text-based video search (cf. chapter 8).

6.3 Multimodal Fusion

The above story-level feature representations provide promising modalities for measuring similarity between story pairs. In prior work, either in high-level concept detection [13, 89] or video search [12, 90], multimodal fusion has shown significant improvement over any single modality. Different feature normalization schemes and fusion strategies are adopted in different applications. However, most of these approaches are in the unit of shots rather in the semantic unit – the story. In the next two chapters, we will demonstrate the significant benefits of these story-level multimodal similarities in terms of cross-domain topic threading and video search

in a large-scale video database.

By fusing multiple modalities, we can augment information exploitation applications across diverse sources, beyond the text baselines. Such representations, though illustrated with broadcast news videos only in the previous sections, can also model the similarities between cross-domain (multimodal) documents (i.e., broadcast news videos, blogs, or web news). For example, Fig. 6.5 illustrates a few examples of a broadcast news video and three web news articles in different languages (e.g., Arabic, English, and Chinese) covering the same topic "Pope sorry for his remarks on Islam." Apparently, the visual duplicates of Pope Benedict XVI are widely used over all the news sources in the same topic. Aside from the text transcripts (if available) or web text tokens alone, the visual duplicates provide another similarity link between broadcast news videos or web news articles and help cross-domain information exploitation.

6.4 Summary

In this chapter, we investigate novel representations and similarities for news stories using multimodal information, including text, visual duplicates, and semantic visual concepts. We rationalize the effectiveness of the proposed representation in terms of visual duplicates. We also present information-theoretic analysis to assess the complexity of each semantic topic and determine the best subset of concepts for tracking each topic. Such story-level representations are hypothesized to augment text-only baselines in applications such as topic threading and video search and will be demonstrated along with novel statistical frameworks for information exploitation applications in the next two chapters.

Chapter 7

Application of Video Similarity – Video Topic Tracking

Videos from distributed sources (e.g., broadcasts, podcasts, blogs, etc.) have grown exponentially. Topic threading is very useful for organizing such vast information resources. Current solutions primarily rely on text features only but encounter difficulty when text is noisy or unavailable. Based on the novel and multimodal story-level similarities discussed in chapter 6, we develop a multimodal fusion framework for estimating relevance of a new story to a known topic. Our extensive experiments using TRECVID 2005 data set (171 hours, 6 channels, 3 languages) confirm that near-duplicates consistently and significantly boost the tracking performance by up to 25%. In addition, we present information-theoretic analysis to assess the complexity of each semantic topic and determine the best subset of concepts for tracking each topic.

7.1 Introduction – Multimodal Topic Tracking

Due to the explosion of Internet bandwidth and television broadcast channels, video streams are easily accessible in many forms such as news video broadcasts, blogs,

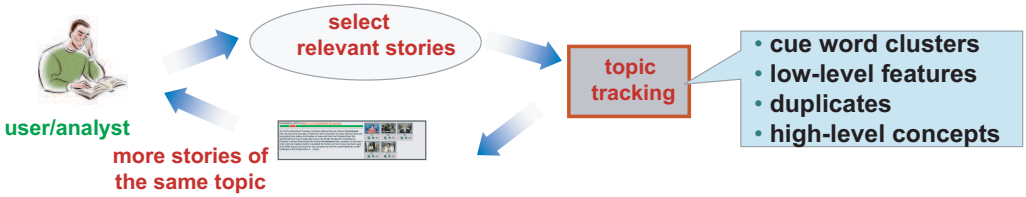


Figure 7.1: Topic tracking to expand the portfolio of the user's interest by utilizing multimodal story-level similarity. The user can prepare certain example stories from unknown media sources or through manual collection.

and podcasts. As news of critical events break (e.g., tsunami or hurricanes), bursts of news stories of the same topic emerge from both professional and amateur sources. Topic threading is an essential task to organize video content from distributed sources into coherent topics for further manipulation in applications like browsing or search.

In the application scenario, a common user or an analyst might have a portfolio of stories that he or she finds interesting, either gathered from unknown media sources or collected manually. They can expand their set of relevant documents by utilizing multimodal context similarity to include or monitor (track) online news from distributed sources, and even avoid the burden of language constraints due to the unreliable text of international news channels. The scenario is sketched in Fig. 7.1.

Current topic threading approaches primarily exploit text information from speech transcripts, closed captions, or web documents. The use of multimodal information such as visual duplicates [1] or semantic visual concepts [3], though not explored before, are also very helpful. There are usually recurrent visual patterns in video stories across sources that can help topic threading. For example, Fig. 6.1 has three example stories[1] from Chinese, Arabic, and English news sources, which

[1]More example stories at "http://www.ee.columbia.edu/∼winston"

cover the same topic. The stories from different channels share a few near-duplicates such as those showing George Bush and Tony Blair, the press conference location, and the audience. Such duplicates are actually effective for news threading across different languages, as is confirmed by our analysis later (cf. Section 6.2.3).

A major research effort for topic threading based on text has been conducted under NIST Topic Detection and Tracking (TDT) event [25], which includes three tasks: (1) story link detection, determining whether two stories discuss the same topic; (2) topic tracking, associating incoming stories with topics that are known to the system; (3) topic detection, detecting and tracking topics that are not previously known to the system. In this chapter, we mainly focus on topic tracking across international broadcast news videos. One representative work of a text-based approach can be found in [26], where authors represent documents as vectors of words, weighted by term-frequency inverse-document-frequency (TF-IDF). The cosine angle is used for measuring document-pair similarity. A modified k nearest neighbor (kNN) approach is then used for classification.

Recently some work started to study new techniques using multimodal information for story topic tracking. Xie *et al.* [27] applied Hierarchical-HMM models over the low-level audio-visual features to discover spatio-temporal patterns, latent semantic analysis to find text clusters, and then fused these multimodal tokens to discover potential story topics. In [28], authors studied the correlation between manually annotated visual concepts (e.g., sites, people, and objects) and topic annotations, and used graph cut techniques in story clustering. In [29], authors addressed the problem of linking news stories across two English channels on the same day, using global affine matching of key-frames as visual similarity. In all of these prior works, neither visual duplicates nor automatically detected visual concepts were used. In addition, comparisons with text-based approaches were not clear.

In this work, we develop novel approaches for story topic tracking using multimodal information, including text, visual duplicates, and semantic visual concepts. We propose a general fusion framework for combining diverse cues and analyze the performance impact by each component. Evaluating on the TRECVID 2005 data set [3], fusion of visual duplicates improves upon the state-of-the-art text-based approach consistently by up to 25%. For certain topics, visual duplicate alone even outperforms the text-based approach. In addition, we propose an information-theoretic method for selecting subsets of semantic visual concepts that are most relevant to topic tracking.

We describe the new multimodal topic tracking framework and story representations in Section 7.2 based on the feature representations, similarity measures, and fusion methods discussed in chapter 6. In Section 7.3, evaluations of the proposed techniques are shown on the TRECVID 2005 benchmark. We present the summary and future work in Section 7.4.

7.2 Topic relevance

Using the story-level representations and similarity measures described in chaper 6, we propose an approach to estimate the relevance score of a new story with respect to a topic. Motivated by [26], we adopt a modified kNN approach to measure the topic relevance score $R_m(s_i)$ of story s_i using modality m . It is defined as follows.

$$R_m(s_i) = \frac{1}{K} \sum_{s_j \in N_K(s_i)} y_{s_j} \cdot \psi_m(s_i, s_j), \tag{7.1}$$

where $y_{s_j} \in \{-1, +1\}$ means relevant or irrelevant to the topic and $N_K(s_i)$ are the K nearest sample stories of s_i, measured with the story-level similarity metric $\psi_m(\cdot, \cdot)$

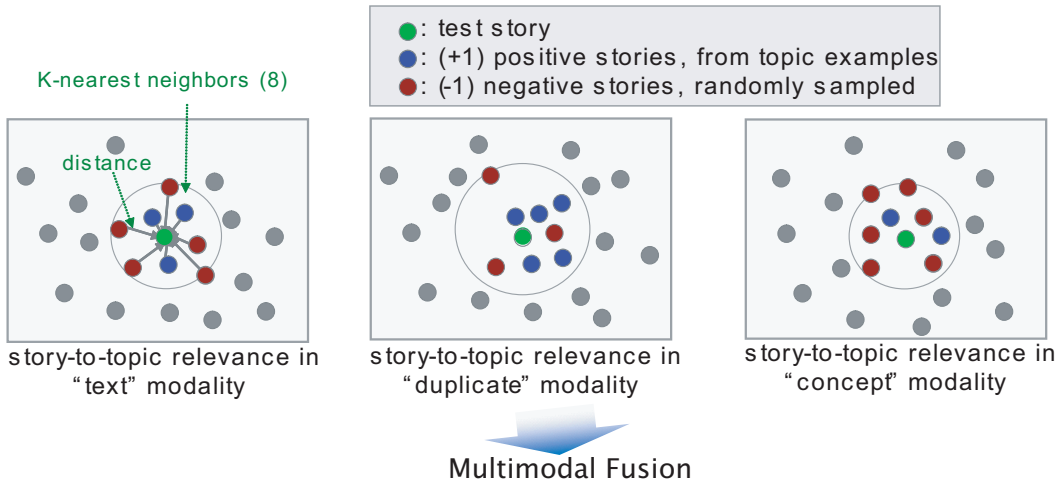


Figure 7.2: The illustration of multimodal (linear)fusion for story to topic relevance from the story-level similarity spaces of different modalities along with user-provided positive and randomly sampled negative example stories. A modified kNN approach is first used in each modality to derive the story-to-topic relevance in the specific feature dimension by considering both the k-neighbor distances (similarities) and labels. Story-to-topic relevances from all modalities are later linearly fused to form a single story-to-topic relevance score for each story.

in modality $m \in \{t : \text{text}, l : \text{low-level visual}, d : \text{duplicate}, c : \text{concept}\}$. Such an approach can benefit from the capabilities of multiple modalities and is illustrated in Fig. 7.2. Note that the positive labels are for the sample stories of interest provided by the user or analyst; whereas, the negative examples are randomly sampled based on the assumption that the true positives are rare. The same assumption is used in many works, such as [34], in multimodal retrieval for large-scale databases.

Basically, the modified kNN model does not simply rely on the counts of positive and negative neighbors but their similarity scores. More sophisticated classification models (e.g., support vector machines) can also be used. However, the main focus of this work is to explore the effectiveness of semantic concepts and visual duplicates. We will pursue the influences of different machine learning methods in the future.

We use a linear weighted fusion method for combining the relevance scores from

different modalities. Such linear fusion model, though simple, has been shown to be one of the most effective approaches to fuse visual and text modalities in video retrieval and concept detection [9, 88]. For story s_i, the fused topic relevance score is

$$R(s_i) = \sum_m w_m \cdot R_m(s_i),$$

where $\sum_m w_m = 1$. The linear weights w_m among modalities are determined empirically based on cross-validation evaluation.

7.3 Experiments

7.3.1 Data set

The data set contains 277 international broadcast news videos from TRECVID 2005 [3], which includes 171 hours of videos from 6 channels in 3 languages (Arabic, English, and Chinese). The time span is from October 30 to December 1, 2004. The story boundaries are from manual annotation, except that 42 Chinese news videos are replaced with automatically detected story boundaries, around 0.84 accuracy (See [9] for more explanations), due to unavailability of manual annotations. There are a total of 4247 stories after commercials are excluded from the original total of 5538. The ASR and MT transcripts are provided by NIST [3]. ASR transcripts are produced by a Microsoft ASR beta system on English and MT transcripts are produced by Virage VideoLogger on Arabic and Chinese. Anchor shots are automatically detected and removed from key-frame sets.

Without official topic annotations, we conducted our own pooling and annotation processes to obtain some topic ground truth. First, an unsupervised IB clustering approach [79] was applied on the ASR and MT transcripts to discover the candidate

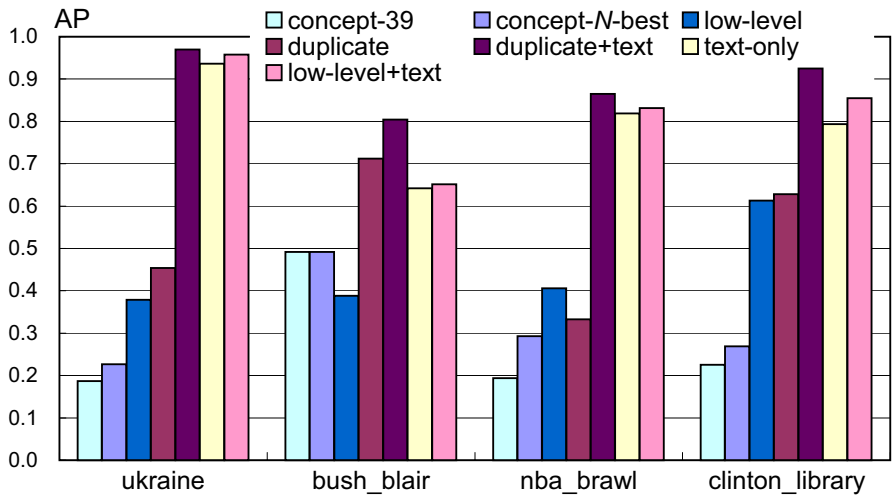


Figure 7.3: Topic tracking performance among different modalities and fusion sets. See explanations in Section 7.3.2.

topics in the corpus. Among them, 4 cross-channel topics are manually selected and then annotated following the guidelines in [25]. The topics are: (1) *ukraine*: Ukrainian presidential election, 74 stories; (2) *bush_blair*: Bush and Blair meet to discuss Mideast peace, 16 stories; (3) *nba_brawl*: NBA players fighting with some viewers in the audience, 29 stories; (4) *clinton_library*: Clinton Presidential Library opens, 25 stories.

