

**Multimodal News Summarization, Tracking and Annotation
Incorporating Tensor Analysis of Memes**

Chun-Yu Tsai

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017

ABSTRACT

Multimodal News Summarization, Tracking and Annotation

Incorporating Tensor Analysis of Memes

Chun-Yu Tsai

We demonstrate four novel multimodal methods for efficient video summarization and comprehensive cross-cultural news video understanding.

First, for video quick browsing, we demonstrate a multimedia event recounting system. Based on nine people-oriented design principles, it summarizes YouTube-like videos into short visual segments (8–12 seconds) and textual words (less than 10 terms). In the 2013 Trecvid Multimedia Event Recounting competition, this system placed first in recognition time efficiency, while remaining above average in description accuracy.

Secondly, we demonstrate the summarization of large amounts of online international news videos. In order to understand an international event such as Ebola virus, AirAsia Flight 8501 and Zika virus comprehensively, we present a novel and efficient constrained tensor factorization algorithm that first represents a video archive of multimedia news stories concerning a news event as a sparse tensor of order 4. The dimensions correspond to extracted visual memes, verbal tags, time periods, and cultures. The iterative algorithm approximately but accurately extracts coherent quad-clusters, each of which represents a significant summary of an important independent aspect of the news event. We give examples of quad-clusters extracted from tensors with at least 10^8 entries derived from international news coverage. We show the method is fast, can be tuned to give preferences to any subset of its four dimensions, and exceeds three existing methods in performance.

Thirdly, noting that the co-occurrence of visual memes and tags in our summarization result is sparse, we show how to model cross-cultural visual meme influence based on normalized PageRank, which more accurately captures the rates at which visual memes are reposted in a specified time period in a specified culture.

Lastly, we establish the correspondences of videos and text descriptions in different cultures by reliable visual cues, detect culture-specific tags for visual memes and then annotate videos in a cultural settings. Starting with any video with less text or no text in one culture (say, US), we select candidate annotations in the text of another culture (say, China) to annotate US video. Through analyzing the similarity of images annotated by those candidates, we can derive a set of proper tags from the viewpoints of another culture (China). We illustrate cultural-based annotation examples by segments of international news. We evaluate the generated tags by cross-cultural tag frequency, tag precision, and user studies.

Contents

Contents	i
List of Figures	v
List of Tables	xiii
Acknowledgements	xv
1 Introduction	1
1.1 Overview	1
1.2 News Dataset and Visual Memes	2
1.2.1 Dataset	2
1.2.2 Visual Memes	4
1.2.3 Near-duplicate Keyframes Clustering	4
1.3 Contributions	6
2 Related Work	9
2.1 Visual to Text Summarization	9
2.2 Multimodal Co-clustering	10
2.3 Multimodal News Analysis	11

2.4	Automatic Tag Recommendation	12
2.5	CCA-based Multimodal Embeddings	13
2.6	Linguistic and Cultural Differences in Human Annotation of Visual Content	14
3	Highly Efficient Video Summarization	17
3.1	Introduction	17
3.2	Efficient Multimodal Video Summarization Algorithm	19
3.2.1	People-oriented Design Principles	20
3.2.2	Video Recounting Pipeline	24
3.3	Results	30
3.4	Discussion	30
4	News Video Summarization by Latent Coherent Patterns	33
4.1	Introduction	33
4.2	PARAFAC Decomposition with Sparse Latent Factors	36
4.3	Algorithm	38
4.3.1	Feature Extraction and Data Tensor	38
4.3.2	Problem Formulation	40
4.3.3	Tensor Decomposition and Dense Tensor Selection	43
4.3.4	Adding soft constraints	45
4.4	Discussions	47
4.4.1	Extraction Accuracy	48
4.4.2	Higher-order Co-clustering	50
4.4.3	Constraints Weighting and Limitations	52

4.5	Conclusion	52
5	Cross-cultural Visual Meme Tracking	55
5.1	Introduction	55
5.2	Cross-cultural Visual Meme Influence and Tracking Algorithm	58
5.2.1	Visual Meme Clustering	58
5.2.2	Visual Meme Graph Construction and Influence Calculation	58
5.3	Discussions	60
5.3.1	Visual Meme Coverage in Different Cultures	61
5.3.2	Visual Meme Influence Correlation	61
5.3.3	Single Visual Meme Tracking	65
5.3.4	Visual Meme and View Count Correlation	66
5.4	Future Work	67
6	Cross-cultural Video Annotation	69
6.1	Culture-specific Tag Detection	69
6.1.1	Embedding Methods	72
6.1.2	Experiments	75
6.1.3	Conclusion	85
6.2	Cross-cultural Video Annotation	86
6.2.1	Data Collection	86
6.2.2	Document Clusters by Visual Similarity	88
6.2.3	Annotation Algorithm	91
6.2.4	Results and Evaluation Methods	100

CONTENTS

6.2.5	Discussion	108
6.2.6	Future work	109
7	Future Work	111
	Bibliography	113

List of Figures

1.1	Flow diagram for near-duplicate frame clustering, in order from (A) to (H). (A) We extract a feature vector (SIFT-BOF or VGG 19-layer) from each keyframe. (B) We concatenate these vectors into a single descriptor matrix, one vector per row. (C)(D) We build the FLANN index on keyframes by comparing the distances between vectors. (E)(F) We binarize distances between vector pairs by thresholding, to form a keyframe similarity graph. (G)(H) We generate clustering results by taking the graph's transitive closure.	5
3.1	An example of system output using the NIST MER recounting tool. The system has located the three minimal 4-second video segments whose semantic classifiers provide total coverage of user query terms, which are themselves displayed as keywords sorted by relevance. (Principles 1,2,8,9)	19
3.2	Semantic ontology of IBM classifiers.	20
3.3	Video browsing tool designed for MER user studies. Timelines illustrating presence of semantic concepts appear under zoomable thumbnails; concepts can be combined according to standard boolean connectives. (Principles 1,4,6)	21

3.4	The system pipeline reduces video data according to human preferences. The analysis starts with complete classifier scores at each sampled frame, focuses on event-specific terms, locates semantic temporal boundaries, trades off specificity for accuracy, selects most relevant semantic segments, then extracts minimal subsegments. (Principles 3,4,5,6,7)	23
3.5	A graph showing the frame-to-frame coherence measure of a video.	25
3.6	Trading off specificity (“reward”) against classifier confidence. Upper left shows part of the ontology tree; yellow nodes are event-specific concept nodes. Upper right is the corresponding keyframe. Bottom table shows ranking of concepts before and after the “rewarding”. (Principle 5)	28
3.7	Results of ten systems in the Trecvid MER competition. Horizontal axis is textual description precision (0=Fail, 5=Excellent); vertical axis is log(speed-up) of decision time compared with video length. Lines indicate two linear regressions (x vs y, y vs x); both show precision increases with speed-up. Our system is circled; speed up is by a factor of 6, the highest reported.	29
4.1	(a)(b) are examples of quad-clusters for the AirAsia Flight Q8501 event. 4-axis modalities: visual-memes, tags, time, and culture.	34
4.2	The PARAFAC decomposition of a three-dimensional tensor capturing visual, verbal, and temporal information in a news video collection.	36
4.3	Selecting and refining high-dimensional co-clusters from the without loss of generality in two dimensions.	38

4.4	This figure summarizes 2 culture-specific quad-clusters. One is Europe, and the other is U.S. Modalities: visual-memes, tags, time and culture.	44
4.5	(a) One of extracted Zika quad-cluster which is dominated by common tags. (b) Add soft constraints to visual-memes and tags with $w_v = 1, w_t = 0.5$. The memes are an exact subset, and the soft constraints have eliminated more common tags, compared to (a).	45
4.6	(a) Our data is a four-dimensional tensor, where each entry of the matrix counts the number of co-occurrences of $\mathbf{v}(i), \mathbf{s}(j), \mathbf{t}(m), \mathbf{c}(l)$. But additionally we can measure inter-video relationships, such as how \mathbf{v} or \mathbf{s} tend to co-occur across videos. (b) To model constraints, we create “virtual” tags and memes. We detect memes with high video-to-video centrality, for example, $\mathbf{v}(1)$. We then create a virtual tag that expresses this relationship to its “nearby” memes (the tag is: “this is the $\mathbf{v}(1)$ group”). We similarly create virtual memes based on high video-to-video tag co-occurrence.	46

- 4.7 Comparing tags of bi-clustering results to tags of tri-clustering results extracted from the same video dataset. (a)(b)(c) show the top 7 tags with their membership scores within the extracted bi-cluster and tri-clusters. (a) Bi-clustering only extracts 1 bi-cluster related to “Ebola virus spread in Sierra Leone”, and its tags are dominated by co-occurrence of most common tags in the system. (b)(c) However, with additional time information, tri-clustering extracts 2 tri-clusters related to “Ebola virus spread in Sierra Leone”. Each of them represents more specific news stories, in which common tags like “virus” and “outbreak” are split, and tags of subevents (“first”, “case”, “patient”, “escape”,...etc.) have higher membership scores. 50
- 4.8 Extracted quad-cluster which is constrained by a group of tags. (a) is a list of high-frequency tags, which co-occur with “quarantine”. We model a constraint on this group of tags by creating a virtual visual meme (Section 4.4). (b) Extracted quad-cluster by adding the constraint. (Here we show tags and visual memes only.) (c) Examples of YouTube videos where the extracted tags and visual memes come from. The images in (c) are keyframes, which are used to represent videos, not necessarily visual memes. 51
- 5.1 A partial illustration of a cross-cultural visual meme graph for the AirAsia Flight Q8501 event. Red nodes: visual memes occurring in U.S.. Blue nodes: visual memes in China. Red and blue nodes overlapping: visual memes in both countries. 56

5.2	Pipeline for generating cross-cultural visual meme graphs. (A) Video preprocessing. We collect videos from YouTube and Baidu, detect shots in each video, then visually cluster these shots, with each shot represented by a keyframe. A visual meme is defined as a collection of two or more near-duplicate keyframes. (B) Simplified visualization of visual meme graphs for the Ebola event, in a culture-versus-time matrix. Each graph displays in a circle the same visual meme nodes, but with only the edges present in culture c at time t , illustrating differences and evolution.	56
5.3	Cross-cultural visual meme influence correlation coefficients for different countries, for different news events, along time. Red curves are Pearson, blue curves are Spearman.	62
5.4	Timing of visual meme propagation. (a) Coverage of American Dr. Kent Brantly: the peak of the U.S. curve is ahead of Europe. (b) Coverage of British nurse William Pooley: the peak of the European curve is ahead of the U.S. . . .	64
5.5	Long-lasting visual meme clusters. (a) Airplane visual meme from AirAsia. (b) One of aid scenes from Ebola.	64
5.6	Log-log plot of visual meme influence (vertical) versus view count (horizontal), showing independence. Both influence and view count also follow log-normal distributions marginally.	67

- 6.1 Left: Two-view Pair-Pair embedding via (deep) CCA. Right: Three-view embedding via GCCA. In the Two-view Pair-Pair embedding, each circle represents an image, and each triangle represents a text description, as a pair is a (image, text) pairing. Here we map pairs from two different cultures (culture 1, culture 2) into a joint embedding space via CCA. In Three-view embedding, we project data of three views (texts in culture 1, texts in culture 2, all images) into a joint embedding space. 72
- 6.2 Near-duplicate keyframes across cultures with different texts. Pair (A) is from AirAsia Flight. Description of US image give more detailed information. Pair (B) is from Zika virus. Description of South America image takes Zika more seriously. 77
- 6.3 We use an image in culture 1 (say, U.S.) to query texts in culture 2 (say, China). We use the texts of near-duplicate images in culture 2 as ground truth to calculate the recall. 81
- 6.4 (a) Keywords co-occurring with “Ebola”, which is an event (E). “Kaci”, “quarantine”, “cdc”, “africa”, are each a possible keyword for a subevent (e_i). (b) Collocated keywords detected for five time-limited subevents. 87
- 6.5 (a) Shows a frame-based document cluster: it contains all $d_j = \langle v_j, t_j \rangle$ containing keyframes in the same keyframe cluster. (b) Shows a video-based document cluster: it contains all frame-based document clusters which have near-duplicate keyframes from the video noted by “*”. 89

- 6.6 (a) Text from Chinese or from English reference video-plus-text documents used to annotate a target video with no tags or text. (b) Text from Chinese video-plus-text used to annotate a target video already annotated with English text. 90
- 6.7 The processing pipeline of cross-cultural tag annotation. (A) Candidate phrases are extracted, and used as queries into an annotated image store. A multiclass SVM is trained to map visual features of these retrieved verification images into these phrases. (B) The SVM associates phrases with each keyframe of the target video. A keyframe's phrase is verified if that target keyframe finds a visual match to a retrieved verification image that also carries the phrase. . . . 91
- 6.8 Keyframes with assigned verifying phrases. **HIT(exact)** indicates algorithm found at least one exact image match among annotated images. **HIT(inferred)** indicates algorithm failed to find an exact image match, but recognized similar image semantics. **NOT** indicates neither was found. Google Translate result is provided in parens for each Chinese word. 94
- 6.9 Image pairs recognized as matching by our algorithm. First row: Keyframes. Second row: Annotated images returned by web search engine. Keyframes (a) (c)(e) and images shown by red circles in (b)(d)(f) are respective pairings. . . . 96