The tracking is conducted per topic with 2-fold cross-validation. Negative data for each topic are randomly sampled from the negative pool and its size is controlled to be 16 times of that of the positive data. Each experiment is repeated 6 times and then the mean of the performance is calculated. We use the average precision (AP), as the official metric in TRECVID, to be the performance metric. AP corresponds to the area under an ideal (non-interpolated) recall/precision curve, given a result list ranked based on the relevance scores (cf. Section 7.2). Note that the AP for the random-guess baseline is $1/16 \approx 0.063$.

7.3.2 Performance and discussions

As shown in Fig. 7.3, the most significant finding is that near-duplicate plays an important role in story tracking. When used alone, its tracking performance has been very impressive (AP from 0.33 to 0.71), compared to text-based approach (AP from 0.64 to 0.93). It even outperforms text approaches for certain topics, such as *bush_blair*, in which near-duplicates are frequent. When near-duplicate is combined with text, it consistently improves the text-only accuracy by up to 25%.

Comparing near-duplicate with low-level features, near-duplicate is superior in most cases, except for the topic *nba_brawl*. We hypothesize that in this case near-duplicate detection may not be accurate because the complex objects and background (many small players and complex audience scene), which may make the parts detection and modeling difficult. Note, even for the *nba_brawl* topic, fusion of text with near-duplicate is still better than fusion of text and low-level features.

Automatic semantic concepts are generally worse than text and visual duplicates due to the limited accuracy of the automatic concept detectors and the availability of specific concepts (e.g., named locations and people), which are essential cues for topic threading. We also found that fusion of concepts with text brings only slight improvements; among them, topic *bush_blair* improves the most (around 20%). However, if there exist specific concepts relevant to the topic, tracking based on concepts is very useful. For example, the "sports" concept is found to be very useful for tracking *nba_brawl* topic, so is "flag-us" concept for topic *bush_blair*. We believe that expanding the concept lexicon beyond the 39 concepts in TRECVID will be very valuable.

For concept-based story tracking, we also compare our TF-IDF weighted representation and cosine similarity with the dice measure used in [28]. The TF-IDF

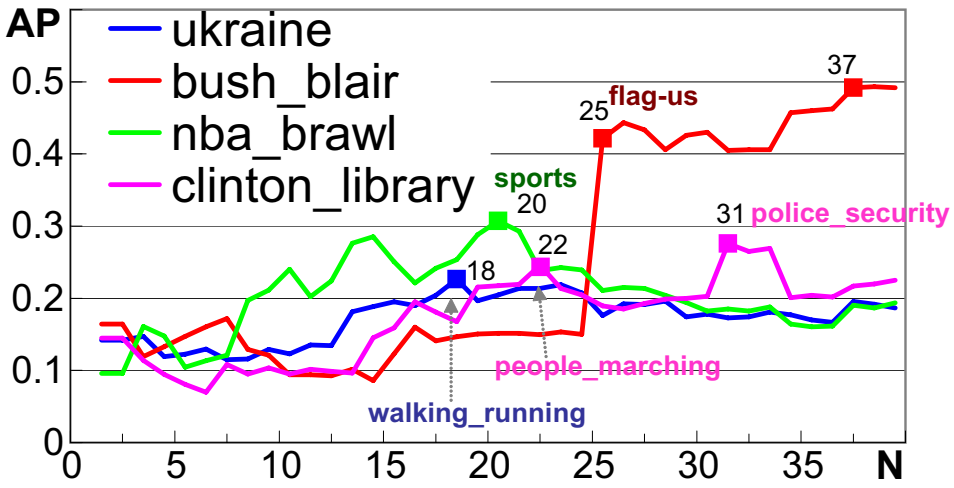


Figure 7.4: Topic tracking performance at variant concept dimensions among the 39 concepts in the order listed in Table 6.2.

method was found to have a performance gain by about 30%.

In addition, we analyze the impact of using subsets of concepts on topic tracking performance. Fig. 7.4 shows the tracking performance when only N most informative concepts were included. The informativeness of each concept is computed using Eq. 6.1, independent of topics. It is interesting to note that different topics reached peak performance at different N values, e.g., 20 for *nba_brawl*, 18 for *ukraine*, and 31 for *clinton_library*. We hypothesize that such differences may be correlated to the diversity of visual content used in each topic and thus may be used to assess the "visual complexity" of each topic. We have found such analysis techniques exciting – to the best of our knowledge, it has been the first work on visual complexity assessment of semantic topics. Finally, some concepts show large influence on specific topics, e.g., "sports" for *nba_brawl*, "flag-us" for *bush_blair*, and "walking-running" for *ukraine*. This confirms the finding mentioned earlier.

7.4 Summary

We propose a novel multimodal topic tracking framework and analyze the contributions of different modalities, including visual near-duplicates and semantic concepts. Visual near-duplicates consistently enhance story tracking across international broadcast news videos, while automatically detected concepts are helpful but require an expanded concept lexicon. In addition, feature selection is necessary to select the correct dimensionality of the concept space and to identify the concepts relevant to each topic.

Chapter 8

Application of Video Similarity – Video Search Reranking

Having observed recurrent images or videos in image search engines, photo sharing sites, and cross-source video databases, we propose to leverage this multimodal similarity between semantic units to improve the initial text query results. The approach is formulated as a random walk problem along the context graph where the graph nodes are news stories. These nodes are connected by the edges weighted with pair-wise multimodal contextual similarities (e.g., visual duplicates, high-level concepts, text tokens, etc.) discussed in chapter 6. The stationary probability of the random walk along the multimodal context graph is used to represent the ranking scores of the reranked results. However, the random walk is biased with reference towards stories with higher initial text search scores – a principled way to consider both initial text search results and their implicit contextual relationships.

We have shown experimentally that the proposed approach can improve retrieval on the average up to 32% relative to the baseline text search method in terms of story-level MAP. Such improvement is achieved without requiring extra efforts to include query expansion, to specify which high-level concepts to use, or to train

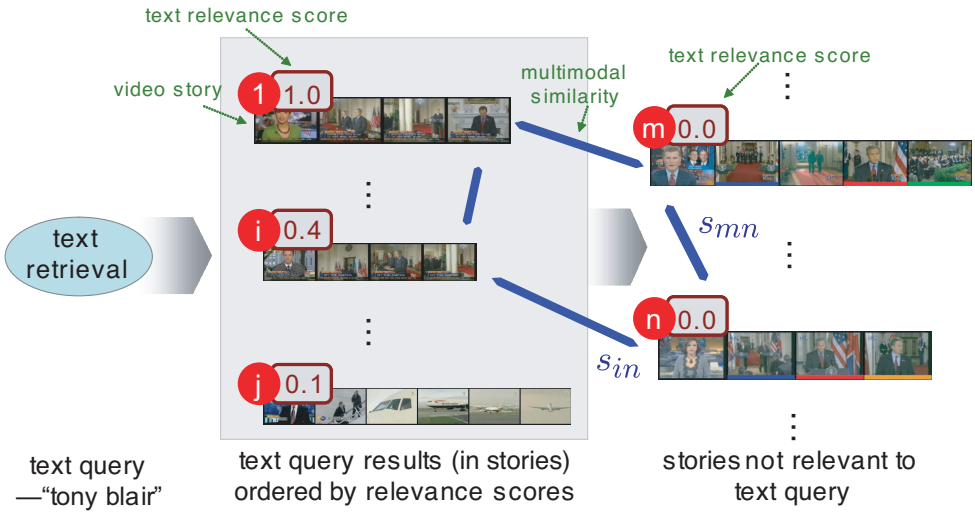


Figure 8.1: Example of a video search that benefits from multimodal story-level similarity on a large-scale video database, even with unreliable text ASR/MT transcripts. Though not relevant to the text query, certain stories can be boosted due to their closeness to some relevant text-query stories by the multimodal similarity (shown in the right panel) consisting of text, visual duplicates, and high-level concepts, etc.

ad-hoc content recognition models, etc. Furthermore, in the people-related queries, which usually have recurrent coverage across news sources, we can have up to 40% relative improvement in story-level MAP. Through parameter sensitivity tests, we also discovered that the optimal weight ratios between text and visual are .15:.85, respectively.

8.1 Introduction

Video and image retrieval has been an active and challenging research area thanks to the continuing growth of online video data, personal video recordings, digital photos, and 24-hour broadcast news. In order to successfully manage and use such enormous multimedia resources, users need to be able to conduct semantic searches over the multimodal corpora either by issuing text keyword queries or providing

example video clips and images (or some combination of the two).

Current successful semantic video search approaches usually build upon the text keyword queries against text associated with the video content, such as speech transcripts, closed captions, and video OCR text. The additional use of other available modalities such as image content, audio, face detection, and high-level concept detection has been shown to improve upon text-based video search systems [9, 12, 14, 15]. However, such multimodal systems tend to get the most improvement through leveraging multiple query example images, applying specific semantic concept detectors, or by developing highly-tuned retrieval models for specific types of queries, such as using face detection and speaker recognition for the retrieval of named persons. In the end, though, it will be quite difficult for the users to acquire example images for their queries. Retrieval by matching semantic concepts, though promising, strongly depends on availability of robust detectors and requires training data. Likewise, it will be difficult for the developers of the system to develop highly-tuned models for every class of query and apply the system to new domains or data sets. It is clear, then, that we need to develop and explore approaches for leveraging the available multimodal cues in the search set without complicating things too much for the system users or developers.

Pseudo-relevance feedback (PRF) [14, 33, 34], is one such tool which has been shown to improve upon simple text search results in both text and video retrieval. PRF is initially introduced in [35], where the top-ranking documents are used to rerank the retrieved documents assuming that a significant fraction of top-ranked documents are relevant. This is in contrast to relevance feedback where users explicitly provide feedback by labeling the top results as positive or negative. The same concept has been implemented in video retrieval. In [14], the authors used the textual information in the top-ranked shots to obtain additional keywords to

perform retrieval and rerank the baseline shot lists. The experiment was shown to improve slightly in TRECVID 2004 video search task [3].

The major assumption for PRF in the previous approaches is that there are some hidden context or topics in the returned documents (or Web pages). PRF-related approaches aim at extracting additional keywords relevant to the (hidden) topics. For example, conducting a search with query terms "tony blair" might get the related keywords "bush washington," implying his visit to US and meet with President Bush, and "chile world trade," the visit to World Trade Organization meeting in Chile. Researchers (e.g., [77]) utilize PRF as a method for query expansion – expanding related query terms from the initial user inputs. Most approaches use ad-hoc methods to select those terms from the large and often erroneous text return sets. Even worse, the approach fails in video search since it can not locate useful visual patterns (e.g., video shots, images) or high-level concepts due to the rareness and low accuracy of the video retrieval system [33]. Thus, most query expansion experiments are restricted to the text modality only.

Across diverse video sources or domains, there are often recurrent patterns sharing similarities either in visual or audio modalities. Even without explicitly discovering these recurrent patterns, we can still utilize these multimodal (soft) "contextual links" to improve the text search results. For example, in Fig. 8.1, an initial text query retrieves certain video stories with text tokens "tony blair;" however, there are still certain relevant stories not retrieved from the text modality. If we consider the contextual similarities in multimodalities (e.g., visual duplicates, text, etc.), we can link such missing stories to the initial text queries.

Such recurrent images or videos are commonly observed in image search engines (e.g., Yahoo or Google) and photo sharing sites (e.g., Flickr). Interestingly, in chapter 6 we have quantitatively analyzed the frequency of such recurrent patterns

(in terms of visual duplicates) for cross-language topic tracking – a large percentage of international news videos share re-used video clips or near duplicates. Such visual patterns were used in chapter 7 for tracking topics and will be the basis for search result reranking in this work.

Based on the same observations, in our previous work [19], we utilized the recurrent video shots in the initial text queries and rerank the search results and yield a significant improvement over the baseline text search. In [19], we measured the recurrent frequency at the shot level which can only match the visual features between shots. Such restriction at the shot level also resulted in deficiency in finding semantic relations.

In this chapter, we propose to utilize the multimodal similarity between semantic units, e.g., video stories, to improve the initial text query results. The video stories are treated as documents. Due to lack of explicit links between documents, the approach is formulated as a random walk over a graph whose nodes represent documents in the search set. These nodes are connected by the edges weighted with pair-wise multimodal contextual similarities (e.g., visual duplicates, high-level concepts, text tokens, etc.) discovered in chapter 6 and our prior work [32]. The stationary probability of the random walk is used to compute the final scores of documents after reranking. The random walk is biased with the preference towards the stories with higher initial text search scores – a principled way to combine both initial text search results and their implicit contextual relationships.

We have shown in the experiments (cf. section 8.4) that the proposed approach can improve the baseline text retrieval up to 32% in story-level MAP, even without using explicit extra efforts to conduct query expansion, to specify which high-level concepts to use, or to train ad-hoc face models. Furthermore, for people-related queries, which usually have recurrent coverage across news sources, we can have

up to 40% relative improvement in story-level MAP. Through parameter sensitivity tests, we also discovered that the optimal text/visual weight ratio for reranking baseline text search results is .15:.85, respectively. It is natural since the most promising contribution should be from the visual modality, which is also illustrated in Fig. 8.1.

The random walk framework is motivated by PageRank [91, 92], the kernel of Google search, which considers the links between web pages. It quantitatively measures the "quality" of the web pages instead of considering the text search "relevance" only. The assumption of PageRank is that when a random surfer continues browsing web pages following the links, the probability that he or she will reside on the page can be used to represent the PageRank score. Intuitively, a web page which is pointed to by an important source page might also be important. The stationary probability of the Web page link graph can be solved by some methods based on eigenvectors.

In section 8.2, we will first introduce the random walk framework and its applications for text retrieval. The extension for video search based on multimodal contextual similarities is described in section 8.3. Section 8.4 demonstrates the experiments using the TRECVID video benchmark. The summary is in section 8.5.

8.2 Random Walk Overview

8.2.1 Random Walk in Text Retrieval

The PageRank algorithm [91], which was first introduced in the Google search engine, utilizes the links between billions of Web pages. The link analysis is treated as a random walk problem based on a link graph, which consists of the link counts between pages. The idea is simulating a web surfer who browses the web pages

continuously and randomly picks a link in the page. At a specific time point, the probability that the web surfer is on a particular page is used to quantitatively measure the Web page quality or the so-called PageRank score. The stationary probability is actually a normalized dominant eigenvector (cf. Definition B..1 in Appendix) derived from the link graph. The authors of [91] argue that such scores can be used to improve the search quality rather than considering search relevance only.

The random walk process is governed by the Markov matrix (See more in Definition B..2 of Appendix) which describes the probability that the surfer clicks on a specific link. Some modifications are needed to guarantee the existence and uniqueness of the stationary probability. For example, we need first to modify the link graph so that a surfer visiting a dangling page (i.e., a page without forward links) can still jump to other pages in the next time step and continue the process iteratively. The other modification is to make sure that the Markov matrix is irreducible or strongly conducted (cf. section 8.2.3).

The Markov matrix and its stationary probability will be introduced in section 8.2.2. Its convergence properties and conditions are discussed in section 8.2.3. Section 8.2.4 will discuss the Power Method, an efficient way to solve the principal eigenvector.

8.2.2 Markov Matrix and Stationary Probability

Assuming we have n nodes (or states) $D = \{d_1, ..., d_n\}$ in the random walk process. A node might correspond to a Web page in the text retrieval problem or a story in the broadcast news video. A stochastic (transition) matrix

$$\mathbf{P} \equiv [p_{ij}]_{n \times n} = [p(j|i)]_{n \times n}, 1 \leq i, j \leq n$$

is used to govern the transition of a random walk process. p_{ij} is the probability that transition from state i to state j occurs. The state probability at time instance k is

$$\mathbf{x}_{(k)} \equiv \left[p_{(k)}(i)\right]_{n\times 1},$$

a column vector of the probabilities residing in the n nodes at the instance. Since being Markovian, the state probability at time instance $\mathbf{x}_{(k)}$ can be modeled by the time instance $\mathbf{x}_{(k-1)}$ and the transition matrix $\mathbf{P}$ such that

$$\mathbf{x}_{(k)}^T = \mathbf{x}_{(k-1)}^T \mathbf{P}$$

or equivalently

$$\mathbf{x}_{(k)} = \mathbf{P}^T \mathbf{x}_{(k-1)}. \tag{8.1}$$

The stationary probability

$$\mathbf{x}_\pi \equiv \lim_{k\to\infty} \mathbf{x}_{(k)}$$

is the state probability of the random walk process as the time instance proceeds to infinity if the convergence conditions are satisfied. Since it is stationary and according to Eqn. 8.1, we can easily state that

$$\mathbf{x}_\pi = \mathbf{P}^T \mathbf{x}_\pi. \tag{8.2}$$

The solution of $\mathbf{x}_\pi$ will be the eigenvectors of $\mathbf{P}^T$. According to Theorem B..1, the stationary probability $\mathbf{x}_\pi$ is exactly the (normalized) dominant eigenvector, the one corresponding to the largest absolute eigenvalue, of $\mathbf{P}^T$.

8.2.3 Convergence of Stationary Probability

Given a square raw matrix $\mathbf{S}$ such as a link graph in PageRank applications [91] or an affinity matrix between nodes, a few preprocessing steps are required to ensure convergent stationary probability. First, we need to normalize the rows by dividing each row with the sum of the entire row and then yield the normalized matrix $\mathbf{P}$, where each row adds up to 1.

In some cases, there might be a few dangling links in certain nodes. These nodes have no outlinks to other nodes or have zeros for the entire row. It is problematic since the random walk process will be trapped when transiting to such nodes. To solve the problem, empirically, these all-zero rows of $\mathbf{P}$ will be replaced with $\mathbf{e}^T/n$, where $\mathbf{e}$ is a n-dimensional column vector with all 1. Then the process will continue to jump to all nodes uniformly with the probability $1/n$. We then produce the matrix $\mathbf{P}'$ from the the normalized matrix $\mathbf{P}$.

Furthermore, the stationary probability exists if and only if the transition matrix is *aperiodic* and *irreducible* [93]. The former condition is easily satisfied in almost all applications such as web documents or broadcast news video stories. The latter, however, is not met for most applications. Reducible Markov chains are those containing sets of nodes in which the chain is eventually trapped. In other words, if reordering the transition matrix $\mathbf{P}$, we can get the following canonical form

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ 0 & \mathbf{P}_{22} \end{pmatrix}.$$

This states that when reaching node set $\mathbf{D}_2$ the random walk process can not reach any nodes within set $\mathbf{D}_1$. Hence, an irreducible Markov chain is one in which every node is eventually reachable from every other node – the states are fully connected.

To make sure the matrix is fully connected, the most common approach is to add a new set of complete outgoing transitions with small transition probabilities to all nodes. The transition probability is controlled by the vector $\mathbf{v}$ and we can then have the modified random walk process as the following:

$$\begin{aligned} \mathbf{v} &\equiv [v(i)]_{n\times 1}, 1 \leq i \leq n \\ \mathbf{E} &= \mathbf{e}\mathbf{v}^T \qquad (8.3) \\ \mathbf{P}'' &= \alpha\mathbf{P}' + (1-\alpha)\mathbf{E}, 0 \leq \alpha \leq 1 \qquad (8.4) \end{aligned}$$

Eqn. 8.4 states that in each time step the process will jump to a random page selected by the vector $\mathbf{v}$ with probability $1-\alpha$, or will follow the stochastic matrix $\mathbf{P}'$ with probability α. Such modifications in Eqn. 8.3 and 8.4 guarantee that the stochastic matrix $\mathbf{P}''$ is fully connected and that a unique stationary probability exists. Note $\mathbf{x}_{(k-1)}^T\mathbf{E} = \sum_{i=1}^{n} p_{(k-1)}(i)\mathbf{v}^T = \mathbf{v}^T$; therefore, a transition based on probability specified by $\mathbf{E}$ always traverses to a fixed state vector $\mathbf{v}^T$.

The small transition probability $\mathbf{v}$, also called the *personalization vector*, may be used to further enable other promising applications besides ensuring the irreducibility of the stochastic matrix. For example, if $\mathbf{v} = \mathbf{e}/n$, the process will uniformly jump to all nodes when not following the links governed by $\mathbf{P}'$. In [94], the authors use $\mathbf{v}$ to embed the relevance of a page to the query for a more intelligent search rather than using the links between pages only; intuitively, a user will jump to the page that has higher query-relevance scores. Such approach is promising for fusing the text relevance score and the stationary probability estimated directly from the link graph [91] or raw feature matrix only. Meanwhile, from a business perspective, search engines can twist $\mathbf{v}$ to give their preferences toward certain pages or web

sources [92]. The authors of [95] tuned the vector $\mathbf{v}$ to produce topic-dependent PageRank scores where they adjust the influence of the links by considering web pages of the same topic.

Once $\mathbf{P}''$ is available, we can then calculate the stationary probability of $\mathbf{P}''$ by deriving its (normalized) dominant eigenvectors of $(\mathbf{P}'')^T$ (cf. Theorem B..1). For efficiency, we need not calculate all the n eigenvectors and eigenvalues but just take advantage of the Power Method described in the next section.

8.2.4 Power Method

To calculate the dominant eigenvector of a stochastic matrix $\mathbf{A}$ (i.e. $(\mathbf{P}'')^T$ in section 8.2.3), we adopted the Power Method which iteratively applies $\mathbf{x}_{(k)} = \mathbf{A}\mathbf{x}_{(k-1)}$ until $\mathbf{x}_{(k)}$ converges. Note that $\mathbf{x}_{(k)}$ needs to be normalized by its 1-norm ($||\cdot||_1$) to make sure that it adds up to 1. Applying this method iteratively can result in convergence to the dominant stationary probability $\mathbf{x}_\pi$. The intuition is as the follows. Since a positive matrix and each of its columns sums to 1, $\mathbf{A}$ has n eigenvalues $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ and their corresponding eigenvectors $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_n\}$. Meanwhile, according to Theorem B..1,

$$\lambda_1 = 1 > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_n|$$

The initial n-dimensional vector $\mathbf{x}_{(0)}$ can be a linear combination of these n eigenvectors:

$$\mathbf{x}_{(0)} = \mathbf{u}_1 + a_2\mathbf{u}_2 + a_3\mathbf{u}_3 + ... + a_n\mathbf{u}_n.$$

Hence,

$$\mathbf{x}_{(1)} = \mathbf{A}\mathbf{x}_{(0)} = \mathbf{u}_1 + a_2\lambda_2\mathbf{u}_2 + a_3\lambda_3\mathbf{u}_3 + ... + a_n\lambda_n\mathbf{u}_n,$$

and then

$$\mathbf{x}_{(k)} = u_1 + a_2\lambda_2^k\mathbf{u}_2 + a_3\lambda_3^k\mathbf{u}_3 + ... + a_n\lambda_n^k\mathbf{u}_n.$$

All the eigenvalues, except for the first, are less than 1. If k is large enough, $\mathbf{x}_{(k)} \approx \mathbf{u}_1$, which is the dominant eigenvector.

Empirically, the initial vector $\mathbf{x}_{(0)}$ can be arbitrary values; i.e. $\mathbf{e}/n$ (randomness) or the personalization vector $\mathbf{v}$. The Power Method takes the input of initial vector $\mathbf{x}_0$, the convergence threshold ϵ_π (i.e., $1E-6$) and the designated stochastic matrix $\mathbf{A}$, which satisfies the criteria discussed in section 8.2.3. The Power Method algorithm is as follows:

(1) $k = 1$;

(2) $\mathbf{x}_{(k)} = \mathbf{A}\mathbf{x}_{(k-1)}$;

(3) $\mathbf{x}_{(k)} = \mathbf{x}_{(k)}/||\mathbf{x}_{(k)}||_1$;

(4) $\delta = ||\mathbf{x}_{(k)} - \mathbf{x}_{(k-1)}||_1$;

(5) $k = k + 1$;

(6) go to (2) unless $\delta < \epsilon_\pi$.

This algorithm is intuitive and an example demonstrating how to normalize the affinity matrix $\mathbf{S}$ and the calculation of the dominant eigenvector is shown in Example B..1 of the Appendix.

8.3 Video Search via Random Walk

There are a few works in the video research community utilizing the random walk framework. The authors of [96] formulate the *normalized cut* problem [97] in video

segmentation as a random walk process. By normalizing the similarity matrix of pair-wise pixels, one obtains the stochastic matrix and then utilizes the second largest eigenvector or sets of eigenvectors [98] to do image segmentation. It differs from our approach as it is used for clustering of pixels; we focus on the high-level contextual similarities between stories; meanwhile, the text modality in our approach is used as a prior in the random walk process. Besides, we are primarily interested in the ranking of the stationary probability rather than the similarities in the spectral (eigenvector) space.

The authors of [99] utilized the same framework to associate image regions to certain keywords by constructing a graph with nodes composed of annotated keywords, images, and their regions. The links between the nodes are binary by thresholding similarities between nodes or the existence of annotated words to certain images.

In addition, He *et al.* proposed an approach called *ImageRank* in [100], where they first use the spectral clustering approach such as [97] to cluster images, sampled from Corel Image Database, into a few classes. In this process, an image is represented by an 11-dimensional Boolean vector, each entry indicating the existence of a specific keyword. Secondly, in each class, three images are selected as the representatives of the class. The criteria is determined by the stationary probability derived from the transition matrix modeled by low-level features between images in the class. The approach does not address the multimodal search problem and the text and visual features are treated independently. In addition, only the image-level similarity is considered, not the similarity between video stories using both visual and text features associated with the video.

In our work, we leverage the recurrent patterns of topics across video sources and develop video search reranking methods by utilizing context similarity and the random walk framework. We formulate the video search as a random walk process

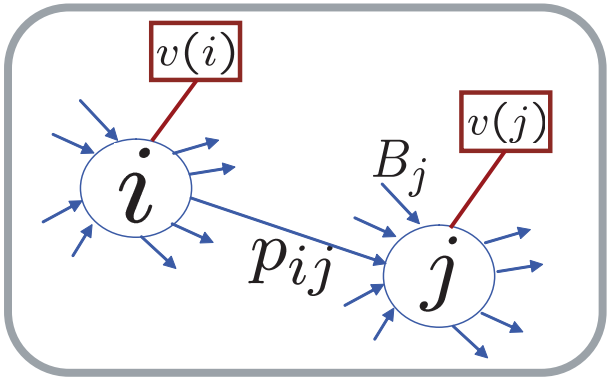


Figure 8.2: Example of a context graph for random walk over nodes (stories); i and j are node index with their original text search scores $v(i)$ and $v(j)$; p_{ij} is the transition probability from node i to j; B_j are edges back to node j.

among these stories connected with multimodal contextual similarities. The solution is the stationary probability when the propagation between nodes converges.

We will introduce how to embed the affinity matrix between stories into the random walk framework in section 8.3.1. We then present variants of methods in section 8.3.2 to incorporate initial text search scores as priors.

8.3.1 Random Walk on Multimodal Story-Level Similarities

Motivated by the observations in video/image searches, where recurrent patterns are commonly seen [32], we explore the multimodal contextual similarities between stories for improvement against the text baseline queries.