6.10	Keyframe verified by semantic match. (a) Keyframe in Ebola video. (b) Image set returned by query term, “charity medecins sans frontieres ebola”. There are no exact images matches, but the semantics (group of people in masks, etc.) are similar. (c) Keyframe in ISIS video. (d) Image set returned by query term, “launched 11 airstrikes overnight ISIS”. There are no exact image matches, but more than 50% of retrieved image are about airstrikes of the same place from different perspectives and with different lighting.	98
6.11	Examples of video annotation produced by our algorithm for four events: Ebola (top left), World Cup (top right), AirAsia (middle both) and ISIS (bottom both). Google Translate is provided for each Chinese tag, with a few corrections for proper names (e.g., “Ka Xige” to “Kassig”, etc.)	99
6.12	Keyframes from videos presented from the perspective of a different culture. The results are retrieved by our cross-cultural tags, from top 20 returns from Google video search. (a) Emory University Hospital from US media source. (b) News conference of National Search and Rescue Agency in Indonesia from Chinese media source. Neither segment appeared in the other country’s news coverage.	100

List of Tables

4.1	Average performance of bi-clusters extraction from sampled Ebola, AirAsia, Zika dataset by different algorithms. (a) Information-Theoretic co-clustering [6]. (b) Spectral co-clustering [21]. (c) Hierarchical co-clustering [40]. (d) Extended PARAFAC with sparse latent factor and dense tensor selection.	48
4.2	Average performance of bi, tri and quad-clusters extracted by extended PARAFAC algorithm from sampled Ebola, AirAsia, Zika dataset.	48
5.1	The average number of visual memes per day for Ebola and AirAsia news events.	61
6.1	Tag detection performance of two-view pair-pair embedding. EU: Europe, CH: China, SA: South America.	82
6.2	Tag detection performance of three-view embedding. EU: Europe, CH: China, SA: South America.	84
6.3	English candidate phrases extracted from news video archives by RAKE algorithm, from “Ebola” news on Sept. 3, 2014, using query: <i>British Ebola survivor William Pooley</i>	93
6.4	Left: Chinese candidate phrase sets extracted from news video archives by Jieba system, from “Ebola” news on Sept. 3, 2014, which is related to <i>British Ebola survivor William Pooley</i> . Right: Google translation into English, except “Plymouth” has been corrected to “Pooley”.	93

6.5	Average tag frequency per document for each event. Loss is calculated as $ origin - translated / origin $	102
6.6	The average goodness of tags, which were ranked 1~5.	105
6.7	Interchangeability of native tags with translated tags, for 20 videos of four international events. User evaluation of interchangeability was on a five-level likert scale, ranked 1~5.	106
6.8	The percentage of related videos retrieved by our cross-cultural tags that were in the top 10 returns from a Web search engine.	107

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor Prof. John R. Kender for the opportunity of my Ph.D study, for his patience and immense knowledge. John has been supportive and has given me the freedom to pursue various projects. He has also provided insightful discussions about the research. His guidance helped me in all the time of research and writing of this thesis.

Besides, I would like to thank the rest of my dissertation committee: Prof. Belhumeur, Prof. Feiner, Dr. Merler and Prof. Cannon, not only for their time and extreme patience, but for their intellectual contributions to my development as a scientist. Thanks Prof. Belhumeur for his insightful comments in my candidacy exam; Thanks Prof. Feiner for his interests in cross-cultural analysis; Special thanks to Michele for always introducing interesting researches to me and also being my committee of candidacy and proposal. To Prof. Cannon, whom I am most appreciative for agreeing to serve on the committee on short notice, and knowing he would probably have less than two weeks to read my thesis.

I would also like to express my gratitude to my parents, 陳惠芬 and 蔡淵黎, for their moral support and warm encouragements. Thanks for their confidence in me has enhanced my ability to complete my degree in the end.

Finally, I would like to thank National Science Foundation (NSF) and IARPA (Intelli-

ACKNOWLEDGEMENTS

gence Advanced Research Projects Activity) for grants that made it possible to complete this thesis. Research in Chapter 3 is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20070; Research in Chapter 4 is supported by the National Science Foundation under Grant Number IIS-1513853. I would like to thank Taiwan government in the form of national scholarship, which enabled me to complete the degree.

Chapter 1

Introduction

1.1 Overview

With the proliferation of devices for capturing and watching videos, video hosting services have gained an enormous number of users. For example, almost one third of the people online use YouTube to upload or review videos [3]. This increasing popularity of Internet videos has accelerated the demand for efficient video browsing and retrieval, the focus of this thesis.

After we review the state of the art in Chapter 2, we show in Chapter 3 the results of experiments on a large collection compiled by NIST (National Institute of Standards and Technology) as part of their Multimedia Event Recounting (MER) evaluation. This work focused on generating a textual human-understandable description for multimedia clips, in order to enable video quick browsing. We design a semantic video browser tool to conduct user studies, and then present nine people-oriented design principles for video summarization. Based on these principles, we developed a recounting pipeline that summarizes 3-min long videos into short clips (8–12 secs) with short human-understandable description (less than 10 terms).

Chapters 4–6 are related to news video analysis. News is one of the most important

categories of online videos, and news videos provide huge amounts of information in our daily life. They are created, remixed, tagged, and maintained on social video repositories. Thus, algorithms are needed to provide viewers quick and comprehensive understanding. We demonstrate algorithms for news analysis in the following categories:

- News video summarization: To distill the most important information from news video sources to produce an abridged version for a particular user or task.
- News video tracking: To identify content that spreads widely and then fades over time scales on the order of days—the time scale at which we perceive news and events. We therefore extend video tracking to a cultural setting.
- News video annotation/tagging: To access the large number of news videos. Videos can only be effectively accessed if the metadata describing them are available in user-friendly structure. We therefore extend video annotation to a particular cultural setting.

1.2 News Dataset and Visual Memes

In order to better demonstrate algorithms for news analysis, in this section we introduce the news video dataset used in our research, and introduce the term “visual memes” used in this thesis.

1.2.1 Dataset

We collected dataset for three international news events:

1. Ebola Virus: News coverage of the West African Ebola virus epidemic outbreak (2013-2016) and its effects around the world. The West African Ebola virus epidemic was the most widespread outbreak of Ebola virus disease (EVD) in history—causing major loss of life and socioeconomic disruption in the region, mainly in the countries of Guinea, Liberia, and Sierra Leone.
2. AirAsia Flight 8501: News coverage of AirAsia Flight 8501 crash. The flight was a scheduled international passenger flight, operated by AirAsia Group affiliate Indonesia AirAsia, from Surabaya, Indonesia, to Singapore. On 28 December 2014, the aircraft crashed into the Java Sea during bad weather, killing all 155 passengers and seven crew on board.
3. Zika Virus: News coverage of Zika Virus outbreak. In early 2015, a widespread epidemic of Zika fever, caused by the Zika virus in Brazil, spread to other parts of South and North America. It is also affecting several islands in the Pacific, and Southeast Asia.

They are long-term news events lasting 2 months to 1 year. For Ebola news event, we have collected about 3100 videos and their metadata, in an approximate 3:1 (US:Europe) ratio, in a date range from 8/21/14 to 11/30/14. For AirAsia Flight 8501 events, we have collected about 1000 videos and their metadata, in an approximate 1:1 (China:US) ratio, in a date range from 12/28/14 to 1/15/15. For Zika Virus, we have collected about 1700 videos and their metadata, in an approximate 7:10 (South America:US), in a date range from 12/01/15 to 2/15/16. Videos sourced from US, Europe, and South American were collected from YouTube, and we verified their posted location in the metadata. Videos

sourced from China were collected from Baidu, the biggest Chinese video search engine in the world, which aggregates videos from Chinese online news channels and from Chinese video sharing websites.

1.2.2 Visual Memes

The definition of “visual memes” comes from the term “memes”. A meme is defined as a cultural unit (e.g., an idea, value, or pattern of behavior) that is passed from one person to another in social settings. News tracking [49] [64] [87] has been an active research topic. In the literature, a “meme” is a frequently re-posted phrase, and a “visual meme” is a frequently re-posted video segment or image. They both act as signatures of topics and events, and their propagation and diffusion over the web has been widely used to monitor the lifespan of a news story.

We define a visual meme as a frequently re-posted news video segments which was first introduced in [87]. Similar to memes, visual memes act as signatures of topics and events, and their propagation and diffusion over the web has been widely used to monitor the lifespan of a news story. The following we demonstrate how we cluster frequently re-posted news video segments as visual memes.

1.2.3 Near-duplicate Keyframes Clustering

We first represent each video segment as a keyframe and use a variant of near-duplicate detection between the keyframes of one video against the keyframes of another. To avoid the cost of the full $\mathcal{O}(N^2)$ matches, we restrict the keyframes to high-entropy I-frames, and

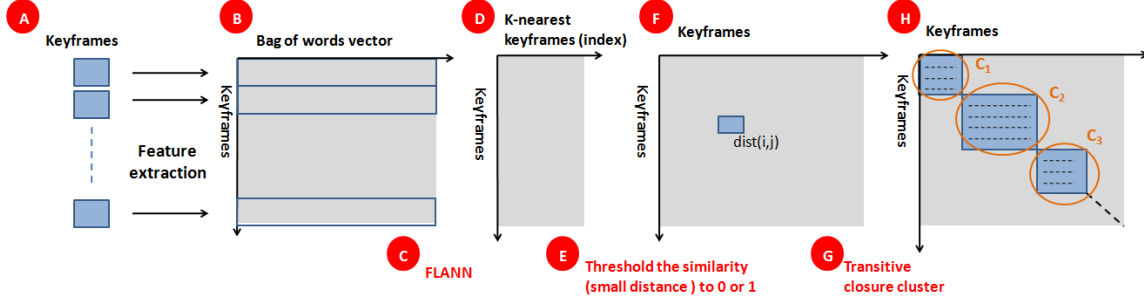


Figure 1.1: Flow diagram for near-duplicate frame clustering, in order from (A) to (H). (A) We extract a feature vector (SIFT-BOF or VGG 19-layer) from each keyframe. (B) We concatenate these vectors into a single descriptor matrix, one vector per row. (C)(D) We build the FLANN index on keyframes by comparing the distances between vectors. (E)(F) We binarize distances between vector pairs by thresholding, to form a keyframe similarity graph. (G)(H) We generate clustering results by taking the graph’s transitive closure.

select one I-frames per shot. (Most such videos are in mp4 format, which makes this easy.)

As shown in Figure 1.1, we then extract visual features (SIFT-BOF or 19-layer VGG [72]), and store each feature vector, from each of these keyframes, from each video in event, as a row in a descriptor matrix, which accumulates a total of m_i such rows. Using the FLANN library [59], we find the K nearest feature vectors to each feature vector, where $K = \sqrt{m_i}$, further limiting time complexity without compromising accuracy.

We now record these distances to these nearest neighbors in a keyframe-to-keyframe similarity matrix, which is then binarized via thresholding to yield a keyframe similarity graph. Its transitive closure is computed via a union-find algorithm [29] to find rough equivalence classes of near-duplicate keyframes. Then, each cluster of near-duplicate keyframes is a visual meme in our system.

1.3 Contributions

In this thesis, we demonstrate four novel multimodal methods for efficient video summarization and comprehensive cross-cultural news video understanding. The major contribution of this thesis are summarized here:

1. For video quick browsing, this thesis is the first to demonstrate a video recounting algorithm that generates multimedia clips with a human-understandable description based on user semantic preferences. These generated descriptions can be at various levels of detail, including a label for the overall theme, but also for important critical components. Our system pipeline reduces video data according to human preferences learned from user studies. The analysis starts with complete classifier scores at each sampled frame, focuses on event-specific terms, locates semantic temporal boundaries, trades off specificity for accuracy, selects the most relevant semantic segments, then extracts minimal subsegments.
2. For news video summarization, this thesis presents an efficient constrained tensor factorization algorithm for integration of multimodal data, and more specifically of language and vision data. This thesis also addresses the issue that both tags and visual memes are polysemic. By representing news videos as four-dimensional tensors (visual memes, tags, time stamps, and cultures), this thesis demonstrates an algorithm that can extract quad-clusters for better understanding news events.
3. For news video tracking, this thesis extends the tracking problem into a cultural setting. Given the fact that visual memes are frequently remixed and reposted in different cultures, we define cross-cultural visual memes “influence index”, which can

captures the rates at which visual memes (video segments) are reposted in a specified time period in a specified culture. We also showcase cross-cultural visual meme tracking by using two specific cultures for each of two specific news events (Ebola and AirAsia), over a relatively short specific expanse of time.