We formulate the solution as a random walk over the *context graph*, where stories are nodes and the edges between them are weighted by multimodal contextual similarities, as illustrated in Fig. 8.2. We use the *context ranking score*, the stationary probability of the random walk over the context graph, to represent the search relevance scores. In this context, we consider both the multimodal similarities or transition probabilities $\mathbf{P}$ between stories and the original (normalized) text search

scores $\mathbf{v}$. In this framework, the state probability $x_{(k)}(j)$ of node j at time instance k is:

$$x_{(k)}(j) = \alpha \sum_{i \in B_j} x_{(k-1)}(i) p_{ij} + (1 - \alpha) v(j), \tag{8.5}$$

where B_j is the set of edges back to node j, p_{ij} is the contextual transition probability from story i to j, and $\alpha \in [0, 1]$ linearly weights two terms.

Eqn. 8.5 is actually an interesting interpretation of the random walk process exemplified in Fig. 8.2 and 8.1. The state probability at node j is influenced by nodes B_j, which are edges back to j. Intuitively, $x_{(k)}(j)$ is parameterized by its neighboring nodes at time instance $k-1$ and its own initial text scores $v(j)$. Both are then linearly fused with weights α and $1-\alpha$ respectively. For the first term of Eqn. 8.5, we consider not only the state probabilities of its neighbors B_j but also their corresponding transition probabilities – how possible it is to reach node j. The second term is the initial text score for node j. Such linear fusion considers the state probabilities (or search context relevances) of its neighbors and its initial text scores. The linear fusion is commonly used in multimodal search and concept detection research [71, 101] and has been shown to be effective.

Eqn. 8.5 also shows that each node propagates its influence to its neighbors at each time step until the whole process converges. The neighbors B_j of node j can be all of the nodes in the context graph or only a subset according to the configurations in different applications (cf. section 8.3.2).

Likewise, augmenting text search with contextual similarities, the relevance of the story is weighted not only by its initial text score but also the importance of its neighbors. The relationship is updated recursively until all nodes in the graph converge. The updating process is the same as the Power Method (cf. section 8.2.4)

and stops when it converges. For each node, the new search relevance score is its stationary probability, if it exists. For example, the stationary probability of node j is

$$x_\pi(j) = \alpha \sum_{i \in B_j} x_\pi(i) p_{ij} + (1-\alpha) v_j. \tag{8.6}$$

Naturally, we have $\sum_j x_\pi(j) = 1$. If we assume $\mathbf{E} = \mathbf{e}\mathbf{v}^T$ and go through some algebra, then we can get

$$\begin{aligned} \mathbf{x}_\pi^T \equiv [x_\pi(j)]_{1\times n} &= \alpha \mathbf{x}_\pi^T \mathbf{P} + (1-\alpha)\mathbf{x}_\pi^T \mathbf{E} \\ &= \mathbf{x}_\pi^T [\alpha \mathbf{P} + (1-\alpha)\mathbf{E}] \end{aligned} \tag{8.7}$$

Eqn.8.7 has almost the same formulation as the PageRank algorithm, introduced in section 8.2.2, if we set $\mathbf{P}'' = \alpha\mathbf{P} + (1-\alpha)\mathbf{E}$ (cf. Eqn. 8.4). The only exception is $\mathbf{P}'$ of Eqn. 8.4, which is modified for the dangling problem mentioned in section 8.2.3.

We need to derive the context graph transition probability from the raw story-level affinity matrix $\mathbf{S}$. Intuitively, we just conduct the normalization and ensure that each row sums to 1. That is,

$$p_{ij} = \frac{s_{ij}}{\sum_k s_{ik}}.$$

A random walk process will more likely jump to the nodes with higher (contextual) similarities. Similar to the PageRank algorithm, we need to handle the same dangling problem and just set those all zero rows of affinity matrix $\mathbf{S}$ as $\mathbf{e}^T$.

The affinity matrix $\mathbf{S}$ is composed of pair-wise similarities between stories. The

story-level similarity consists of visual duplicates and text (in terms of "cue word clusters") similarities, which are linearly fused. Such modalities and fusion are the same as those discussed in chapter 6. In our following experiments, variant modality weights between visual duplicates and text are tested experimentally and will be further discussed in section 8.4.3.

8.3.2 Improving Text Search with Context Ranking

In this work, our goal is to utilize the context graph between the stories to improve the text search baseline in video retrieval. The problem is formed as a random walk among the story-level (contextual) similarity graph. The context ranking score or the stationary probability $\mathbf{x}_\pi$ in the context graph is one of the criteria to rank the stories. The original (normalized) text search scores can be another cue to fuse the results. There are some promising configurations and baselines for the proposed frameworks. These methods vary in the aspects of: (1) Which stories form the nodes of the context graph? The whole set of stories (documents) or those with text relevant scores only? (2) How to use the initial text search results? Fused with the context ranking scores or acting as the prior in the personalization vector $\mathbf{v}$ for the random walk process? To maximize the performance of these promising methods and modalities, we test several combinations of the approaches in this section. The experiments results are shown in section 8.4.2.1.

- Full-Ranking (**FR**): In *full-ranking*, all documents (stories) in the database are used to construct the context graph for the random walk framework. The personalization vector is $\mathbf{v} = \mathbf{e}/n$, uniform over all stories. The stationary probability, then, is used as the context ranking score. The assumption behind FR is that those stories having the most recurrences will be given higher

context ranking scores. This is not the best setup since it does not utilize the initial text search scores but it does provide a baseline for comparison.

- Full-Ranking and Text Score (**FRTS**): The context rank scores, derived in FR, are then fused with the the initial text scores by averaging the two, which generally is comparable with those with varying weights [71].

- Full-Ranking with Text Prior (**FRTP**): Same as FR, except that the initial text score is used as the personalization vector $\mathbf{v}$ or the random walk prior. The second term, $v(j)$, in Eqn. 8.5 is no more equal across all stories but biased by the initial text scores. Such prior will put more preferences towards those stories with higher initial text search scores. The effect is modulated by the linear weighting factor α.

- Partial Ranking (**PR**): In *partial ranking*, only stories in the database that have initial text search scores higher than some threshold are used to construct the context graph. The multimodal similarity graph $\mathbf{S}$ is first built based on this subset of stories relevant to the query terms. A uniform personalization vector $\mathbf{v}$ over the subset of stories, as FR, is used. The assumption is to consider those text-query-relevant stories only and then order them by the recurrence frequency – more recurrent stories gain more relevance, or higher stationary probability, from the partial context graph.

- Partial-Ranking and Text Score (**PRTS**): Same as PR, except the context ranking score is further fused with the initial text scores by averaging the two.

- Partial-Ranking with Text Prior (**PRTP**): Same as PR, except that the initial text search score is used as the personalization vector $\mathbf{v}$ for the random walk

prior. The assumption is to consider the recurrent patterns within the text-query-relevant stories only and biased with the prior of the initial text search scores.

- Partial-Ranking with Text Prior and Link Reduction by K Nearest Neighbors (**PRTP-KNN**): In the previous method PRTP, the context graph built over the subset of relevant stories is a complete graph or composed of all edges between the nodes in the graph. The authors of [99] had discovered that selecting the top-K most significant edges between nodes can further boost the random walk performance. Likewise, in this method, we select the top-K edges, originating from each node and having highest similarities. Note that to discover the effects of K, we experiment with varying values of K in the following experiments.

The breakdowns of the experimental results on two different initial text sets are compared and discussed in section 8.4.

8.4 Experiments

8.4.1 Data set

The experiments are conducted over the TRECVID 2005 data set [3], which contains 277 international broadcast news videos and accumulates 171 hours of videos from 6 channels in 3 languages (Arabic, English, and Chinese). The time span is from October 30 to December 1, 2004. The ASR and machine translation (MT) transcripts are provided by NIST [3]; ASR transcripts are produced by a Microsoft ASR beta system on English and MT transcripts for foreign channels are produced by Virage VideoLogger on Arabic and Chinese.

In our prior work [19], we have shown that the best approach to text searches over speech recognition transcripts from international broadcast news videos is to use the transcripts from the entire story. This makes sense since the true semantic relationships between images and the text transcripts exist at the story level: if a concept is mentioned in the text it is likely to appear in the video somewhere within the same story, but unlikely to appear in the next story or the previous one. Story boundaries can be extracted with reasonable accuracy automatically by analyzing the visual characteristics of the image content and the speech of the anchorperson. Different story boundary detectors are trained separately for each language – English, Chinese, and Arabic, in order to capture specific production styles in each source. The performance, evaluated using TRECVID metrics (F1[1]), is 0.52 in English, 0.87 in Arabic, and 0.84 in Chinese [65, 76]. Our experiments have shown that associating text within each story outperforms the fixed-window based approach consistently with approximately 15% improvement in terms of MAP in TRECVID 2003, 2004, and 2005. Using manually annotated story boundaries offers an additional 5-10% increase in performance. Note that automatically detected boundaries are used in this thesis.

The video search ground truth provided in TRECVID 2005 is at the shot level. Since we are evaluating at the story level, we have to convert the shot-level ground truth into story-level ground truth. A story is considered positive if it contains at least one shot that is labeled as positive for a particular query. It is an "OR" operation among shots in the same story. This is intuitive since browsing stories in a story level is more natural and a story is a semantic coherent unit; a story should be labeled as positive once a target object or shot is discovered in the story.

[1] $F1 = \frac{2 \cdot P \cdot R}{P+R}$, where P and R are precision and recall rates defined in TRECVID [3] story boundary detection task. F1 equals the precision and recall values when they are equal.

For example, a user might be interested in news related to Chinese President Hu Jintao and thus any story that contains at least one shot of the person should be considered relevant.

For measuring performance, we adopted non-interpolated average precision (AP), which corresponds to the area under a (non-interpolated) recall/precision curve. Since AP only shows the performance of a single query, we use mean average precision (MAP), which is simply the mean of APs for multiple queries, to measure average performance over sets of different queries in a test set. See more explanations in [3].

Natural language processing tools such as stemming, part-of-speech tagging, etc, are first applied to the ASR/MT transcripts provided by NIST. We conduct the experiments on two text-based video search sets, "baseline-text" and "example-text." The former is from the baseline text search approach based on Okapi method [19]. The latter is motivated by multi-bag support vector machines (MB-SVM) in [34]. We sample the ASR/MT transcripts from stories associated with the example video clips as positive examples and randomly sample other stories in the database for pseudo-negatives. A discriminative model (e.g., SVM) is trained to classify other stories in the test set. The process is repeated several times (with different random samples of negative data) and the (positive) distances to the margin plane of these SVM models are then fused to compute the final relevance scores. In TRECVID 2005 data set, the "example-text" approach significantly outperforms the "baseline-text" (cf. Tab. 8.1). See more explanations in [102].

methods	baseline-text MAP	%	example-text MAP	%
initial text search	0.204	-	0.274	-
Full Ranking (FR)	0.109	-46.8	0.109	-60.3
Full Ranking and Text Score (FRTS)	0.238	16.5	0.302	10.2
Full Ranking with Text Prior (FRTP)	0.210	2.8	0.280	2.1
Partial Ranking (PR)	0.196	-4.1	0.240	-12.3
Partial Ranking and Text Score (PRTS)	0.255	24.5	0.309	12.9
Partial Ranking with Text Prior (PRTP)	0.271	**32.5**	0.333	**21.6**
PRTP-KNN	0.271	32.5	0.333	21.6

Table 8.1: The performance (MAP at depth 20) and relative improvements (%) from the initial text search results at different methods (cf. section 8.3.2). Note that in PRTP-KNN the best results among variant K (number of neighbors) are shown. We have fixed $\alpha = 0.8$ for all methods.

8.4.2 Performance and discussions

8.4.2.1 Experiments of variant methods

The performances of variant methods proposed in section 8.3.2 are shown in Table 8.1. Apparently, FR method, which considers only the recurrent frequency over 2580 automatically detected stories has the worst performance among the methods. It is understandable since the recurrence frequency does not necessarily match the search topics although it is interesting to see that frequency alone produces a non-trival search results with MAP at .109. However, in the FRTS method, averaging the context ranking and the initial text search scores does improve the performance and by 16.5% and 10.2% for two different search tools respectively. It confirms that the recurrence frequency, measured by the stationary probability over the multimodal context graph, is a nice criterion for ranking the story search. However, the FRTP method has almost no improvement since its random walk context graph contains all the stories in the database and the initial text search scores are used as the preference in the personalization vector $\mathbf{v}$. The random walk over a large graph,

even with text search prior, might not converge to the search targets.

Clearly the partial ranking – context ranking over a subset stories with positive text-based search relevance scores – outperforms the methods using the full set of stories. It might be that the search task is still dominated by the text modality. Filtering with text search scores to obtain a subset of relevant stories guarantees the performance of adding context ranking no worse than the baseline performance.

Even in the partial ranking, considering only recurrence frequency on those relevant stories does not bring any gains, as shown in Table 8.1. However, in PRTS, averaging the context ranking score $\mathbf{x}_\pi$ and the initial text search scores does improve the average performance over the 24 queries. The most significant improvement comes from the PRTP method, where the personalization vector $\mathbf{v}$ is used as the random walk prior which enforces the random walk process not only going through the multimodal similarities link but also taking into account the initial text search scores (cf. Eqn. 8.6). The average relative improvement over all queries are significant – 32.5% and 21.6% in the two sets "baseline-text" and "example-text" respectively. Our conjecture is that the random walk framework utilizes the recurrent stories across sources, while the initial text scores are used as a prior for the personalization vector $\mathbf{v}$ in the random walk process. More experiment breakdowns are explained in section 8.4.2.2.

In Table 8.1, the PRTP-KNN method, which selects the K most relevant neighbors for each node in the context graph, does not improve the performance in both text-based search sets. Meanwhile, for both data sets, the performances degrade when K is small (below 100 on the average). This is interesting since it implies that the required connectivity, in order to exploit the power of the context graph, can not be too small for this application domain. However, reducing the neighborhood size is still desirable since it translates into less computation load in executing the

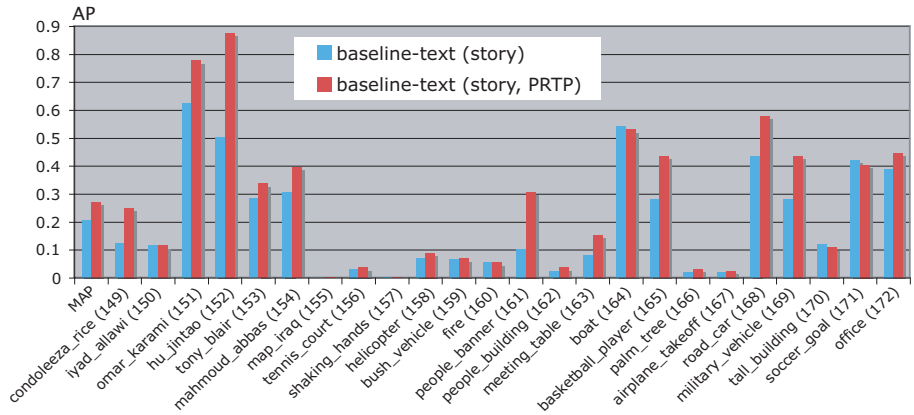


Figure 8.3: Performance (in story-level AP) of PRTP at depth 20 across topics based on the text-base search set "baseline-text." The relative MAP improvement over all queries is 32.45% and that over named-people queries (149-154) is 40.7%.

iterative random walk procedure.

In the following experiments, to investigate the performance breakdowns in more details, we experiment only on the PRTP method.

8.4.2.2 Analysis of PRTP Performance

The breakdowns of the PRTP performance across topics using the two text search methods, "baseline-text" and "example-text," are shown in Fig. 8.3 and 8.4 respectively. The overall (story-level) MAP improvement in "baseline-text" is from 0.204 to 0.271 (32.5%) and 0.274 to 0.333 (21.6%) in "example-text." The former is larger since it initially has a lower performance and has more space for improvement. More interestingly, the relative improvements in the people-related queries (149 to 154) are significant in both sets and are 40.7% and 19.1% respectively. Generally the improvement is larger for people-based queries than general queries since these people-related topics, in most cases, are reported in major political news events worldwide. Even with poor ASR/MT, the use of recurrent patterns, especially the visual duplicates, greatly improves the story ranking performance. The same be-

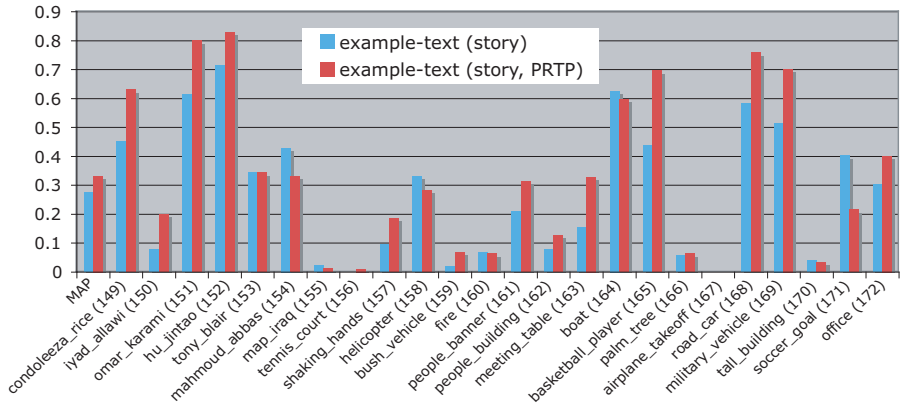


Figure 8.4: Performance (in story-level AP) of PRTP at depth 20 across topics based on text-base search set "example-text." The relative MAP improvement over all queries is 21.6% and that over named-people queries (149-154) is 19.1%.

havior was also observed and confirmed in [32], where the visual duplicates are used to augment multimodal topic tracking.

The PRTP method improves almost all queries with many queries showing significant gains, as depicted in Fig. 8.3 and 8.4. Even for the few queries that did not benefit, none has significant loss. Besides those people queries, salient improvements also come from queries such as "basketball_player" (165). It was related to an important topic of "NBA brawl," in which the news about the basketball players fighting with the fans are covered across channels. So does the query "miltary_vehicle" (169), which consists largely of Iraq-related news stories. Another one is "people_banner" (161), though it includes no specific objects, it is covered in a few events (from different topics) in the news collections. Because of these cross-source relations, the inclusion of contextual relationships indeed improve the precisions over the text-base queries.

Compared to the traditional techniques of text-based query expansion, our proposed method is advantageous since it does not require explicit discovery of new words or visual models which are often needed in expanded queries. In contrast,

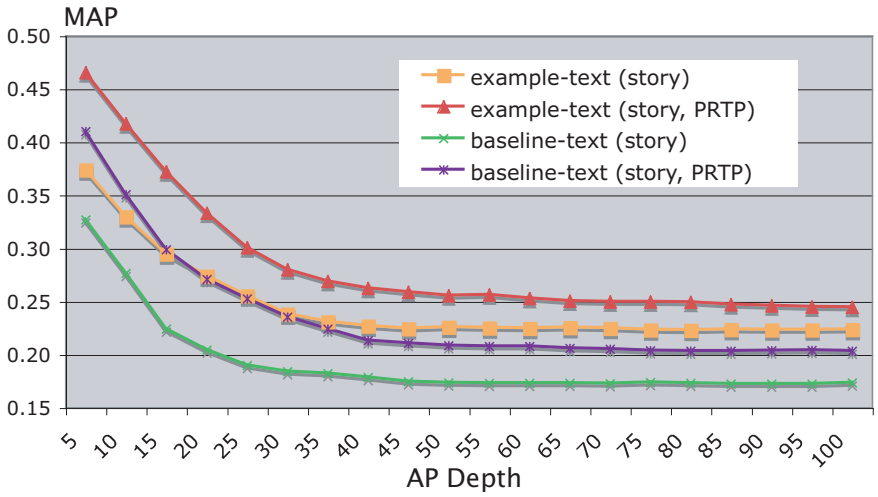


Figure 8.5: Consistent performance improvements for PRTP method evaluated at variant depths in both text search sets.

our method leverages the underlying similarity among documents and uses them to establish context and refine the initial search scores.

The previous discussions are based on the AP at depth 20, which is a reasonable number of results that users like to browse in typical search scenarios. In such case, an improvement from AP 0.50 to 0.87 is significant, as in the query "hu_jintao" (152) shown in Fig. 8.3. Nevertheless, we are also interested in the performance at different AP depths. The results are plotted in Fig. 8.5, which shows clear MAP improvements over both text search sets. The relative improvement decreases as the evaluation depth increases since the results are gradually dominated by irrelevant stories.

8.4.3 Parameter Sensibility

8.4.3.1 Visual Modality Significance

The significance of the contributions from the visual modality has been of great interest to researchers. The context graph of the random walk process includes linearly

weighted combination of contributions from visual duplicates and text similarity. To analyze the impacts of each individual modality, we compare the performance with different text weights ranging from 0.0 to 1.0 with an increasing step of 0.05 and plot the results in Fig. 8.6. Interestingly, we discover that the best weight for text is 0.15 in both "example-text" and "baseline-text." The dominant weight (0.85) for the visual modality is consistent with what we have anticipated. Our conjecture is that the initial text search has already used the text information in computing the relevance scores in response to the query. Therefore, additional gain by measuring cross-document similarity will probably come from the visual aspect (e.g., via visual duplicate detection), rather than the text modality.

Other interesting observations are related to the extreme cases of visual only and text only, shown as the two end points of each curve in Fig. 8.6. The performances are almost the same as the text-base search sets if the context graph considers text modality solely (i.e., text weight = 1). However, when using the context graph in visual duplicates only (i.e., text weight = 0), the PRTP performance still achieves significant improvement (+26.1%) in the "baseline-text" and slightly (+6.5%) in "example-base" text-base search sets. It confirms the significant contribution of visual similarity relationship in context ranking.

8.4.3.2 α sensibility

The use of α in Eqn. 8.5 plays an important role in balancing the personalization vector $\mathbf{v}$ (i.e., the text search prior in this experiment) and the multimodal contextual relationships modeled in the Markov matrix. The value of α also impacts the performance and convergence speed [99, 103]. With α close to 1, the random walk process relies almost entirely on the story similarities and ignores the text search prior; hence, the performance degrades sharply. Based on our empirical compar-

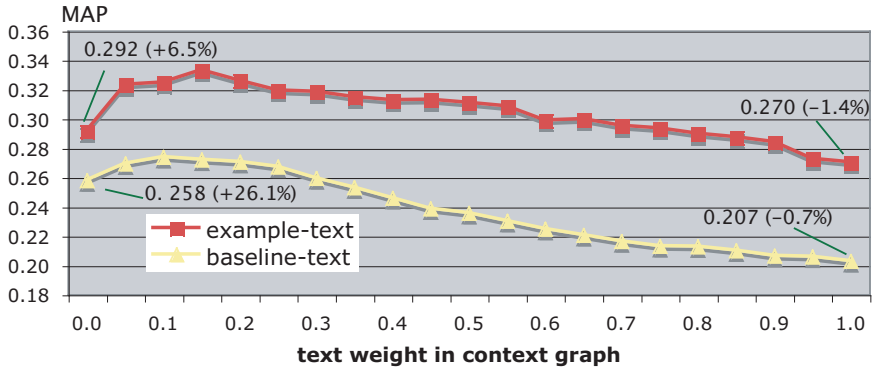


Figure 8.6: PRTP on both text search sets with variant text weights, where $\alpha = 0.8$ in the random walk procedure. The numbers in the parentheses are the relative improvement from their corresponding text search results. The best performance of the ratio of text vs. duplicates is around .15:.85.

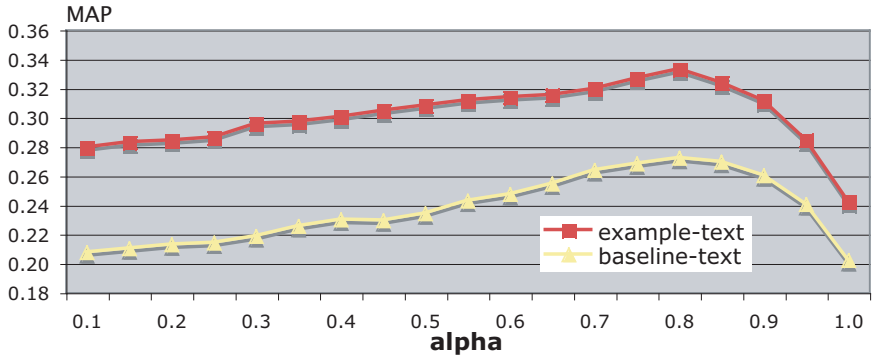


Figure 8.7: PRTP on both text search sets with variant α in Eqn. 8.6, where the text weight in the context graph is fixed at 0.15. The best performance in each set is when $\alpha = 0.8$.

isons, the best setup for both data sets is $\alpha = 0.8$, which is close to but slightly less than what were reported in large-scale web applications. The best value reported in [103] was around 0.9 and 0.85 in [92]. The authors of [99] reported that the performance reaches a plateau as α grows from 0.5 to 0.9.

8.4.4 Related Work

Our prior work [19] proposed a promising reranking approach in the image level based on the observation that search targets usually occur multiple times with high

similarity and scatter at the top ranks of initial text search results. Thus, we revised the reranking scores to favor those images with high feature density and high initial text scores. However, it measured the recurrence frequency at the image level which can only match the visual features between image pairs. Such restriction at the image level also resulted in deficiency in finding semantic relations.

In text summarization, the authors of [104] proposed an approach, LexRank, for computing sentence importance based on the concept of eigenvector "centrality" in a graph representation of sentences. The nodes of the graph are sentences from documents of the same topic. The similarity between sentence pairs is the cosine similarity between two corresponding word vectors. The approach is to utilize the random-walk method to find the most visited sentences in the collection and use the visit probability as the criteria for selecting the most salient sentences for text summarization. Though it adopted the same Power Method, the similarity graph is computed based on the text features only and the uniform distribution is adopted for the personalization vector $\mathbf{v}$.

Our reranking processes are based on the initial text search results. Such methods utilize the initial search scores as some type of pseudo supervision in order to further improve the search accuracy. Therefore, they are related to approaches such as Transductive Learning [66, 105] and Co-training [106]. These two paradigms consider the problem of using a large number of unlabeled samples to boost performance of a learning algorithm when only a small set of labeled examples are available. However, in the search reranking problem, we can not exactly locate what positive data might be in the initial text search results. Instead, only the (noisy) search relevance scores from the initial text search are available. Use of such text search scores in reranking multimedia documents has not been explored before and is a novel area we investigate in the reranking methods presented here.

8.5 Summary

We propose to utilize the multimodal similarity between semantic units (news video stories) to improve the initial text query results. The approach is formulated as a random walk problem over a graph with the nodes representing stories in the search set and edge indicating pair-wise multimodal contextual similarities (e.g., visual duplicates, high-level concepts, text tokens, etc.). The stationary probability of the random walk over the graph is used to compute the scores of the reranked results. In addition, the random walk is biased with the preference towards the stories with higher initial text search scores – a principled way to consider both initial text search results and their implicit contextual relationships.