4. For news video tag analysis and video annotation, this thesis extends the problem into a cultural setting, too. We first show that tags in different cultures have significant usage differences, and then we demonstrate an algorithm, which annotates the entire news video from different cultural points of view. We evaluate the annotations by cross-cultural tag frequency, tag precision, and user studies.

Chapter 2

Related Work

This thesis is related to several research fields: multimodal video summarization, co-clustering, constrained matrix factorization, multimodal news analysis, visual meme tracking, multimodal embedding and cross-culture analysis. Below, we review a few representative works in six specific related domains: visual to text summarization, multimedia co-clustering, news event analysis, automatic tag recommendation, multimodal embedding and cultural differences in human annotation of visual content.

2.1 Visual to Text Summarization

Precise visual content description is useful for many applications such as search and browsing. Commercial search engines have been using some content-relevant texts, such as those extracted from captions/contexts of online videos and scripts of movies and TV series. There are many researches in this category. Tan *et al.* [77] explore the technical feasibility of the challenging issue of precisely recounting video contents, and use both visual and audio concepts to summarize videos. Then, various research papers focus on generating accurate and short sentence description [7][18][44][92]. Besides, Sadovnik *et al.* [67] develop method for creating the most efficient textual description that can discriminate one image from a group of images by ranking the different items in the target image by discrim-

inability (category, relationship and color) and salience. In this thesis, we demonstrate an efficient video summarization algorithm that helps users rapidly understand video contents.

2.2 Multimodal Co-clustering

Co-clustering algorithms are a kind of unsupervised clustering, which have been used to mine the latent relationships between different variables[21][20][22][69]. For example, bi-clustering (two-way co-clustering) simultaneously groups the rows and columns of a matrix to produce coherent submatrices. In nearly all research fields, there are relatively few papers on tri-clustering, or higher dimensional co-clustering.

In multimedia research, virtually all work is on bi-clustering, with few exceptions. In [85], Xiao *et al.* performed story clustering by exploiting the duality between textual and visual concepts through spectral co-clustering. Cai *et al.* [14] introduced information theoretic co-clustering to exploit potential grouping trends among low-level acoustic features and mid-level representations like audio effects; this provided a more accurate similarity measure for comparing auditory scenes. Essid *et al.* [26] applied a Nonnegative Matrix Factorization (NMF) algorithm to histograms of audio-visual feature counts, first to jointly discover latent structuring patterns, and then to track their activations along time. Irie *et al.* [35] proposed a novel Bayesian co-clustering model that jointly estimates the underlying distributions of image descriptors and textural words. However, none of this work has considered co-clustering in dimensions higher than two.

Work more related to this thesis is that of Lin *et al.* [52], in which they analyze multirelational structures in social media streams. However, they do not target news events, nor do

they deal with sparse data. Recently, Chu *et al.* [16] has proposed a video co-summarization technique that exploits visual co-occurrence across multiple videos. They avoid the difficulty that traditional spectral co-clustering algorithms fail to capture sparsely co-occurring shots. Instead, they model clustering as a problem of finding maximal bi-cliques, which is less sensitive. Since our data tensor is sparse, too, and has four modalities, we propose and extend a related but more efficient parallel factorization algorithm [63].

2.3 Multimodal News Analysis

There are many possible references for news tracking and analysis, and many of them use visual and textual features together. Wu *et al.* [84] measured novelty and redundancy in cross-lingual broadcast news by using near-duplicate visual detection, named entity matching, and other textual information. Dumont *et al.* [25] extracted visual, textual, and audio features, and a temporal context, to develop a TV news shot boundary detector. Jou *et al.* [39] extracted who, what, when, and where from news data, and studied temporal trends of news topics. Messina *et al.* [57] proposed Hyper Media News, which is a fully automated platform for the large scale analysis and production of multimodal news content. Wang *et al.* [86] proposed the integration of multimodal features (lexical, audio, and video) using Conditional Random Fields for the segmentation of broadcast news stories. Similar to these researches, we propose to use the multimodal features of image, text, time, and culture. However, our goal is to mine latent coherent patterns from high-dimensional multimodalities, and all dimensions are handled uniformly and in parallel.

2.4 Automatic Tag Recommendation

Tag recommendation systems can be roughly categorized into model-based and data-driven approaches [91]. A model-based approach is useful when there are no human annotated data. It defines a set of visual concepts, and proceeds as a concept detection task [37, 50, 66]. There is much (monocultural) work in this area. As examples, Qi *et al.* [66] first used binary classification to detect each individual concept in a concept set, and then fused multiple concepts. Li *et al.* [50] trained hundreds of semantic concept classifiers, using example pictures from each concept, by statistical modeling and optimization techniques, and performed high speed annotation for online pictures. Jiang *et al.* [37] proposed active context-based concept fusion, for effectively improving the accuracy of semantic concept detection in images and videos. However, model-based annotating methods with predefined visual vocabularies usually are monolingual only and can not recognize in advance what the human concerns are in a cross-cultural setting.

A data-driven approach is based on visual content similarity. There is much (monocultural) work in this area. Li *et al.* [51] proposed a neighbor voting algorithm for image tagging, which learns tag relevance by accumulating votes from visual neighbors. Siersdorfer *et al.* [70] used visual redundancies to connect videos, and proposed several tag propagation methods for automatically obtaining richer video annotation. Zhao *et al.* [93] provided a novel solution for fast near-duplicate video retrieval, and showed the effectiveness of this classifier-free approach. However, they did not use additional online annotated images or text to verify and further infer cross-cultural tags.

Other efforts neither train concept models nor match visual content similarity (again

monoculturally). Yao *et al.* [91] explored user search behavior through click-through data, which is largely available and freely accessible by search engines. He learned video relationships, and applied these relationships, for an economic way of annotating online videos. Sigurbjörnsson *et al.* [71] selected the top-N co-occurring tags of a candidate tag, and re-ranked them based on a number of attributes (co-occurrence, stability, descriptiveness). Sun *et al.* [76] built tag relation graphs from the collective knowledge embedded in the tag-tag co-occurrence pairs, and recommended the tags by leveraging IR techniques. Larson *et al.* [48] extracted tags from video metadata and speech, and used IR divergence models to order them.

2.5 CCA-based Multimodal Embeddings

Canonical correlation analysis (CCA) and its kernel version (KCCA) maximize the correlation in the latent space. Hodosh *et al.* [32] directly employs KCCA for matching images and captions. In [30], Gong *et al.* builds two layers of CCA: the first layer transfers information from a large external dataset with 1 million image-caption pairs, and the final latent space is learned in the second layer of CCA. Andrew *et al.* [4] extend CCA in deep learning frameworks; Yan *et al.* address the issues in DCCA in order to produce state of the art performance on matching images and text [89]. Generalized canonical correlation analysis (GCCA) was proposed to learn embeddings for more than two-views of data. Benton *et al.* [9] use weighted GCCA to learn vector representations of social media users that best accounts for all aspects of a user's online life; they were then able to identify users who behaved similarly.

2.6 Linguistic and Cultural Differences in Human

Annotation of Visual Content

As noted by Popescu *et al.* [65], although the total number of multimedia sources on the Web exceeds several billions, the query space is unequally covered, especially for languages other than English. People tend to annotate the multimedia source by their native languages. To solve this problem, they proposed the MLFLICKR system, which is a multilingual query platform over FLICKR. First, they translated the query into different languages, and further verified the returned result by their visual content. In [11], Bergsma *et al.* noted that users naturally label their images as they post them online, providing an explicit link between a word and its visual representation. Because images are labeled with words in many languages, they were able to generate word translations by finding pairs of words that have a high visual similarity between their respective image sets. Clough *et al.* [17] indicated that the language used to express the associated texts or textual queries should not affect retrieval. He participated in the CLEF cross-language image retrieval campaign (ImageCLEF) to explore the use of both text and content-based retrieval methods for cross-language image retrieval. Besides the language differences, the literature in perception and cognition suggests that people in different cultures allocate attention differently when viewing image and videos. Dong *et al.* [24] found that for Europeans or Americans, the first tag was more likely assigned to the main objects than by Chinese; but for Chinese, the first tag was more likely assigned to the overall description or relations between objects in the images. However, this work did not compare cross-cultural semantic differences in tag selection.

To the best of our knowledge, our proposed work is the first to consider linguistic and cultural differences together in candidate tag selection, and to automatically generate cross-cultural tags for the videos of difference language sources.

Highly Efficient Video Summarization

3.1 Introduction

In order to help users rapidly understand video contents, there has been a growing focus on multimedia event recounting. This provides visual and textual evidence, such as short clips and phrases, for a video's human event content, even for videos without any captions, or human annotated text.

Unlike traditional video summarization techniques which aim to produce a complete and concise summary, multimedia event recounting only highlights important human-event content, usually a small fraction of the entire video. Watching and reading these recountings can facilitate and speed the human search process. However, providing efficient recountings (clip selection and text formation) with a good match to human preferences and limitations is still difficult.

Previous works have indicated that using natural language to formalize video semantics can help user gain useful information relevant to their demands. Khan *et al.* [45] and Barbu *et al.* [7] proposed algorithms to produce sentential descriptions for video using pre-defined sentence generating rules. One of the 2011 Multimedia Challenges proposed by industry sponsors was to automatically describe a video showing an excerpt of a public event. The

TRECVID 2012 evaluation added a new task, Multimedia Event Recounting (“MER”), which aims to produce a recounting that summarizes the key evidence of the event for each clip that a parallel Multimedia Event Detection task system deems positive. Tan *et al.* [77] explored the technical feasibility of precisely recounting video contents, and they found that the current visual and audio concept classification is able to provide very useful clues, although they did not explore the time costs involved. A topic-oriented multimedia summarization system proposed by Ding *et al.* [23] summarized the important information in a video belonging to a certain topic by using topic-oriented visual and audio concepts. In [92], Yu *et al.* demonstrated a recounting approach that recovers the contribution of each semantic evidence towards the event classification, but again without explicit regard to any cost measure.

Different from those works mentioned above, our recounting approach is based on nine people-oriented design principles, derived from several user studies conducted by four researchers. The result is a highly efficient MER, attaining a sixfold speed-up in time to video recognition and decision, without sacrificing quality of textual description. The system is based on a purely semantic approach to video, in which visual events are judged by more than 1400 semantic concept classifiers. These are organized in an ontology tree, divided into different “facets” (subtrees) that reflect different categories of linguistic “thematic roles”. Each subtree refines the semantics from general to specific (ex:Activity → People Activity → Demonstration), allowing great flexibility in describing video contents.

For each video, our system selects a small number of significant concepts that a user has specified are important descriptors of an event, and we evaluate the corresponding concept classifiers on a sampling of frames of the video. The video is therefore summarized as a

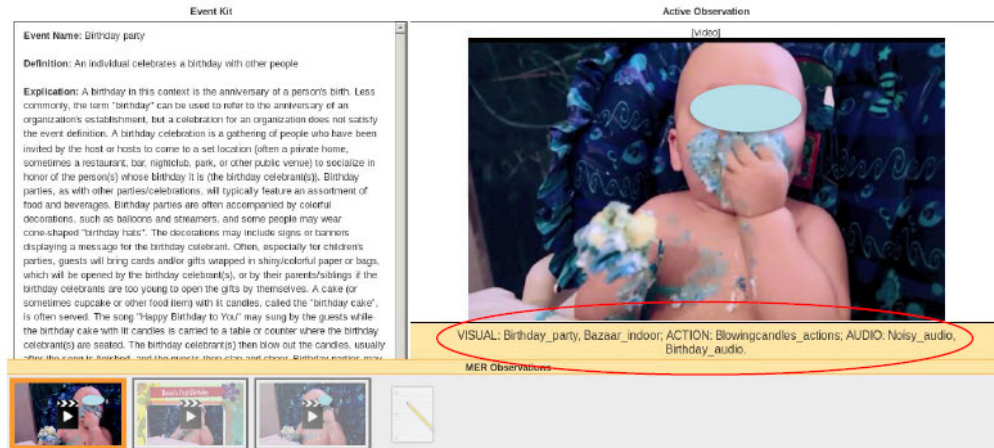


Figure 3.1: An example of system output using the NIST MER recounting tool. The system has located the three minimal 4-second video segments whose semantic classifiers provide total coverage of user query terms, which are themselves displayed as keywords sorted by relevance. (Principles 1,2,8,9)

matrix of (some) concept scores over (some) frames, typically of about 20 concepts over about 75 frames for a typical 2.5 minute video. We use this matrix to segment the video, by analyzing frame-to-frame semantic “coherence”, based on studies of user preferences and limitations. A greedy algorithm then selects a small sorted list of representative segments. Based on studies of human attention, we select 4-second snippets from each selected segment, and then produce the final recounting by listing in order the relevant concept names that appear within it. Figure 3.1 is an example of our recounting results.