We have shown in the experiments that the proposed approach can improve the baseline text search results by up to 32% in story-level MAP. The proposed method does not require extra processes of query expansion or use of predefined content recognition models. Additionally, in the people-related queries, which usually have recurrent coverage across news sources, we can achieve up to 40% relative performance improvement. Through parameter sensitivity tests, we also discovered that the optimal text/visual weight ratio for multimodal context ranking .15:.85, respectively. The optimal value for the α parameter is about 0.8, consistent with the results found in web page ranking applications.

Chapter 9

Conclusions and Future Work

We first summarize the work presented in the thesis. Specifically, we have presented an automatic framework for video indexing and retrieval on large-scale video databases, which have been applied to topic threading, video search, and video story segmentation. We then discuss a few potential areas of future research.

9.1 Thesis Summary

To tackle the challenges of video indexing and retrieval over large-scale video databases, we thoroughly address four fundamental problems: (1) How to select and fuse a large number of heterogeneous multimodal features from image, speech, audio, and text? (2) How to automatically discover and model mid-level features from a large set of raw features? (3) How to model similarity between multi-modal documents such as news videos or multimedia web documents? (4) How to explore unsupervised or pseudo-supervised situations in video search to boost performance of automatic search? We approach such challenging problems with the following solutions.

9.1.1 Feature fusion and selection from heterogeneous modalities

We investigate issues pertaining to multimodal fusion and selection in broadcast news video story segmentation, which has been shown to be essential for video indexing, summarization, and topic tracking. In this context, we applied and extended the Maximum Entropy statistical model to systematically select and effectively fuse diverse features from multiple levels and modalities, including visual, audio, and text. We have included various features such as motion, face, music/speech discrimination, speech rapidity, high-level text segmentation information, prosody, etc. We also incorporated some novel features such as syllable cue terms (in Mandarin) and significant pauses for international broadcast news video segmentation. The statistical fusion model is used to automatically discover salient features that are most useful for the detection of story boundaries. We also introduced a novel feature wrapper to address diverse feature data types – different time scales, discrete vs. continuous, periodic vs. aperiodic, etc. We demonstrated excellent performance (F1 = 0.76 for ABC news, 0.73 for CNN news, and 0.90 for Mandarin news). Our method was applied on broadcast news video segmentation in TRECVID 2003, and 2004, achieving top performance in booth years.

9.1.2 Automatic discovery of mid-level features

Abstraction of raw features into mid-level features has been shown promising for improving performance in semantic content classification. However, in most systems, such mid-level features are selected manually. To automate the process, we propose an information-theoretic framework, to automatically discover adequate mid-level features via a method based on semantic-consistent clusters. Our approach is based on the principle of mutual information maximization, through which optimal clusters

are discovered to preserve the highest mutual information about the semantic labels. We adopted and extended the Information Bottleneck (IB) technique, which has been applied to text document classification before. In order to handle the diverse features from video, we also extended the original IB method to the case of high-dimensional continuous features.

We apply these discovered semantically-consistent clusters as new mid-level representations for the video segmentation problem in chapter 4. It achieves significant performance gain over representations derived from conventional clustering techniques and even the mid-level features selected manually. It is the first work to remove the dependence on the manual process in choosing the mid-level visual features and the huge labor cost involved in annotating the training corpus for the mid-level feature detectors. We have tested this method in story segmentation and reported strong performance gains in TRECVID 2004 and 2005. It was the only method shown effective for segmenting the diverse videos from multiple international channels in 2005.

We also show that the discovered mid-level features can be used to significantly improve the video search quality. The IB-based clusters provide a novel way for denoising the raw search scores from initial text search, and constraining the visual reranking step. The method improves upon the text-search baseline by up to 23%, and is comparable with other sophisticated models, which use highly tuned models for each specific query type.

9.1.3 Semantic threading across video sources

Adequate models and computational methods for video similarity at the story level are important for video topic tracking and video story search. Current solutions primarily rely on text features and therefore encounter difficulty when text is noisy

or unavailable. We extend the text modality and propose new representations and similarity measures for news videos combining low-level visual features, visual near-duplicates, and high-level semantic concepts automatically detected in videos.

Contrary to similarity measures defined over fixed windows or shots, here we consider feature representations at the semantic level – video stories, which usually has variable lengths. We address the many-to-many pair-wise similarity measures in such story-level representations.

Based on the novel story-level similarities, we develop a multimodal fusion framework for estimating relevance of a new story to a known topic. Our extensive experiments using TRECVID 2005 data set (171 hours, 6 channels, 3 languages) confirm that near-duplicates consistently and significantly boost the tracking performance by up to 25%. Visual duplicates alone even outperform text-only approaches in certain topics. In addition, we present information-theoretic analysis to assess the complexity of each semantic topic and determine the best subset of concepts for tracking each topic.

9.1.4 Graph-based random walk for video search reranking

Having observed recurrent images or videos in image search engines, photo sharing sites, and cross-source video databases, we propose to leverage this multimodal similarity between semantic units to improve the initial text query results. The approach is formulated as a random walk problem along the context graph where the graph nodes are news stories. These nodes are connected by the edges weighted with pair-wise multimodal contextual similarities (e.g., visual duplicates, high-level concepts, text tokens, etc.) discussed in chapter 6. The stationary probability of the random walk along the multimodal context graph is used to represent the ranking scores of the reranked results. However, the random walk is biased with preference

towards stories with higher initial text search scores – a principled way to consider both initial text search results and their implicit contextual relationships.

We have shown experimentally that the proposed approach can improve retrieval on the average up to 32% relative to the baseline text search method in terms of story-level MAP. Such improvement is achieved without requiring extra efforts to include query expansion, to specify which high-level concepts to use, or to train ad-hoc content recognition models, etc. Furthermore, in the people-related queries, which usually have recurrent coverage across news sources, we can have up to 40% relative improvement in story-level MAP. Through parameter sensitivity tests, we also discovered that the optimal weight ratios between text and visual are .15:.85, respectively.

9.2 Future Directions

Development of a principled framework to support video indexing and retrieval on large-scale databases is challenging. While we have presented in this thesis some promising solutions, there remain many open issues and exciting opportunities for further advancement.

9.2.1 Considering Temporal Information

In chapter 2, we consider the video classification problem (i.e., story boundary detection) in the local fixed windows surrounding the candidate points only. Another perspective, making use of the temporal cues such as the change of genres over time or story length distributions, provides a different way of formulation. Similar observations are considered in works such as [24]. Recently, authors of [107] utilized the temporal transitions in high-level concept detection and demonstrated improved

accuracy in detecting concepts that have temporal dependency.

Meanwhile, such temporal effects might be considered when we estimate the correlations of topics across sources. Video stories of the same topic tend to follow certain temporal patterns. We believe that understanding of such temporal relationships can facilitate more reliable document similarities when compared to those using content similarity only (cf. chapter 6). Similar ideas utilizing temporal priors for text document threading were proposed in [108] and [109], where the former used a fixed time window over all topics while the latter incorporated both context and time information for different topics.

9.2.2 Speed-up of IB framework

We had demonstrated the effectiveness of the IB framework in both automatic mid-level feature discovery (cf. chap 4) and deriving information preserving clusters for video search reranking (cf. chapter 5). One of the bottlenecks of the framework is the complexity associated with the kernel density estimation [62], which requires $O(N^2)$ time complexity for N images. One promising solution is to speed up KDE through some approximation process such as the *N-body* method [110], which skips pairwise distances that are far away and can thus reduce the complexity to linear time in certain cases.

Meanwhile, in the video search reranking, we have found that most of the relevant results are in the first few clusters and users usually are interested in the first few pages of the search results. Therefore, in order to improve efficiency, we can just rerank the images or videos in the first few clusters produced by the IB clustering process. Such flexibility provides great potential for optimizing the tradeoffs between accuracy and efficiency.

9.2.3 Balance of visual and semantic consistence

The IB framework presented in chapter 3, though preserving the semantic consistence in each cluster, does not result in consistent clusters in terms of visual similarity. Thus, when inspecting the images within each cluster, we often see images of different appearances. As a consequence, it is not a visually coherent clustering approach. However, by combining semantic and feature distortion, we can add more flexibility to the proposed framework in other video applications. The feature distortion can be measured by conventional metrics discussed in Rate Distortion Theory [38] and then fused with the semantic distortion in Equation 3.4 for joint optimization.

The extended framework may be used to summarize images in a large returned set in search applications. Images/videos with similar relevance and appearance may be organized (clustered) together to summarize the search result pages or provide a concise presentation format.

9.2.4 Scalable multimodal topic tracking

We have seen the exponential growth of digital media across domains and demonstrated promising results in cross-domain topic tracking. However, scalability of the proposed technique to hundreds of video channels or web sources is still a significant challenge.

To address the scalability issue, first, we might develop hierarchical clustering techniques to group sources based on geographical scopes and topical of each source. This will allow us to discover topics within individual groups first then link topics across groups. The computational complexity can thus be reduced significantly. Second, we might leverage rich resources available on online portals such as Google

news. By using the topics of the online sites as seeds, we can quickly link the stories in the video broadcast to the topics in the online sites using our multi-modal linking techniques mentioned in chapter 7.

9.2.5 Photo/video relevance and summarization

The huge collections of videos and photos from the public sharing sites or personal collections are growing exponentially. Efficient browsing tools are needed to explore such large data sets. Recently, there are a few interesting works addressing such problems. Rother *et al.* proposed "AuthoCollage" [111], which uses an automatic procedure to construct a visually appealing collage from a collection of input images. The resulting collage should be representative of the collection, able to capture the main theme. Some other works try to summarize the photos near a famous tourist spot by utilizing 3D model visualization [112] or the map-based navigation over a large collection of geo-referenced ones [113].

Such approaches are based on a set of provided images relevant to a trip or location interesting to a user. However, to summarize the images or videos from the search engines, we need to (1) determine the *relevance* of the images, (2) select *representative* images from the collection, and then (3) summarize them with clear themes. We might utilize the further extension of the IB framework discussed in section 9.2.3 to pursue tasks (1) and (2). Then, those relevant and representative images can be used to compose a meaningful and concise summarization of retrieved videos or photos.

9.2.6 Exploiting social media

Social media – online tools and platforms that people use to share opinions, insights, experiences, and perspectives with others – has gained great popularity recently,

such as Flickr (photo sharing), YouTube (video sharing), etc. In these systems, we easily find recurrent visual patterns due to the common interest among users. More interestingly, some of those shared media also have human annotations such as locations, events, dates, people, etc. Furthermore, von Ahn *et al.* proposed ESP Game [114], an entertaining web-based game that motivates users to help annotate content in images. The annotation quality might vary among users; however, such annotation mechanism has attracted great attention from researchers in object recognition, search, and human computer interface.

By utilizing the annotations over images or videos in social media, we can develop new methods to improve image/video search within the social media or across other domains (e.g., searching photos of personal collections). The main idea is to extend the context reranking framework proposed in chapter 8, where the multimodal similarity between the annotated images and non-annotated ones can be used to help improve the search performance over the unlabeled data set.

Additionally, the annotations from online sources can be further used to boost automatic metadata generation (i.e., object categorization or high-level concept detection) by avoiding the costly and time-consuming processes of manual annotation. For example, authors of [115] try to learn object categories from Google image searches in an automatic manner. A more interesting work is from [116], where authors even proposed a system for predicting the effects on classifier accuracy if labels from online sources are used instead of those from manual processes.

However, such annotations are mostly erroneous. A reranking process (e.g., the IB ranking technique in chapter 5) can be used to improve the quality of image labels. Moreover, we may incorporate the constrains of visual consistence in IB reranking so that we can select representative images for each category. Such images covering comprehensive views can then be used to train automatic classifiers, which

are expected to be more accurate than classifiers trained over the raw image set.

Bibliography

[1] Dong-Qing Zhang and Shih-Fu Chang, "Detecting image near-duplicate by stochastic attributed relational graph matching with learning," in *ACM Multimedia*, New York, 2004.

[2] Milind R. Naphade, Lyndon Kennedy, John R. Kender, Shih-Fu Chang, John R. Smith, Paul Over, and Alex Hauptmann, "A light scale concept ontology for multimedia understanding for trecvid 2005," Tech. Rep., IBM, 2005.

[3] *TRECVID: TREC Video Retrieval Evaluation*, http://www-nlpir.nist.gov/projects/trecvid/.

[4] Peter Lyman, Hal R. Varian, Kirsten Swearingen, Peter Charles, Nathan Good, Laheem L. Jordan, and Joyojeet Pal, "How much information 2003?," Tech. Rep., Berkeley University, 2003.

[5] *Google*, http://www.google.com.

[6] *Yahoo! News*, http://news.yahoo.com/.

[7] *flickr*, http://www.flickr.com/.

[8] *YouTube*, http://www.youtube.com/.

[9] Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Lexing Xie, Akira Yanagawa, Eric Zavesky, and Dong-Qing Zhang, "Columbia University TRECVID-2005 video search and high-level feature extraction," in *TRECVID Workshop*, Waishington DC, 2005.

[10] *Informedia Project*, http://www.informedia.cs.cmu.edu/.

[11] *IBM Marvel*, http://domino.research.ibm.com/comm/research_projects.nsf/pages/marvel.index.html.

[12] Alexander G. Hauptmann and Michael G. Christel, "Successful approaches in the TREC Video Retrieval Evaluations," in *ACM Multimedia 2004*, New York, 2004.

[13] Milind R. Naphade and John R. Smith, "On the detection of semantic concepts at TRECVID," in *ACM Multimedia*, New York, 2004, pp. 660–667.