3.2 Efficient Multimodal Video Summarization

Algorithm

Our system is based on the close connection between IBM ontology of semantic classifiers (Figure 3.2) and the functional aspects of natural language, and is informed throughout by

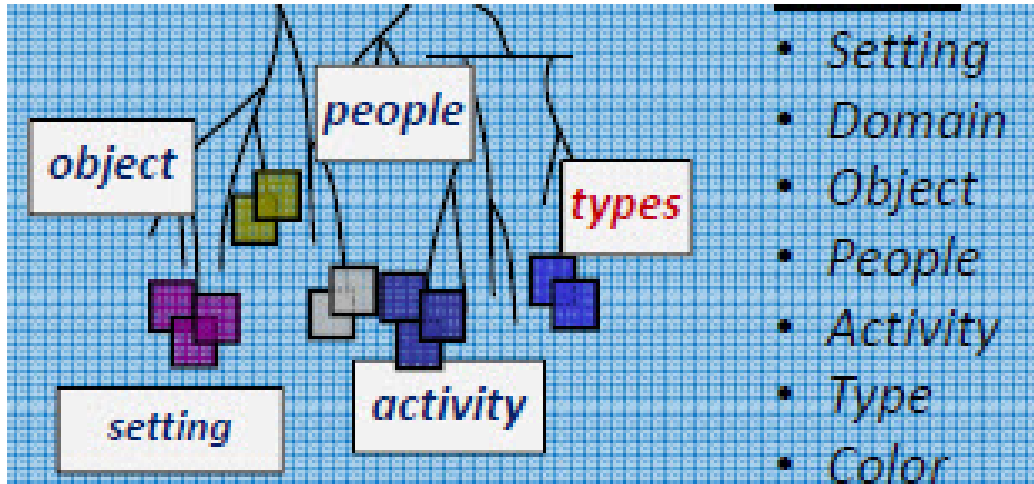


Figure 3.2: Semantic ontology of IBM classifiers.

human psychology and user studies of visual perception. We present nine people-oriented design principles based on user studies, followed by the processing pipeline based on them. This work was assisted by a semantic video browser tool, that accelerated our user studies about human preferences for video data reduction; see Figure 3.3.

3.2.1 People-oriented Design Principles

1. **Semantic ontology.** People tend to view and describe the world in a number of categories and at various levels of generality. Our visual, sound, and motion classifiers are part of an ontology tree that has approximately 1400 concepts[15]. This tree is actually a forest of “facets”, each of which reflecting a category of thematic roles, such as “people”, “actions”, “setting”, and each has multiple levels of specificity (“animal”, “vertebrate”, “mammal”, “dog”, etc.)
2. **Recounting is not Detection.** People generally do not edit YouTube-like videos; there is usually both great redundancy and few shot boundaries. We therefore sepa-

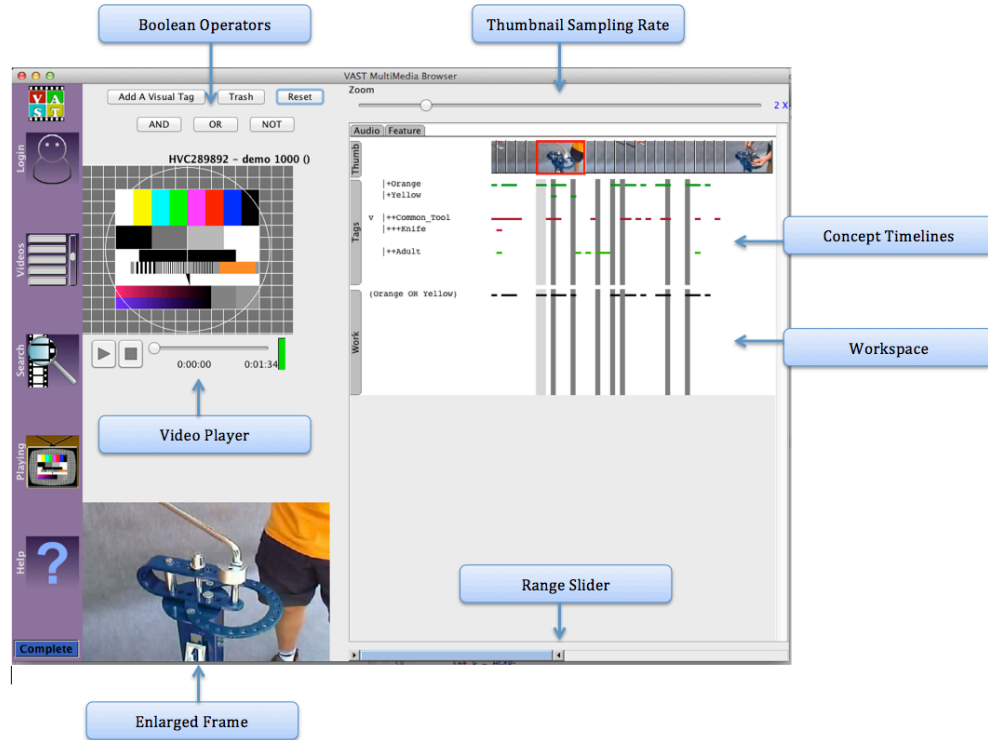


Figure 3.3: Video browsing tool designed for MER user studies. Timelines illustrating presence of semantic concepts appear under zoomable thumbnails; concepts can be combined according to standard boolean connectives. (Principles 1,4,6)

rated the detection task, which finds examples of events based on global information distributed over the entire video, from recounting task, which persuades people of the correctness of detection based on local information that is usually perceivable within seconds, and usually through short/visual segmentation. In [2], Multimedia event detection is not multimedia event recounting.

3. **Small queries.** People tended to use short query strings in our video library browser, and not the full 1400 concepts our system has available. They rarely used proper nouns. So, our recounting describes video contents only using the concepts given in the query strings. Our system therefore begins with a very lightweight matrix of classifier scores to process: usually between two and 30 rows, one per classifier, and

about 75 columns, one for every two seconds of video.

4. **Semantic segmentation.** People’s attention span is limited by short term memory, and this is reflected in well-documented distributions of edited shot lengths, of about 4 seconds. But even in YouTube videos, shot-like semantic segments can be found that are defined by temporal clusters of event-specific semantics, which correspond to units of attention. Our segmentation of videos for recounting is therefore purely based on semantic coherence.
5. **Description focus.** People have a good sense on how to trade off specificity against accuracy. We have devised and tuned an algorithm for doing so, based on a measure of the probability of classifier reliability, and on an approximate measure of information gain within the ontology tree.
6. **Video segment amount and rank.** People tend to remember events differentially, by what makes them distinguished from other events. This often means that the second best video snippet is more salient to a viewer, since the first best snippet tends to win on points, but not on uniqueness. We have tuned our video segment sorting and selection algorithm to reflect this.
7. **Quick subsequences.** People are really good at understanding imagery. Through experimentation, we found that video snippets rarely needed more than four seconds to be an effective proof of event; usually it uses just two. Our algorithm currently finds the best four seconds in a semantic segments to serve as the snippet, although we are aware this is only a heuristic approximation.
8. **Short text.** People have a limited ability to absorb textual concepts. We determined a small fixed limit (seven) to the number of concept words in each textual description.

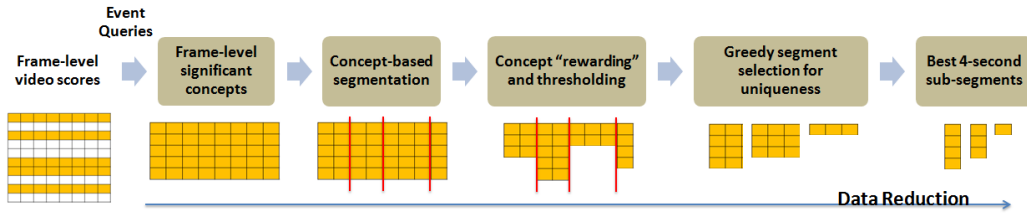


Figure 3.4: The system pipeline reduces video data according to human preferences. The analysis starts with complete classifier scores at each sampled frame, focuses on event-specific terms, locates semantic temporal boundaries, trades off specificity for accuracy, selects most relevant semantic segments, then extracts minimal subsegments. (Principles 3,4,5,6,7)

9. **Textual output.** People tend to ignore “function” words (articles, prepositions, auxiliary verbs, etc.) when they read. So, our text output generator also does. Each semantic segment with high classifier responses generates a short word list for each snippet, and uses typography to make its point. A typical output, for example, is: “VISUAL: Birthday_party, Bazaar_indoor; ACTION: Blowingcandles_actions; AUDIO: Noisy_audio, Birthday_audio.”

3.2.2 Video Recounting Pipeline

Based on the above nine design principles, our video recounting pipeline is illustrated in Figure 3.4.

A. Frame-level significant concepts

It is costly to use all 1400 potential concept classifiers to represent each video. Most concepts have little to do with a specific video, and some may fire in error. As stated in the design principle of small queries, we only use those significant concepts selected by the user during the search process. Typically, users specify less than 1% of the available concepts during their search, and our frame-level score matrix for each video immediately reflects these preferences. We note our textual recounting of a video is immediately a subset of a user’s own query terms, which aids in the rapid recognition by a user of a video’s relevance. At this point, we have reduced the full video to one describable by a small number of terms.

B. Concept-based segmentation

According to the design principle of semantic segmentation, we know that people’s attention spans are limited, and focused on the semantic coherence of a shot or shot-like partition. Therefore, we take advantage of the memory-based segmentation model in [43] and apply it to the video’s frame-level concept information. First, we calculate frame-to-frame coherence

$$Coherence(f_i, f_j) = \left(1 - \frac{d_{f_i, f_j}}{C}\right) e^{\frac{-|f_i - f_j|}{\sigma}} \quad (3.1)$$

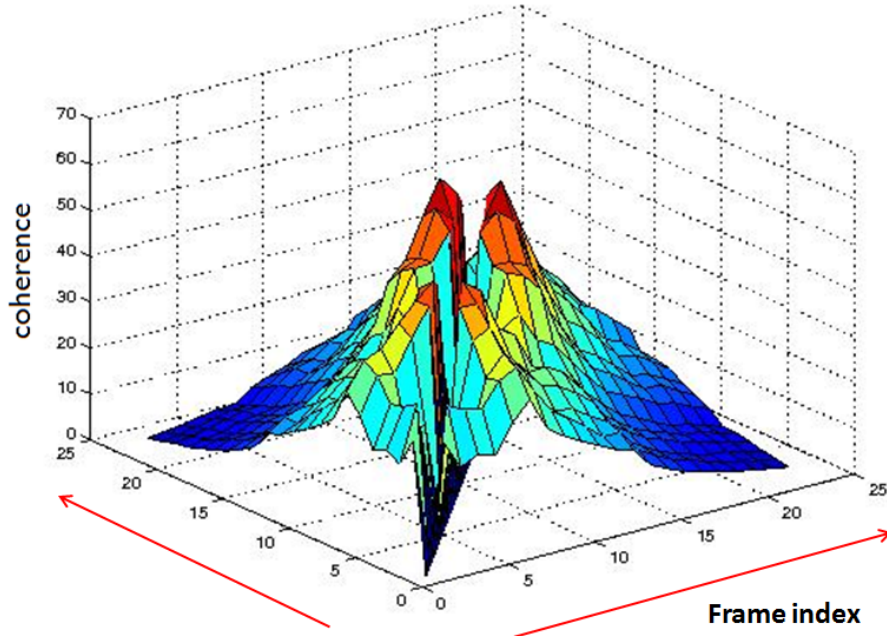


Figure 3.5: A graph showing the frame-to-frame coherence measure of a video.

where f_i, f_j are frame's index (one frame sample every 2 seconds), d_{f_i, f_j} is the L1 distance of concept vectors between f_i and f_j , C is the number of concepts and σ is a parameter to smooth the coherence according to the f_i, f_j 's chronological distance and the limits of viewers' short-term memory. (Figure 3.5) Then we sum the total frame-to-frame coherence across each frame boundary and recognize local minimums as potential scene boundaries.

Although videos with high production values and professional editing tend toward shorter shot lengths, the usual YouTube video is both low in visual quality and poor in temporal structure. In experimenting with different σ values, we found that it was necessary in this domain to set σ to 7 frame samples, which is approximately half as large than what was previously reported, reflecting the relatively lighter cognitive load. The exploration empirically found that as the model of the short-term memory

buffer model increased in size, semantic segments stabilized at about 28 seconds of memory recall; beyond that, the resulting average number of segments per video did not change. Empirically, this is large enough to detect significant semantic changes, without ignoring important semantic details. Because this smoothing still can give small local minima, we suppress minima within a window size ± 3 frames (a window of about 14 seconds). This ensure that there is, on average, only one semantic segment boundary in this 14-second window. Ground truth on a number of videos in a number of event categories showed that videos tended to have an average segment length of about 15 to 17 seconds. At this point, we have reduced the video to one represented by a small number (on average, about 10) of segments, each describable by a small number of terms.

C. Concept “rewarding” and thresholding

We have noted that a rich concept base tends to be have many correlations between concepts, but with subtle differences in the level of specificity. For example, in a “Town Hall Meeting” event video, the concept classifiers in the Activity ontology subtree (Activity, then People_activity, then Demonstration) correlate highly. A human recounting would not report all of these concepts; people innately know how to select amongst these concepts properly. The design principle of description focus indicates that people were seen to both search for, and describe, videos using “middle level” concepts, and that they prefer to describe videos using accurate super-concepts (“definitely animal”) over approximate sub concepts (“maybe dog”). To trade-off specificity and accuracy, and based on the fact that concepts at leaves of ontology give more information, we borrow the “reward” idea mentioned in [19], al-

though their paper does not address an ontology as rich as ours. Through user studies, we were able to re-formulate their “rewarding equation” to adjust our concept scores to better reflect user preferences. We reward both more specific concepts, and more strongly detected concepts:

$$ConceptScore = (Reward + \lambda) * Prob(concept) \quad (3.2)$$

For our ontology tree, we choose $\lambda = 2$, and the Reward is given by a concept’s tree depth (although it can be modeled by any other monotonically increasing function, such as information gain). After re-calculating each concept score in the segment, we sort the adjusted values, pick a threshold, and throw out the concepts whose scores are below this threshold value. Figure 3.6 describes the results after the “rewarding” stage. At this point, each segment is describable by terms that imitate human specificity preferences.

D. Greedy segment selection for uniqueness

We next want to find the minimum set of segments that covers relevant specific concepts. Since this problem is known to be NP-complete, we apply a greedy algorithm. Let C_e be the set of relevant concepts of event e , R_e be the set of concepts from selected segments, and C_s be the set of concepts in a segments. Greedy iteration:

- a) Select the segment with the most number of concepts in $C_e - R_e$. (Break tie : give priority to the one with higher average concept scores.)
- b) $R_e \leftarrow R_e \cup C_s$. If $C_e \neq R_e$, repeat from (a).

Generally, the covering set is very small, as most videos are highly redundant. At

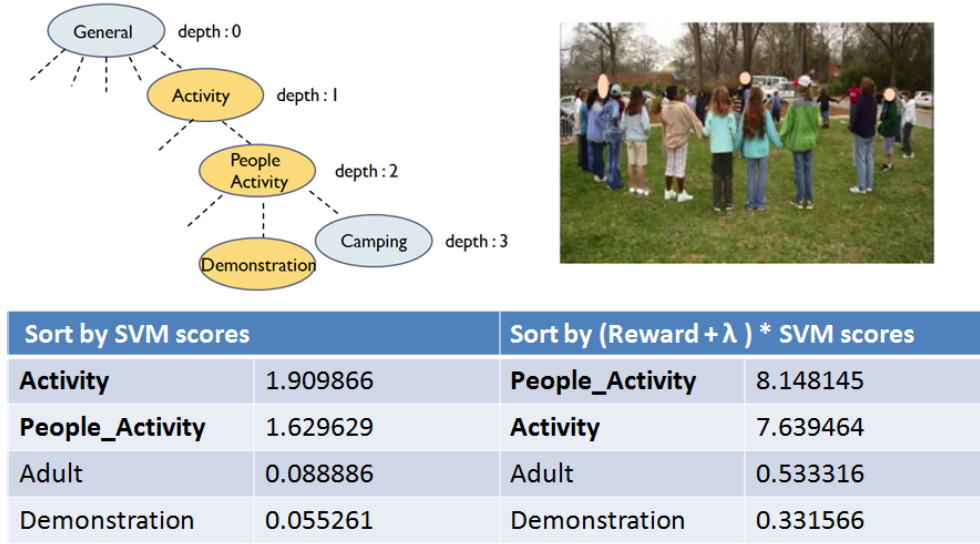


Figure 3.6: Trading off specificity (“reward”) against classifier confidence. Upper left shows part of the ontology tree; yellow nodes are event-specific concept nodes. Upper right is the corresponding keyframe. Bottom table shows ranking of concepts before and after the “rewarding”. (Principle 5)

this point, we have reduced the video to one represented by a very small number (on average, about 3) of segments, each describable by terms that imitate human specificity preferences, and whose union reflect the totality of the search query terms.

E. Best 4-second sub-segments

According to design principle of quick subsequence, we found through user studies that at most a 4-second snippet of each covering segment provides sufficient evidence to a user, as long as it is a well-chosen snippet. We therefore select that 4-second interior of a segment with the highest concept density.

$$interior_{i^*} = \arg \max_{interior_i} Ave(ConceptScores)_{interior_i} \quad (3.3)$$

At this point, we have reduced the video to one represented by a very small number

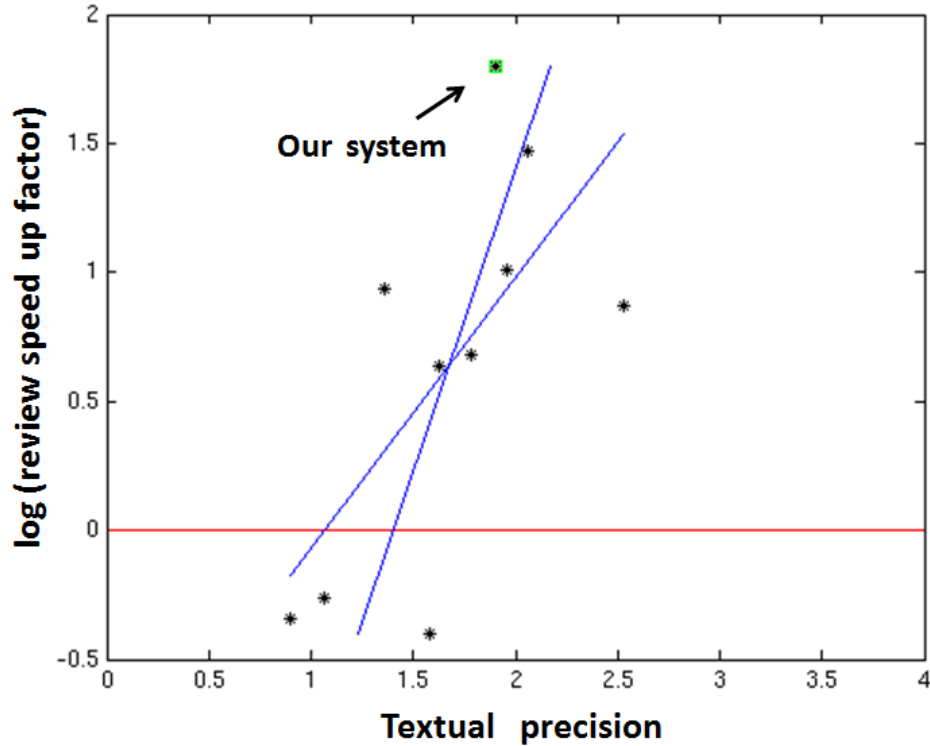


Figure 3.7: Results of ten systems in the Trecvid MER competition. Horizontal axis is textual description precision (0=Fail, 5=Excellent); vertical axis is $\log(\text{speed-up})$ of decision time compared with video length. Lines indicate two linear regressions (x vs y, y vs x); both show precision increases with speed-up. Our system is circled; speed up is by a factor of 6, the highest reported.

(on average, about 3) of very short (4-second) video snippets, each describable by a terms that imitate human specificity preferences, and whose union reflect the totality of the search query terms. After finding these snippets, we generate the text description for each of them. Noting that people ignore function words anyway, we compose the covering concept words into a simple key-word-based list as the final recounting result.

3.3 Results

Our system was part of the NIST Trecvid 2013 MER competition[2], which drew ten international entries. Three performance measures were published: accuracy (whether a given video had a specified human event), text precision (how well the description matched video content), and review time speed-up (ratio of time to view snippets compared to total video time length). A statistical analysis showed that accuracy was an independent factor uncorrelated to the other two measures, but precision was strongly correlated with speed-up: faster systems were also more correct. See Figure 3.7; our system placed first in speed without sacrificing much clarity.

3.4 Discussion

The design principles extracted from user studies allowed us to approach MER in a user-centric way. Textual description output is in the same terms as the query input; video snippets are derived from high-level concept symbols rather than low-level visual signals. Throughout, high speed-ups are possible due to an understanding of human cognitive limits, and we tried to stay close to their edge.

Some further research, however, is needed to better tune and generalize the system. We have preliminary evidence that four seconds is appropriate for convincing users of actions, but may be excessive for static concepts such as object presence. We suspect that the number of video segments necessary to persuade the user increases linearly with video length, as if there were some cognitive conservation law at work in the creation of the videos themselves: a video usually is longer because it is more conceptually complex. Lastly, we believe it

may be possible to pre-segment videos semantically independently of queries, although preliminary results suggest that it is necessarily less accurate.

News Video Summarization by Latent Coherent Patterns

4.1 Introduction

News videos contain a huge volume of daily information, and the number of online news archives has been growing at an exponential rate. Researchers to date have focused on providing viewers with concise and chronological views of news events, through methods of topic detection and tracking. Most of this research uses textual and visual features to represent news events, and to approximately track and detect news topics across time. However, a news document usually contains many more useful multimodal features, such as audio, or metadata about country of origin, news category, commentator, urgency, etc. Usually the data and metadata form coherent but latent patterns. Extracting and labeling these patterns can provide users with a more comprehensive, but also more compact, way for indexing, summarizing, and understanding news events. This is particularly true of long term international events, whose coverage varies by culture, and whose multiple topics receive varying emphasis at different times.

Co-clustering algorithms are a kind of unsupervised clustering, which have been used to mine the latent relationships between different variables[21][20][22][69]. For example, bi-clustering (two-way co-clustering) simultaneously groups the rows and columns of a

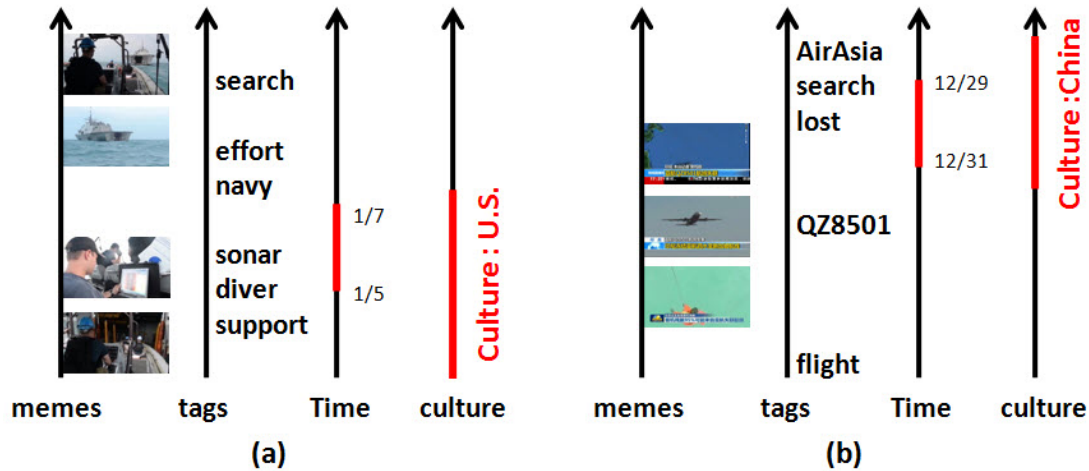


Figure 4.1: (a)(b) are examples of quad-clusters for the AirAsia Flight Q8501 event. 4-axis modalities: visual-memes, tags, time, and culture.

matrix to produce coherent submatrices. Co-clustering has been used in diverse research areas, such as gene co-expression, network traffic analysis, and social network mining. But only a small fraction of the research has leveraged tri-clustering or higher-order co-clustering algorithms [5]. Particularly in the domain of multimedia, most co-clustering work is restricted to bi-clustering [14][16][26][35][85].

In this chapter, we first mine latent coherent patterns from up to four different kinds of modalities: visual memes (frequently re-posted video segments), verbal tags, time stamps, and culture; as noted by [16], the work is challenged by the data sparsity of the co-occurring patterns of visual memes. Additionally, we note that a given visual meme can be annotated with differing verbal tags, and a given verbal tag can annotate different visual memes; both memes and tags are polysemic. Further, a meme or tag may be posted over discontinuous intervals of time and across the culture. Lastly, it may be necessary to add constraints (usually “soft” constraints) to any such algorithm, in order to better compensate for these

difficulties, so that extracted quad-clusters can be better tuned to be more consistent with other available information, such as additional data about visual meme similarity, about tag co-occurrences, about specific culture, or about known preferences for time intervals more common to news coverage.