[14] Tat-Seng Chua, Shi-Yong Neo, Ke-Ya Li, Gang Wang, Rui Shi, Ming Zhao, Huaxin Xu, Qi Tian, Sheng Gao, and Tin Lay Nwe, "TRECVID 2004 search and feature extraction task by NUS PRIS," in *TRECVID Workshop*, Waishington DC, 2004.

[15] Arnon Amir, Janne Argillandery, Murray Campbell, Alexander Haubold, Giridharan Iyengar, Shahram Ebadollahi, Feng Kang, Milind R. Naphade, Apostol (Paul) Natsev, John R. Smith, Jelena Tesic, and Timo Volkmer, "IBM research TRECVID-2005 video retrieval system," in *TRECVID Workshop*, Waishington DC, 2005.

[16] *MyLifeBits Project*, http://research.microsoft.com/barc/mediapresence/MyLifeBits.aspx.

[17] Lexing Xie, Lyndon Kennedy, Shih-Fu Chang, Ajay Divakaran, Huifang Sun, and Ching-Yung Lin, "Discovering meaningful multimedia patterns with audio-visual concepts and associated text," in *IEEE International Conference on Image Processing*, Singapore, 2004.

[18] Dong-Qing Zhang, Ching-Yung Lin, Shi-Fu Chang, and John R. Smith, "Sematnic video clustering across sources using bipartite spectral clustering," in *IEEE International Conference on Multimedia & Expo*, Taipei, Taiwan, 2004.

[19] Winston H. Hsu, Lyndon S. Kennedy, and Shih-Fu Chang, "Video search reranking via information bottleneck principle," in *ACM Multimedia*, Santa Babara, CA, USA, 2006.

[20] Alexander G. Hauptmann and Michael J. Witbrock, "Story segmentation and detection of commercials in broadcast news video," in *Advances in Digital Libraries*, 1998, pp. 168–179.

[21] Wei Qi, Lie Gu, Hao Jiang, Xiang-Rong Chen, and Hong-Jiang Zhang, "Integrating visual, audio and text analysis for news video," in *7th IEEE Intn'l Conference on Image Processing*, 2000.

[22] Z. Liu and Q. Huang, "Adaptive anchor detection using on-line trained audio/visual model," in *SPIE Conf. Storage and Retrieval for Media Database*, San Jose, 2000.

[23] Stanley Boykin and Andrew Merlino, "Machine learning of event segmentation for news on demands," *Communication of the ACM*, vol. 43, no. 2, February 2000.

[24] Lekha Chaisorn, Tat-Seng Chua, Chun-Keat Koh, Yunlong Zhao, Huaxin Xu, Huamin Feng, and Qi Tian, "A two-level multi-modal approach for story segmentation of large news video corpus," in *NIST TRECVID 2003 Workshop*, Gaithersburg, MD, 2003.

[25] LDC, "Tdt3 evaluation specification version 2.7," 1999.

[26] Yiming Yang, Jaime G. Carbonell, Ralf D. Brown, Thomas Pierce, Brian T. Archibald, and Xin Liu, "Learning approaches for detecting and tracking news events," *IEEE Intelligent Systems*, vol. 14, no. 4, 1999.

[27] Lexing Xie, Lyndon Kennedy, Shih-Fu Chang, Ching-Yung Lin, A. Divakaran, and H. Sun, "Discover meaningful multimedia patterns with audio-visual concepts and associated text," in *ICIP*, Singapore, 2004.

[28] John R. Kender and Milind R. Naphade, "Visual concepts for news story tracking: Analyzing and exploiting the nist trecvid video annotation experiment," in *CVPR*, San Diego, 2005.

[29] Yun Zhai and Mubarak Shah, "Tracking news stories across different sources," in *ACM Multimedia*, Singapore, 2005.

[30] Ramesh Sarukkai, "Video search: Opportunities and challenges," in *Multimedia Information Retrieval Workshop (MIR)*, Singapore, November 2005.

[31] John Battelle, *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*, Portfolio Trade, 2006.

[32] Winston H. Hsu and Shih-Fu Chang, "Topic tracking across broadcast news videos with visual duplicates and semantic concepts," in *International Conference on Image Processing (ICIP)*, Atlanta, GA, USA, 2006.

[33] Rong Yan, Alexander Hauptmann, and Rong Jin, "Multimedia search with pseudo-relevance feedback," in *International Conference on Image and Video Retrieval*, Urbana-Champaign, IL, USA, 2003.

[34] Apostol Natsev, Milind R. Naphade, and Jelena Tesic, "Learning the semantics of multimedia queries and concepts from a small number of examples," in *ACM Multimedia*, Singapore, 2005, pp. 598–607.

[35] Jaime G. Carbonell, Yiming Yang, Robert E. Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee, "Translingual information retrieval: A comparative evaluation," in *International Joint Conference on Artificial Intelligence*, 1997.

[36] Winston H. Hsu and Shih-Fu Chang, "A statistical framework for fusing mid-level perceptual features in news story segmentation," in *IEEE International Conference on Multimedia and Expo*, 2003.

[37] Winston H. Hsu, Shih-Fu Chang, Chih-Wei Huang, Lyndon Kennedy, Ching-Yung Lin, and Giridharan Iyengar, "Discovery and fusion of salient multimodal features towards news story segmentation," in *IS&T/SPIE Electronic Imaging: Storage and Retrieval Methods and Applications for Multimedia*, San Jose, CA, 2004.

[38] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[39] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[40] Doug Beeferman, Adam Berger, , and John Lafferty, "Statistical models for text segmentation," *Machine Learning*, vol. 34, no. special issue on Natural Language Learning, pp. 177–210, 1999.

[41] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tur, and Gokhan Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, pp. 127–154, 2000.

[42] M. Franz, J. S. McCarley, S. Roukos, T. Ward, and W.-J. Zhu, "Segmentation and detection at IBM: Hybrid statistical models and two-tiered clustering broadcast news domain," in *Proceedings of TDT-3 Workshop*, 2000.

[43] Di Zhong, *Segmentation, Index and Summarization of Digital Video Content*, Ph.D. thesis, Graduate School of Arts and Sciences, Columbia University, 2001.

[44] Hualu Wang and Shih-Fu Chang, "A highly efficient system for automatic face region detection in mpeg video," *IEEE Transactions on Circuits and Systems for Video Technology (CSVT)*, vol. 7, no. 4, 1997.

[45] 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.

[46] Dongqing Zhang, "Face detection in news video with gemoetric active contours," Tech. Rep., IBM T. J. Watsom Research Center, 2003.

[47] Anil K. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall, New Jersey, 1989.

[48] Barry Arons, "Pitch-based emphasis detection for segmenting speech recordings," in *International Conference on Spoken Language Processing*, Yokohama, Japan, 1994.

[49] Jacqueline Vaissiere, "Language-independent prosodic features," in *Prosody: Models and Measurements*, Anne Cutler and D. Robert Ladd, Eds., pp. 53–66. Springer, Berlin, 1983.

[50] Hari Sundaram, *Segmentation, Structure Detection and Summarization of Multimedia Sequences*, Ph.D. thesis, Columbia University, 2002.

[51] *The Snack Sound Toolkit*, http://www.speech.kth.se/snack/.

[52] S. Dharanipragada, M. Franz, J. S. McCarley, S. Roukos, and T. Ward, "Story segmentation and topic detection in the broadcast news domain," in *1999 DARPA Broadcast News Workshop*, 1999.

[53] Berlin Chen, Hsin-Min Wang, and Lin-Shan Lee, "Discriminating capabilities of syllable-based features and approaches of utilizing then for voice retrieval of speech information in mandarin chinese," *IEEE transactions on speech and audio processing*, vol. 10, no. 5, 2002.

[54] J. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, no. 1-2, pp. 89–108, 2002.

[55] Arnon Amir, Marco Berg, Shih-Fu Chang, Winston H. Hsu, Giridharan Iyengar, Ching-Yung Lin, Milind Naphade, Apostol (Paul) Natsev, Chalapathy Neti, Harriet Nock, John R. Smith, Belle Tseng, Yi Wu, and Donqing Zhang, "Ibm research trecvid-2003 video retrieval system," in *NIST TRECVID 2003 Workshop*, Gaithersburg, MD, 2003.

[56] Gang Wu and Edward Chang, "Adaptive feature-space conformal transformation for imbalanced-data learning," in *International Conference on Machine Learning*, Washington DC, 2003.

[57] Lekha Chaisorn, Tat-Seng Chua, , Chun-Keat Koh, Yunlong Zhao, Huaxin Xu, Huamin Feng, and Qi Tian, "A two-level multi-modal approach for story segmentation of large news video corpus," in *TRECVID Workshop*, Washington DC, 2003.

[58] Teuvo Kohonen, *Self-Organizing Maps*, Springer, Berlin, third edition, 2001.

[59] Noam Slonim, Nir Friedman, and Naftali Tishby, "Unsupervised document classification using sequential information maximization," in *25th ACM international Conference on Research and Development of Information Retireval*, 2002.

[60] Noam Slonim and Naftali Tishby, "Agglomerative information bottleneck," in *Neural Information Processing Systems (NIPS)*, 1999.

[61] Shiri Gordon, Hayit Greenspan, and Jacob Goldberger, "Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations," in *International Conference on Computer Vision*, 2003.

[62] David W. Scott, *Multivariate Density Estimation : Theory, Practice, and Visualization*, Wiley-Interscience, 1992.

[63] Greg Hamerly and Charles Elkan, "Learning the k in k-means," in *Neural Information Processing Systems (NIPS)*, 2003.

[64] Pinar Duygulu and Alexander Hauptmann, "What's news, what's not? associating news videos with words," in *International Conference on Image and Video Retrieval*, Dublin City University, Ireland, 2004.

[65] Winston Hsu, Lyndon Kennedy, Shih-Fu Chang, Martin Franz, and John Smith, "Columbia-IBM news video story segmentation in trecvid 2004," Tech. Rep. ADVENT #207-2005-3, Columbia Universiry, 2005.

[66] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[67] Winston H. Hsu and Shih-Fu Chang, "Generative, discriminative, and ensemble learning on multi-modal perceptual fusion toward news video story segmentation," in *IEEE International Conference on Multimedia and Expo (ICME)*, Taipei, Taiwan, June 2004.

[68] Vojtech France and Vaclav Hlavac, "Statistical pattern recognition toolbox for matlab," Tech. Rep., Czech Technical University, 2004.

[69] R. Fergus, P. Perona, and A. Zisserman, "A visual category filter for google images," in *Proceedings of the 8th European Conference on Computer Vision, Prague, Czech Republic*, May 2004.

[70] Lyndon Kennedy, Paul Natsev, and Shih-Fu Chang, "Automatic discovery of query class dependent models for multimodal search," in *ACM Multimedia*, Singapore, November 2005.

[71] Kieran Mc Donald and Alan F. Smeaton, "A comparison of score, rank and probability-based fusion methods for video shot retrieval," in *International Conference on Content-Based Image and Video Retrieval (CIVR)*, Singapore, 2005.

[72] Xiaoyong Liu and W. Bruce Croft, "Cluster-based retrieval using language models," in *SIGIR*, Sheffield, South Yorkshire, UK, 2004.

[73] Hugo Liu, "Montylingua: An end-to-end natural language processor with common sense," in *http://web.media.mit.edu/∼hugo/montylingua*.

[74] Alias-i, "Lingpipe named entity tagger," in *http://www.alias-i.com/lingpipe/*.

[75] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau, "Okapi at TREC4," in *Text REtrieval Conference*, 1992, pp. 21–30.

[76] Winston H. Hsu and Shih-Fu Chang, "Visual cue cluster construction via information bottleneck principle and kernel density estimation," in *International Conference on Content-Based Image and Video Retrieval (CIVR)*, Singapore, 2005.

[77] Tat-Seng Chua, Shi-Yong Neo, Hai-Kiat Goh, Ming Zhao, Yang Xiao, Gang Wang, Sheng Gao, Kai Chen, Qibin Sun, and Tian Qi, "TRECVID 2005 by NUS PRIS," in *TRECVID Workshop*, Waishington DC, 2005.

[78] Jun Yang, Rong Yan, and Alexander G. Hauptmann, "Multiple instance learning for labeling faces in broadcasting news video," in *ACM Multimedia*, Singapore, 2005.

[79] Noam Slonim and Naftali Tishby, "Document clustering using word clusters via the information bottleneck method," in *SIGIR*, Athens, Greece, 2000.

[80] Wei-Ying Ma and B. S. Manjunath, "Texture features and learning similarity," in *Computer Vision and Pattern Recognition*, 1996.

[81] Reinhard Diestel, *Graph Theory*, Springer, 3 edition, 2005.

[82] Yuxin Peng and Chong-Wah Ngo, "Clip-based similarity measure for hierarchical video retrieval," in *ACM SIGMM international workshop on Multimedia information retrieval*, 2004, pp. 53–60.

[83] Yuxin Peng and Chong-Wah Ngo, "EMD-based video clip retrieval by many-to-many matching," in *International Conference on Content-Based Image and Video Retrieval*, Singapore, 2005, pp. 71–81.

[84] Timo Volkmer, John R. Smith, and Apostol (Paul) Natsev, "A web-based system for collaborative annotation of large image and video collections," in *ACM Multimedia*, Singapore, 2005, pp. 892–901.

[85] Milind Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis, "Large-scale concept ontology for multimedia," *IEEE Multimedia Magazine*, vol. 13, no. 3, July-September 2006.

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

[87] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, "A practical guide to support vector classification," Tech. Rep., National Taiwan University, 2005.

[88] Arnon Amir, Janne O Argillander, Marco Berg, Shih-Fu Chang, Martin Franz, Winston Hsu, Giridharan Iyengar, John R Kender, Lyndon Kennedy, Ching-Yung Lin, Milind Naphade, Apostol (Paul) Natsev, John R. Smith, Jelena Tesic, Gang Wu, Rong Yan, and Donqing Zhang, "IBM research TRECVID-2004 video retrieval system," in *NIST TRECVID 2004 Workshop*, Gaithersburg, MD, 2004.

[89] Belle L. Tseng, Ching-Yung Lin, Milind Npahade, Apostol Natsev, and John R. Smith, "Normalized classifier fusion for semantic visual concept detection," in *IEEE Internaltional Conference on Image Processing*, Barcelona, 2003.

[90] Kieran McDonald and Alan F. Smeaton, "A comparison of score, rank and probability-based fusion methods for video shot retrieval," in *International Conference on Content-Based Image and Video Retrieval*, Singapore, 2005, pp. 61–70.

[91] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, "The pagerank citation ranking: Bringing order to the web," Tech. Rep., Stanford Digital Library Technologies Project, 1998.

[92] Amy N. Langville and Carl D. Meyer, "A survey of eigenvector methods for web information retrieval," *SIAM Review*, vol. 47, no. 1, pp. 135–161, 2005.

[93] Gilbert Strang, *Linear Algebra and Its Applications*, Brooks Cole, 4 edition, 2005.

[94] Matthew Richardson and Pedro Domingos, "The intelligent surfer: Probabilistic combination of link and content information in pagerank," in *Advances in Neural Information Processing Systems*, Cambridge, MA, 2002.

[95] Taher H. Haveliwala, "Topic-sensitive pagerank," in *International WWW Conference*, Honolulu, Hawaii, USA, 2002.

[96] Marina Meila and Jianbo Shi, "Learning segmentation with random walk," in *Neural Information Processing Systems Conference (NIPS)*, 2001, pp. 873–879.

[97] Jianbo Shi and Jitendra Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[98] Andrew Y. Ng, Michael Jordan, and Yair Weiss, "On spectral clustering: Analysis and an algorithm," in *Neural Information Processing Systems Conference (NIPS)*, 2002.

[99] Jia-Yu Pan, Hyung-Jeong Yang, Christos Faloutsos, and Pinar Duygulu, "Gcap: Graph-based automatic image captioning," in *International Workshop on Multimedia Data and Document Engineering*, Washington, DC, USA, 2004.

[100] Xiaofei He, Wei-Ying Ma, and Hongjiang Zhang, "Imagerank : spectral techniques for structural analysis of image database," in *IEEE International Conference on Multimedia and Expo*, 2003.

[101] Craig A.N. Soules and Gregory R. Ganger, "Connections: Using context to enhance file search," in *SOSP*, Brighton, United Kingdom, 2005.

[102] Shih-Fu Chang, Winston Hsu, Wei Jiang, Lyndon Kennedy, Akira Yanagawa, Dong Xu, and Eric Zavesky, "Columbia University TRECVID-2006 video search and high-level feature extraction," in *TRECVID Workshop*, Waishington DC, 2006.

[103] Sepandar D. Kamvar, Taher H. Haveliwala, Christopher D. Manning, and Gene H. Golub, "Extrapolation methods for accelerating pagerank computations," in *International WWW Conference*, Budapest, Hungary, 2003.

[104] Gunes Erkan and Dragomir R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, vol. 22, 2004.

[105] Thorsten Joachims, "Transductive inference for text classification using support vector machines," in *16th International Conference on Machine Learning*, 1999, pp. 200–209.

[106] Avrim Blum and Tom Mitchell, "Combining labeled and unlabeled data with co-training," in *Annual Workshop on Computational Learning Theory*, 1998, pp. 92–100.

[107] Shahram Ebadollahi, Lexing Xie, Shih-Fu Chang, and John R. Smith, "Visual event detection using multi-dimensional concept dynamics," in *IEEE International Conference on Multimedia and Expo (ICME)*, Toronto, Canada, July 9-12 2006.

[108] Yiming Yang, Tom Pierce, and Jaime Carbonell, "A study on retrospective and on-line event detection," in *SIGIR*, Melbourne, AU, 1998.

[109] Zhiwei Li, Bin Wang, Mingjing Li, and Wei-Ying Ma, "A probabilistic model for retrospective news event," in *SIGIR*, Salvador, Brazil, 2005.

[110] Alexander G. Gray and Andrew W. Moore, "N-body problems in statistical learning," in *Neural Information Processing Systems Conference (NIPS)*, 2001, pp. 521–527.

[111] Carsten Rother, Lucas Bordeaux, Youssef Hamadi, and Andrew Blake, "Autocollage," *ACM Transactions on Graphics (SIGGRAPH Proceedings)*, vol. 25, no. 3, pp. 847–852, 2006.

[112] Noah Snavely, Steven M. Seitz, and Richard Szeliski, "Photo tourism: Exploring photo collections in 3d," *ACM Transactions on Graphics (SIGGRAPH Proceedings)*, vol. 25, no. 3, pp. 835–846, 2006.

[113] Alexandar Jaffe, Mor Naaman, Tamir Tassa, and Marc Davis, "Generating summaries and visualization for large collections of geo-referenced photographs," in *Multimedia Information Retrieval Workshop (MIR)*, Santa Barbara, CA, October 2006.

[114] Luis von Ahn, Ruoran Liu, and Manuel Blum, "Peekaboom: a game for locating objects in images," in *ACM CHI*, Montral, Qubec, Canada, August 2006.

[115] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman., "Learning object categories from google's image search," in *CVPR*, San Diego, 2005.

[116] Lyndon Kennedy, Shih-Fu Chang, and Igor Kozintsev, "To search or to label?: Predicting the performance of search-based automatic image classifiers," in *Multimedia Information Retrieval Workshop (MIR)*, Santa Barbara, CA, October 2006.

Appendix

A. Temporal Sequence Operators

During candidate determination, multi-modal fusion, or feature development in chapter 2, we apply many temporal operations on time instances. To explain with these operations, we define two temporal sequence operators, *OR* $\oplus_\epsilon$ (Equation 9.1) and *AND* $\odot_\epsilon$ (Equation 9.2), with a fuzzy window ϵ,

$$\begin{aligned} A \oplus_\epsilon B &= A \cup B - \{z | z \in B, \exists\ a \in A, |z - a| < \epsilon\}, \\ &= A \cup B - B \odot_\epsilon A. \end{aligned} \tag{9.1}$$

$$A \odot_\epsilon B = \{z | z \in A, \exists\ b \in B, |z - b| < \epsilon\}, \tag{9.2}$$

$A, B = \{z_n \in \mathbb{R}\}_{n \geq 1}$ represent two sequences of time points such as (i) and (ii) in Figure 9.1.

The "AND" operator locates those points from two time sequences coinciding with each other within a fuzzy window and keeps the time points from the first operand. An example is shown in (iii) of Figure 9.1. The "OR" operator combines time points from two sequences but removes the duplications from the second operand within the fuzzy window. An example is illustrated in (iv) of Figure 9.1. Both operators are not commutative.

We also need a filter operation to filter time instances within certain intervals

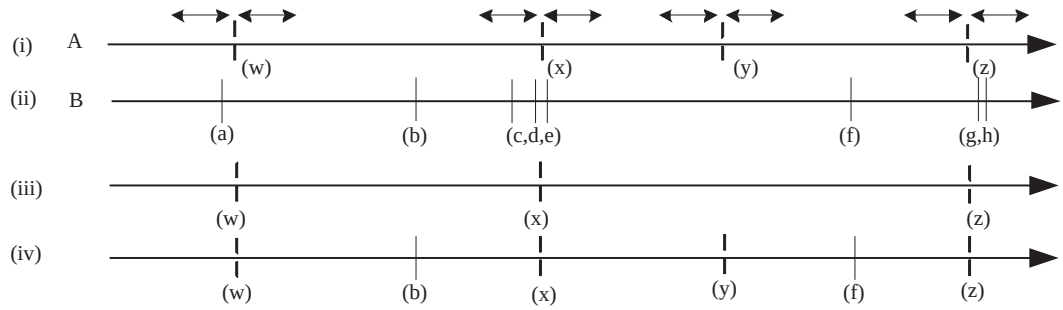


Figure 9.1: Example of temporal operators: (i) the sequence A is composed of time points $\{w, x, y, z\}$; (ii) B is the time sequence with points $\{a, ..., h\}$; (iii) the *AND* operation $A \odot_\epsilon B$ yields $\{w, x, z\}$; (iv) the *OR* operation $A \oplus_\epsilon B$ unifies sequences A and B into $\{w, b, x, y, f, z\}$ by removing duplications within the fuzzy window ϵ, indicated by "$\longleftrightarrow$."

with a fuzzy window ϵ. The filter function $\Psi_\epsilon(A, S)$ is defined in the following,

$$\Psi_\epsilon(A, S) = \{z | z \in A, \exists\ i, z \in [s_i - \epsilon, e_i + \epsilon], (s_i, e_i) \in S\}, \tag{9.3}$$

where $S = \{(s_i, e_i)\}$ is composed of segments defined by starting point s_i and end point e_i. The filter operation is to locate those points in A falling within the range of segments in S with a fuzzy window ϵ.

With these two operators and the filter, we could easily combine or filter the occurrence of features from different modalities. For example, we use $\oplus_\epsilon$ to combine audio pauses and shot boundaries as candidate points (Section 2.2.1). We also compactly represent the boundary performance metrics with $\odot_\epsilon$ in Equations 2.8 and 2.9. We later use the filter $\Psi_\epsilon(\cdot)$ to locate feature points in a specific region or state (Section 2.4.8). From our experiments, these operations are shown to be important for story boundary segmentation (Section 2.5.3).

B. Eigenvectors and Stationary Probability

Definition B..1. *Let λ_1, λ_2 , ..., and λ_n be the eigenvalues of an $n \times n$ matrix A, whose corresponding eigenvectors are $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_n$. λ_1 is called the* ***dominant (or principal) eigenvalue*** *of A if*

$$|\lambda_1| > |\lambda_i|, i = 2, ..., n$$

The eigenvectors corresponding to λ_1 are called the dominant (or principal) eigenvectors of A .

Definition B..2. *A random (Markov) process is governed by the* ***Markov (transition) matrix*** *where each column of the matrix adds up to 1 and the matrix has no negative entries.*

Theorem B..1. *A Markov matrix $\mathbf{A}$ has the following properties:*

(i) $\lambda_1 = 1$ *is an eigenvalue*

(ii) its principal eigenvector $\mathbf{u}_1$ is nonnegative and it is a steady (stationary) state since $\mathbf{A}\mathbf{u}_1 = \mathbf{u}_1$

(iii) the other eigenvalues satisfy $|\lambda_i| \leq 1$

(iv) if any power of $\mathbf{A}$ has all positive entries, then $|\lambda_i|, i = 2, ..., n$ are below 1. The product $\mathbf{A}^k\mathbf{x_0}$ approaches a multiple of $\mathbf{u}_1$, when k increases. Note that $\mathbf{x}_0$ is any arbitrary vector of n dimensions.

Proof. See proof in pp. 267, [93]. □

The above theorem states the existence of the stationary probability for the stochastic matrix and the convergence of the Power Method which can be utilized

to calculate the dominant eigenvector (i.e., stationary probability) of a Markov matrix. Meanwhile, the proof of the Theorem provides some insight into the rate of convergence of the Power Method. That is, if the eigenvalues of $\mathbf{A}$ are ordered so that

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_n|,$$

then the power method will converge quickly if $|\lambda_2|/|\lambda_1|$ is small, and slowly if $|\lambda_2|/|\lambda_1|$ is close to 1 [103].

Example B..1. *Calculate the stationary probability from a similarity matrix.*

If we have a discrete link graph $\mathbf{S}$ of 7 nodes, which indicates the relations among nodes $\{d_1, d_2, ..., d_7\}$. For example, node d_1 has links to node d_2, d_3, d_4, d_6. A value of 1 is used to represent the existence of a link.

$$\mathbf{S} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

After the normalization, which ensures that the entire row adds up to 1, we can

have the transition matrix,

$$\mathbf{P} = \begin{pmatrix} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

According to Theorem B..1, the stationary probability of $\mathbf{P}$ is actuallythe dominant eigenvector of $\mathbf{P}^T$ within a scalar vector. A common solution is to use off-the-shelf toolkits (e.g., MATLAB) to calculate the entire eigenvalues and then locate the dominant eigenvector. From the "eig" command of MATLAB, we can have six eigenvalues

$$|\lambda| = \{1.0, 0.50, 0.50, 0.32, 0.17, 0.0\},$$

where the dominant eigenvalue is 1 as proved by Theorem B..1. The corresponding dominant eigenvector is

$$\mu_1 = [0.6995, 0.3829, 0.3240, 0.2430, 0.4123, 0.1031, 0.1399]^T.$$

After the normalization, divided by its 1-norm $||\mu_1||_1$, we can derive the stationary probability

$$\mathbf{x}_\pi = [0.3035, 0.1661, 0.1406, 0.1054, 0.1789, 0.0447, 0.0607]^T.$$

For efficiency, we can get the dominant eigenvector through the Power Method

(cf. Section 8.2.4) and then derive the (normalized) dominant eigenvector

$$\mathbf{x}'_\pi = [0.3034, 0.1662, 0.1406, 0.1054, 0.1789, 0.0447, 0.0607]^T.$$

From this example, we can find that the both solutions $\mathbf{x}_\pi$ and $\mathbf{x}'_\pi$ are almost the same except some minor differences due to numerical rounding errors. Meanwhile, in this example, by checking the stationary probability $\mathbf{x}_\pi$, the most probable states for the random walk are $d_1 > d_5 > d_2 > d_3 > d_4 > d_7 > d_6$.