This work is inspired by PARAFAC with sparse latent factors [63], which is evaluated by up-to three dimensional tensors, and also inspired by constrained matrix factorization [53]. We propose an algorithm which can select more dense tensors to improve extraction accuracy, and can be further tuned by synthesizing “virtual” meme, tag, time, and culture constraints in a natural way, resulting in more compact tri-clusters and quad-clusters. Our algorithm applies the iterative approximation technique of [63] in extremizing an objective function for matrix factorization, in which sparsity is actually an advantage. Figure 4.1 shows an example of one of our extracted four-way co-clusters. Our algorithm has the following advantages: (1) Allows higher-dimensional co-clustering on multiple features with a natural definition of cluster quality. (2) Allows for “soft” (partially overlapping) co-clusters to accommodate the interrelatedness and reuse of visual memes and verbal tags in the news domain. (3) Allows simple refining or biasing of the co-cluster results through constraints formalization. (4) Tolerates data sparsity. (5) Integrates time and culture sensitivity. Neither temporal tracking nor cultural differencing need to be discovered by a post-processing of the more traditional visual-verbal bi-clusters. Instead, all dimensions are handled uniformly and in parallel.

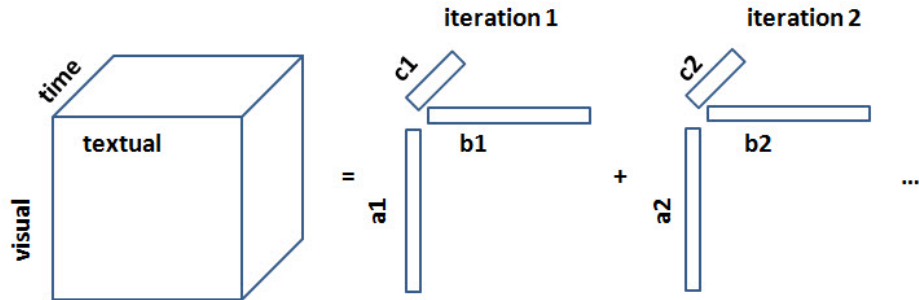


Figure 4.2: The PARAFAC decomposition of a three-dimensional tensor capturing visual, verbal, and temporal information in a news video collection.

4.2 PARAFAC Decomposition with Sparse Latent

Factors

The original PARAFAC tensor decomposition is a higher order analogue to the matrix singular value decomposition SVD, but the singular vectors produced by PARAFAC are not generally orthonormal as in the case of SVD [94]. For example, a rank- K approximation of the three-way tensors $\underline{\mathbf{X}}$ is as follows:

$$\underline{\mathbf{X}} \cong \sum_{k=1}^K \mathbf{a}_k \circ \mathbf{b}_k \circ \mathbf{c}_k \quad (4.1)$$

where \circ denotes the outer product; \mathbf{a}_k , \mathbf{b}_k , \mathbf{c}_k are vectors. The PARAFAC decomposition approximates the tensor by the sum of K rank-1 outer products shown in Figure 4.2.

The sparsity of the latent factors in multimedia news videos is a key assumption of this work. The co-clustering task selects weighted items along each dimension that co-occur; in our domain that corresponds to finding memes, tags, times (and, if four dimensional, cultures) that produce a large volume of coherent coverage of a specific news sub-event.

Therefore, \mathbf{a}_k , \mathbf{b}_k and \mathbf{c}_k are necessarily binary and sparse. For example, in our news event examples, we find only some memes and tags co-occur at only some times and only in a specific culture; see Figure 4.1.

Papalexakis *et al.* [63] demonstrated a greedy algorithm variant of PARAFAC that works effectively when all latent factors are sparse. And, by penalizing non-zero elements using an l_1 penalty, the method not only suppressed noise but also improved the separability of the co-clustering results. The method is formalized as the solution of the following constrained tensor optimization problem, with F indicating the Frobenius norm, and with each λ indicating a regulation parameter that trades off sparsity for goodness of least-squares fit:

$$\min_{\mathbf{a}_k, \mathbf{b}_k, \mathbf{c}_k \geq 0} \left\| \underline{\mathbf{X}} - \sum_{k=1}^K \mathbf{a}_k \circ \mathbf{b}_k \circ \mathbf{c}_k \right\|_F^2 + \lambda_a \sum_k \|\mathbf{a}_k\|_1 + \lambda_b \sum_k \|\mathbf{b}_k\|_1 + \lambda_c \sum_k \|\mathbf{c}_k\|_1$$

Each successive factor in the decomposition is obtained iteratively by the following three steps:

1. Initialize \mathbf{a}_k , \mathbf{b}_k , \mathbf{c}_k by the method of non-negative alternating least-squares (NN-ALS).
2. Update the approximations of \mathbf{a}_k , \mathbf{b}_k , \mathbf{c}_k by the method of alternating Lasso [78].
3. At convergence, replace $\underline{\mathbf{X}}$ by the residual tensor $\underline{\mathbf{X}} - \mathbf{a}_k \circ \mathbf{b}_k \circ \mathbf{c}_k$, and repeat.

This method tends to extract the most dominant co-cluster first, and less dominant co-clusters in order. However, neither the original PARAFAC nor this greedy variant accommodate any additional information or constraints that are both very useful and readily ob-

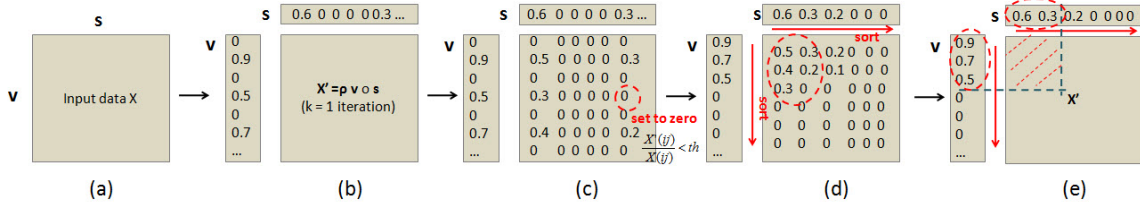


Figure 4.3: Selecting and refining high-dimensional co-clusters from the without loss of generality in two dimensions.

tainable in this domain of news events. Prior knowledge, such as how memes and tags tend to co-occur, or biased preferences, such as a need to focus on a particular time period or culture, or on clusterings that are particularly large or dense, are not easily accommodated.

4.3 Algorithm

We start by extending parallel factor decomposition with sparse latent factors [63] to 4 dimensions. Then we propose a way to select and refine the high-dimensional tensors. Then we propose a way to formalize additional soft constraints on visual memes, tags, timestamps and cultures, which reflect specific interests in news video understanding.

4.3.1 Feature Extraction and Data Tensor

We represent multimodal features of news videos by $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \dots, \mathbf{f}_n\}$, where each \mathbf{f} indicates a feature vector of a modality. In this chapter, n equals up to 4, but higher dimensions are possible. The four modalities are:

Visual memes. We define a visual meme as a frequently re-posted video segment that starts and ends at shot boundaries. In order to find visual memes from our collected video repository, first we extract high-entropy I-frames from each video, remove frames contain-

ing anchorpersons [88], perform shot detection over this reduced frame set, and then select one keyframe to represent each shot. Then, we cluster these selected keyframes by near-duplicate matching, using SIFT-BoF and color histograms. Each visual meme is defined as a set of two or more near-duplicate keyframes extracted from a news video corpus.

Verbal Tags. We derive verbal tags from the titles and the descriptions in video metadata. For text that is not in English, we translate text into English using Microsoft Translate API [1]. Then, we apply a standard NLP processing pipeline of tokenization, stopword removal, and lemmatization, to extract tags from these texts.

Time stamps. We extract published date information from video metadata. We ignore time information, as our experiments show it to be of too fine a grain to be useful.

Cultures. We verify the video metadata and obtain the cultural origins of our collected videos.

After extracting features from each modalities, we represent them as follows: $\mathbf{F} = \{\mathbf{v}, \mathbf{s}, \mathbf{t}, \mathbf{c}\}$, where \mathbf{v} is a vector of visual memes indices, whose i -th visual-meme is $\mathbf{v}(i)$, with $|\mathbf{v}| = I$. Similarly, \mathbf{s} is a vector of tags indices, whose j -th tag is $\mathbf{s}(j)$ with $|\mathbf{s}| = J$. \mathbf{t} is a vector of date indices in chronological order, whose m -th timestamp is $\mathbf{t}(m)$ with $|\mathbf{t}| = M$. Lastly, \mathbf{c} is a vector of culture indices, whose l -th culture is $\mathbf{c}(l)$ with $|\mathbf{c}| = L$. L equals 4 in our entire dataset because we collect videos originating from US, Europe, South America and China. We model the co-occurring relationship of these four kinds of features in \mathbf{X} , which is a four-dimensional tensor of size $I \times J \times M \times L$. The entry $\mathbf{X}(i, j, m, l)$ counts the number of videos in which feature $\mathbf{v}(i), \mathbf{s}(j), \mathbf{t}(m), \mathbf{c}(l)$ mutually co-occur.

4.3.2 Problem Formulation

Due to the repetitive nature of the coverage of a news event, every day produces a number of videos. However, the distribution of memes and tags on any given day tend to be sparse, compared to the universe of memes and tags used for that event. Tags are easier to generate, so they are generally richer than memes; for example, in our Ebola dataset, the average video has two memes but seven tags. As in other text analysis research, we eliminate high frequency tags (they do not help discriminate between clusters) and low frequency tags (they tend to be isolated). We first extend the PARAFAC with sparse latent factors [63] to four dimensions. Using their notation, let $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_k]$, $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_k]$, $\mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_k]$, $\mathbf{C} = [\mathbf{c}_1, \dots, \mathbf{c}_k]$, where k is the number of quad-clusters that are to be extracted from $\underline{\mathbf{X}}$. Then their algorithm, once extended to our data, in more detail becomes:

$$\begin{aligned} & \min_{\{|\rho_k| \leq \check{\rho}, 0 \leq \mathbf{v}_k, \mathbf{s}_k, \mathbf{t}_k, \mathbf{c}_k \leq 1\}_{k=1}^K} \left\| \underline{\mathbf{X}} - \sum_{k=1}^K \rho_k \mathbf{v}_k \circ \mathbf{s}_k \circ \mathbf{t}_k \circ \mathbf{c}_k \right\|_F^2 + \\ & \lambda_v \sum_k \|\mathbf{v}_k\|_1 + \lambda_s \sum_k \|\mathbf{s}_k\|_1 + \lambda_t \sum_k \|\mathbf{t}_k\|_1 + \lambda_c \sum_k \|\mathbf{c}_k\|_1, \quad (4.2) \\ & \text{where } \check{\rho} = \max_{i,j,m,l} |\underline{\mathbf{X}}(i, j, m, l)|. \end{aligned}$$

where ρ_k now represents a scaling factor necessary for normalization, and \circ denotes the outer product. This formulation, on data such as ours, has been shown by its original authors to be more effective in extracting a large number of possibly overlapping co-clusters than do more general matrix factorization methods. It also enforces the compactness of the decomposition results by penalizing the number of non-zero elements in the factors. This

constrained minimization problem can be solved iteratively given in more detail in the while loop in Algorithm 1.

Algorithm 1 While loop for updating 4-dimensional tensors

```

1: while change in cost >  $\epsilon$ , explained in [63] do
2:    $\mathbf{v} = \min_{\{0 \leq v \leq 1\}} \|\underline{\mathbf{X}} - \rho \mathbf{v} \circ \mathbf{s} \circ \mathbf{t} \circ \mathbf{c}\|_F^2 + \lambda_v \sum_i |\mathbf{v}(i)|$ 
3:    $\rho = \min_{\{0 \leq \rho \leq \rho_{max}\}} \|\underline{\mathbf{X}} - \rho \mathbf{v} \circ \mathbf{s} \circ \mathbf{t} \circ \mathbf{c}\|_F^2$ 
4:    $\mathbf{s} = \min_{\{0 \leq s \leq 1\}} \|\underline{\mathbf{X}} - \rho \mathbf{v} \circ \mathbf{s} \circ \mathbf{t} \circ \mathbf{c}\|_F^2 + \lambda_s \sum_j |\mathbf{s}(j)|$ 
5:    $\rho = \min_{\{0 \leq \rho \leq \rho_{max}\}} \|\underline{\mathbf{X}} - \rho \mathbf{v} \circ \mathbf{s} \circ \mathbf{t} \circ \mathbf{c}\|_F^2$ 
6:    $\mathbf{t} = \min_{\{0 \leq t \leq 1\}} \|\underline{\mathbf{X}} - \rho \mathbf{v} \circ \mathbf{s} \circ \mathbf{t} \circ \mathbf{c}\|_F^2 + \lambda_t \sum_m |\mathbf{t}(m)|$ 
7:    $\rho = \min_{\{0 \leq \rho \leq \rho_{max}\}} \|\underline{\mathbf{X}} - \rho \mathbf{v} \circ \mathbf{s} \circ \mathbf{t} \circ \mathbf{c}\|_F^2$ 
8:    $\mathbf{c} = \min_{\{0 \leq r \leq 1\}} \|\underline{\mathbf{X}} - \rho \mathbf{v} \circ \mathbf{s} \circ \mathbf{t} \circ \mathbf{c}\|_F^2 + \lambda_c \sum_r |\mathbf{c}(r)|$ 
9:    $\rho = \min_{\{0 \leq \rho \leq \rho_{max}\}} \|\underline{\mathbf{X}} - \rho \mathbf{v} \circ \mathbf{s} \circ \mathbf{t} \circ \mathbf{c}\|_F^2$ 
10:  Select_dense_tensor( $\mathbf{v}, \mathbf{s}, \mathbf{t}, \mathbf{c}$ )
    
```

Several authors have considered the problem of tuning the λ regularization parameters for the usual linear regression version of Lasso, assuming the availability of training data for cross-validation, or that the data comes from a known distribution. However, the PARAFAC method selects each λ solely based on the input data array $\underline{\mathbf{X}}$. For each dimension, it derived an upper bound on that dimension's λ , and empirically found that choosing each λ to be a small percentage (typically, 0.1%) of the maximum worked very well. For example, the upper bound of our λ_t can be set, using Matlab-like notation, as

$$\lambda_t^* = 2\bar{\rho}\bar{v}\bar{s}\bar{c}\max_m(\|\underline{\mathbf{X}}(:, :, m, :)\|_2) \quad (4.3)$$

where \bar{v} , \bar{s} , and \bar{c} is the expected number of non-zeros elements in \mathbf{v} , \mathbf{s} , and \mathbf{c} . By the

same way, we can calculate the upper bound of λ_v^* , λ_s^* , and λ_c^* also. We note that a visual meme, on average, co-occurs with 5.72 other visual memes; a tag, on average, co-occurs with 7.01 other tags; visual memes are rarely reposted after three days, and 75% news subevents are cultural-specific. Thus \bar{v} , \bar{s} , \bar{t} , \bar{c} can be simply set to 5.72, 7.01, 3, and 1.25.

4.3.3 Tensor Decomposition and Dense Tensor Selection

There are a number of state-of-the-art metrics that can be used to score the quality of a tensor, such as density, size, concentration, or contrast, as discussed in [36]. They proposed to identify dense sub-tensors based on metrics reflecting their problem of interest. For our needs, we select and refine the tensor by searching for the largest robust tensors that meet a density threshold.

Figure 4.3 illustrates how we extract and then refine a quad-cluster (a rank-1 tensor) taken from $\underline{\mathbf{X}}$. We start with feature co-occurrence counts $\underline{\mathbf{X}}$ in Figure 4.3 (a). We apply the algorithm to produce the sparse quad-cluster $\underline{\mathbf{X}}'$ in Figure 4.3 (b). Then, to “clean” this tri-cluster, for each element in $\underline{\mathbf{X}}'$, we set it to 0 if $\frac{\underline{\mathbf{X}}'(i,j,m,l)}{\underline{\mathbf{X}}(i,j,m,l)} < \theta$ in (Figure 4.3) (c). (We set $\theta = 0.2$ here.) To compactify the factor basis vectors, we permute $\underline{\mathbf{X}}'$ by sorting each of the dimensions, \mathbf{v} , \mathbf{s} , \mathbf{t} , and \mathbf{c} , independently, which heuristically gathers the most significant non-zero components of $\underline{\mathbf{X}}'$ “into the corner”; see Figure 4.3 (d). Lastly, we perform a greedy search over each element’s $\frac{\underline{\mathbf{X}}'(i,j,m,l)}{\underline{\mathbf{X}}(i,j,m,l)}$ value, looking to see if that element’s indices should be incorporated into the vectors that make up the outer product. The goal is to find the largest compact four-dimensional “corner” tensor that meets the density criterion Figure 4.3 (e).

The indices of this tensor specify the sub-vectors of \mathbf{v} , \mathbf{s} , \mathbf{t} and \mathbf{c} that index into the memes, tags, times, and cultures that make up the quad-cluster. Each element in these vectors indicates the fractional membership in factor k , where $\sum_{k=1}^K v(i)_k \leq 1$, $\sum_{k=1}^K s(j)_k \leq 1$, $\sum_{k=1}^K t(m)_k \leq 1$ and $\sum_{k=1}^K c(r)_k \leq 1$. This resulting tensor factor describes and summarizes a significant latent event pattern. We then update $\underline{\mathbf{X}} = \underline{\mathbf{X}} - \underline{\mathbf{X}}'$, and repeat the



Figure 4.4: This figure summarizes 2 culture-specific quad-clusters. One is Europe, and the other is U.S. Modalities: visual-memes, tags, time and culture.

process to extract the next quad-cluster until the norm of the new residual tensor is smaller than an ϵ , or until the residual tensor itself does not change. Some extraction examples are in Figure 4.4.

Ideally, the computation of the quality of the factor should be based on some understanding of the prior distribution of the individual features. Jiang *et al.* [36] assume that events are randomly distributed across the tensor data, and define quality metrics based on a Poisson distribution. But we note that visual memes, tags, and time of news video more often follow a power-law distribution.

Our ability to derive the parameters of such distributions is limited by the amount of data that can be gathered under stationary conditions. In particular, we note that repeated visual memes, which are the most difficult dimension to populate, tend to be sparse and must be taken over large intervals of time. We intend to address this more thoroughly in future work. In the meantime, we use raw feature co-occurrence counts as the basis for our density calculation.

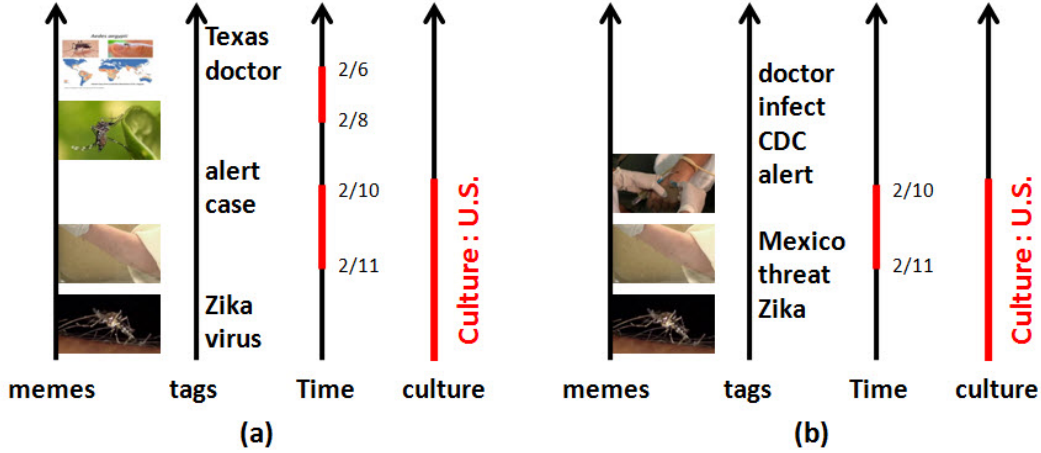


Figure 4.5: (a) One of extracted Zika quad-cluster which is dominated by common tags. (b) Add soft constraints to visual-memes and tags with $w_v = 1, w_t = 0.5$. The memes are an exact subset, and the soft constraints have eliminated more common tags, compared to (a).

4.3.4 Adding soft constraints

We note that Figure 4.5(a) is dominated by common tags. We therefore add soft constraints to our tensor, in order to vary the effect of one or other of its dimensions on the clustering. For example, we might want increase the importance of visual (or verbal, or temporal, or cultural) relationships. Figure 4.6 (a) shows how we generate such constraints and express them as “virtual” memes, tags, times or cultures.

Constraint modeling. We derive intra-modality constraints by studying each dimension separately. We express memes, tags, times or culture by nodes in their own weighted graph. We draw a weighted edge between two nodes that counts the number of videos that they co-occur in. For example, memes that have many such edges are important, and the group of memes they tie together act as if they they shared a tag. So, we invent a tag, a virtual one. For visual memes, we define the importance of a node by its centrality in this weighted graph [61]. $C_{v(i)} = \sum_p^I w_{v(i)v(p)}$, where $C_{v(i)}$ is the centrality of $v(i)$. $w_{v(i)v(p)}$ is the weight

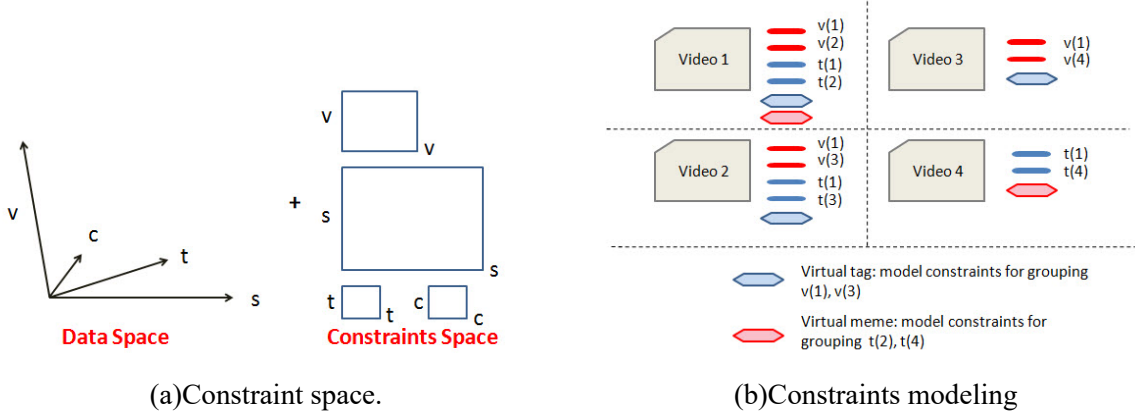


Figure 4.6: (a) Our data is a four-dimensional tensor, where each entry of the matrix counts the number of co-occurrences of $\mathbf{v}(i), \mathbf{s}(j), \mathbf{t}(m), \mathbf{c}(l)$. But additionally we can measure inter-video relationships, such as how \mathbf{v} or \mathbf{s} tend to co-occur across videos. (b) To model constraints, we create “virtual” tags and memes. We detect memes with high video-to-video centrality, for example, $\mathbf{v}(1)$. We then create a virtual tag that expresses this relationship to its “nearby” memes (the tag is: “this is the $\mathbf{v}(1)$ group”). We similarly create virtual memes based on high video-to-video tag co-occurrence.

of edge of $v(i)$ and $v(p)$. If the central visual meme is $v(i)$, the tag would be “the $v(i)$ group”. Sometimes the “meaning” of “the $v(i)$ group” is obvious from its content (e.g. “a group of doctors”). We select the N most central visual memes, and create virtual tags that express which other memes they are related to. An example is in Figure 4.6 (b).

We then add these N virtual tags into the original tensor \mathbf{X} , where they work as additional soft constraints. We apply the same procedure to model soft constraints for tags by adding virtual memes, where they represent a “missing” meme that the co-occurring tags strongly hint at. Creating a virtual time is similar; if times appeared linked because the memes or tags show they have a commonality (particularly if there is a temporal interruption in the news reporting and a news event then “reappears”), then additional virtual memes or tags are added to the central times. We can model culture constraints by a similar way.

Constraints weighting. The weight of each of these virtual constraints is derived algorithmically.

mically, based on their co-occurrence across videos. These numbers generally are higher than the elements of the tensor. So, for convergence, after we create $\underline{\mathbf{X}}$ with constraints, we need to normalize $\underline{\mathbf{X}}$ to $[0, 1]$ by $\max(\underline{\mathbf{X}})$. We also need to adjust the maximum element of the tensor space bounded by the virtual memes and virtual tags to 1. Then, having balanced the real with the virtual, we are free to define four weights w_v, w_s, w_t, w_c , each $w \in [0, 1]$, to separately weight the effect of these four kinds of constraints, depending if we want to emphasize visual, textual, temporal or culture connections. Figure 4.5 (b) shows the effect of virtual meme constraints on fine-tuning the quad-clustering result; note that the tags are more informative.

4.4 Discussions

In this work, we detail a tensor factoring algorithm that allows a user to add soft constraints to each dimension of features separately. We use our algorithm to run **bi-clustering** on two-dimensional tensors (visual memes and tags), **tri-clustering** on three-dimensional tensors (visual memes, tags, time), and **quad-clustering** (visual memes, tags, time, and culture) for evaluation. We have collected about 3100 videos and their metadata, in an approximate 3:1 (US:Europe) ratio for the Ebola news event, in a date range from 8/21/14 to 11/30/14; about 1000 videos and their metadata, in an approximate 1:1 (China:US) ratio for the AirAsia Flight 8501 event, in a date range from 12/28/14 to 1/15/15; and about 1700 videos for the Zika news event and their metadata, in an approximate 7:10 (South America:US), in a date range from 12/01/15 to 2/15/16. Videos sourced from US, Europe, and South American were collected from YouTube, and we verified their posted location in the metadata. Videos

Algorithm	(a)	(b)	(c)	(d)
Average F1 score	0.24	0.35	0.42	0.51

Table 4.1: Average performance of bi-clusters extraction from sampled Ebola, AirAsia, Zika dataset by different algorithms. (a) Information-Theoretic co-clustering [6]. (b) Spectral co-clustering [21]. (c) Hierarchical co-clustering [40]. (d) Extended PARAFAC with sparse latent factor and dense tensor selection.

Extended PARAFAC	bi	tri	quad
Average F1 score	0.51	0.62	0.63

Table 4.2: Average performance of bi, tri and quad-clusters extracted by extended PARAFAC algorithm from sampled Ebola, AirAsia, Zika dataset.

sourced from China were collected from Baidu, the biggest Chinese video search engine in the world, which aggregates videos from Chinese online news channels and from Chinese video sharing websites. We obtained 5546 distinct visual memes and 8031 tags in total. Our largest data tensor represented one month of data of Ebola news event, with size $\mathbf{v} \times \mathbf{s} \times \mathbf{t} \times \mathbf{c} = 1326 \times 2563 \times 31 \times 2$. Extracting 10 quad-clusters from this data matrix without optimization is within 5 minutes on one machine with an Intel Xeon L5420 CPU running at 2.5GHz on 16 GB. We find it acceptable for our experiments, even though we have programmed no specific optimizations, such as parallelizing the vector update processes.

4.4.1 Extraction Accuracy

Because both memes and tags are polysemic, extracting co-clusters from our sparse dataset is much more challenging. Without a suitable high-dimensional co-clustering multimedia algorithm to compare our results with, we simply compare our algorithm against well-

known bi-clustering algorithms by calculating average F1 score[90], which is defined as:

$$\frac{1}{2} \left(\frac{1}{|C^*|} \sum_{C_i \in C^*} F1(C_i, \hat{C}_{g(i)}) + \frac{1}{|\hat{C}|} \sum_{\hat{C}_i \in \hat{C}} F1(C_{g'(i)}, \hat{C}_i) \right) \quad (4.4)$$

where C^* is the set of ground truth co-clusters, and \hat{C} is the set of extracted co-clusters. This measure first determines which $C_i \in C^*$ corresponds to which $\hat{C}_i \in \hat{C}$. Then it defines average F1 score to be the average of the F1-score of the best-matching ground truth co-cluster to each extracted co-cluster, and the F1-score of the best-matching extracted co-clusters to each ground truth co-clusters. g and g' in the above equation represent best-matching functions, where $g(i) = \underset{j}{argmax} F1(C_i, \hat{C}_j)$, and $g'(i) = \underset{j}{argmax} F1(C_j, \hat{C}_i)$, and F1 is harmonic mean of precision and recall.

We note that our smallest dataset, AirAsia flight Q8501, only has 5 salient co-clusters. In order to fairly compare the accuracy, for each news event we first sample and verify a subset of videos with 5 co-clusters (subevents) as the ground truth dataset. We then randomly add several unrelated videos as noise, at about 10% of the size of ground truth dataset. Our method works better than others on extracting co-clusters from sparse and overlapping meme-tag tensors mostly because our algorithm handles all dimensions uniformly, tolerates sparsity, and selects dense tensors (Table 4.1).

We also note that high-dimensional co-clustering gives better accuracy (Table 4.2). There is a 10% improvement from bi to tri-clustering. However, there is only 1% improvement from tri to quad-clustering. This is partly because 75% of news subevents are cultural-specific, so they have some specific visual memes, tags and timestamps. And partly because, even for the same subevent, different cultures favor different visual memes and

Tags of Bi-clustering results related to “Ebola virus spread in Sierra Leone”.	
(a) Bi-cluster	
virus(0.93) outbreak(0.72) spread(0.71) life(0.62) Sierra(0.54) Leone(0.54) deadly(0.32)	

Tags of Tri-clustering results related to “Ebola virus spread in Sierra Leone”.	
(b) Tri-cluster 1	(c) Tri-cluster 2
first(0.63) case(0.63) grow(0.63) fear(0.53) Senegal(0.52) virus(0.44) outbreak(0.38)	patient(0.71) escape(0.60) quarantine(0.59) panic(0.57) Liberia(0.57) virus(0.42) outbreak(0.31)

Figure 4.7: Comparing tags of bi-clustering results to tags of tri-clustering results extracted from the same video dataset. (a)(b)(c) show the top 7 tags with their membership scores within the extracted bi-cluster and tri-clusters. (a) Bi-clustering only extracts 1 bi-cluster related to “Ebola virus spread in Sierra Leone”, and its tags are dominated by co-occurrence of most common tags in the system. (b)(c) However, with additional time information, tri-clustering extracts 2 tri-clusters related to “Ebola virus spread in Sierra Leone”. Each of them represents more specific news stories, in which common tags like “virus” and “outbreak” are split, and tags of subevents (“first”, “case”, “patient”, “escape”,...etc.) have higher membership scores.

tags, which already partially act similarly to a culture dimension.

4.4.2 Higher-order Co-clustering

We compare the results of tri-clustering and quad-clustering of the same Ebola dataset. We note that quad-clustering is able to extract more detailed subevents such as Obama’s speech, favored by US media, and the actions of Médecins Sans Frontières (MSF), favored by Eu-


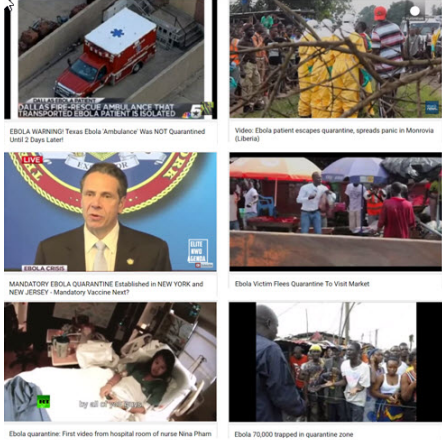
(a)Constrained tag group	(b)Extracted quad-cluster		(C)Examples of YouTube videos where extracted visual memes and tags come from
	Tags	Visual memes	
quarantine (central tags) nurse doctor vaccine patient	quarantine nurse doctor vaccine patient victim hospital Sierra escape virus Dallas risk Liberia CDC Africa		

Figure 4.8: Extracted quad-cluster which is constrained by a group of tags. (a) is a list of high-frequency tags, which co-occur with “quarantine”. We model a constraint on this group of tags by creating a virtual visual meme (Section 4.4). (b) Extracted quad-cluster by adding the constraint. (Here we show tags and visual memes only.) (c) Examples of YouTube videos where the extracted tags and visual memes come from. The images in (c) are keyframes, which are used to represent videos, not necessarily visual memes.

rope news media. We also compare bi-clustering and tri-clustering with the same set of video data. In general, tri-clustering extracts fewer and more accurate co-clustering results than bi-clustering. We also note that there are many news subevents of Ebola that carry common tags (“virus”, “spread”, “outbreak”), but they are actually stories with a different focus and with different timestamps. Without time information, bi-clustering tends to co-cluster all memes and tags, even though they are separable in time. We ran bi-clustering and tri-clustering on the same video dataset. Figure 6.4 shows a comparison between tags of a bi-cluster and two tri-clusters.

4.4.3 Constraints Weighting and Limitations

Adding soft constraints into the data tensor can sharpen the co-clustering results, as shown in Figure 4.5. If we weight virtual meme constraints heavily, we are able to extract visual memes corresponding to the tags that generate those virtual memes.

In addition, soft constraints can also be used to fine-tune co-clustering results for a more comprehensive news understanding. For example, if we want to know a group of visual memes and tags related to a keyword “quarantine” in Ebola news event, we can select a list of keywords co-occurring with “quarantine”, and put an ad-hoc constraint to tie them together. The resulting quad-cluster is in Figure 4.8, which can be considered to be a summarization of “quarantine” in Ebola news.

However, adding too many constraints can hide the latent patterns of original data matrix. In general, recognizing important tags would then become more difficult, since a tag with high centrality (importance score) is usually a more common tag, which would not help to obtain a more informative co-cluster. We will investigate how to better utilize soft constraints in the future.

4.5 Conclusion

This chapter presents an efficient constrained tensor factorization algorithm for integration of multimodal data, and more specifically of language and vision data. By representing news videos as four-dimensional tensors (visual memes, tags, time stamps and cultures), the proposed algorithm can extract quad-clusters for better understanding news events. We show that this method is more accurate for extracting bi-clusters from two-dimensional

sparse (visual memes and tags) tensors, compared to other well-known bi-clustering algorithms. We also show that using more modalities can increase co-cluster accuracy and extract more detailed news subevents. Our algorithm also allows simple refining or biasing of the co-cluster results through constraints formalization, which can fine-tune the co-clustering results for more comprehensive news understanding.

Cross-cultural Visual Meme Tracking

5.1 Introduction

News tracking (text-only, visual-only, or both) [49] [64] [87] has been an active research topic in recent years. In the literature, a “meme” is a frequently re-posted phrase, and a “visual meme” is a frequently re-posted video segment or image which was first introduced in [87]. They both act as signatures of topics and events, and their propagation and diffusion over the web has been widely used to monitor the lifespan of a news story. While some journalism research has (mostly manually) investigated the impact of cultural differences in meme analysis, visual meme analysis is still limited to approaches that are not culturally based. Additionally, IR researchers have found that the coverage of text-based news is known to be negatively influenced by the physical distance of consumers from the location of the event. But visual information by its very nature is much less dependent on language, and therefore it is much more likely to become viral across cultures.

With the continued rapid growth of online video data, the common usage of only a few universal video sharing platforms (YouTube, Baidu), and the universal importance of human-interest international news events, it becomes imperative to understand how visual memes can cross cultural language barriers. Understanding these viral paths will enable

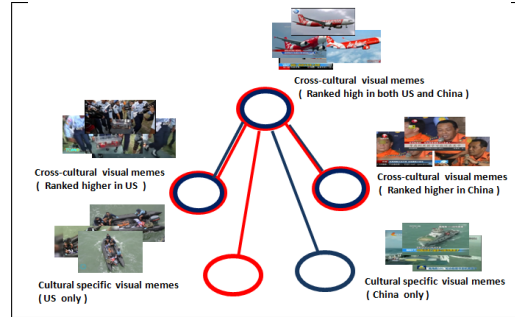


Figure 5.1: A partial illustration of a cross-cultural visual meme graph for the AirAsia Flight Q8501 event. Red nodes: visual memes occurring in U.S.. Blue nodes: visual memes in China. Red and blue nodes overlapping: visual memes in both countries.

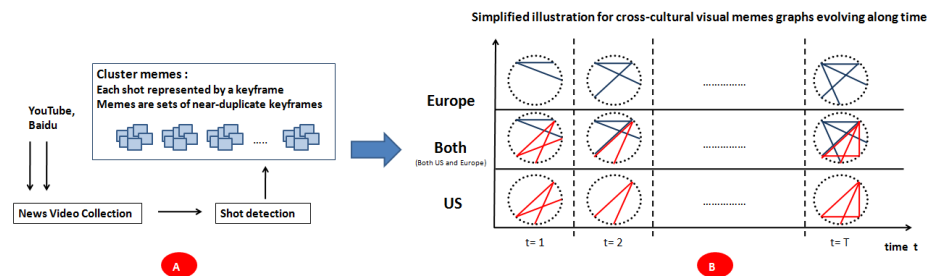


Figure 5.2: Pipeline for generating cross-cultural visual meme graphs. (A) Video preprocessing. We collect videos from YouTube and Baidu, detect shots in each video, then visually cluster these shots, with each shot represented by a keyframe. A visual meme is defined as a collection of two or more near-duplicate keyframes. (B) Simplified visualization of visual meme graphs for the Ebola event, in a culture-versus-time matrix. Each graph displays in a circle the same visual meme nodes, but with only the edges present in culture c at time t , illustrating differences and evolution.

users in a given culture to find more complete information about an international event by querying the search engines of another culture, particularly for those aspects of an event that are not in the news coverage of the user’s culture. We have developed representations and methods to recognize, track, measure, and query these cross-cultural visual memes. These tools show interesting patterns of visual meme propagation not previously reported (Figure 5.1).

We propose a new way to use the PageRank algorithm to model cross-cultural visual meme influence, in order to capture the rates that a visual meme will be re-posted in a

specific time period in a particular culture. We have collected videos from YouTube (the primary source for U.S. and European videos) and from Baidu (the primary source for Chinese videos). We use generic text queries as a pre-filter for content on a given topic, then build visual meme co-occurrence graphs to calculate cross-cultural visual meme influence. In this chapter we specifically showcase two international events, the Ebola crisis and the AirAsia crash, from three viewpoints, showing the tracking of their memes, and discussing their cultural differences.

Prior literature such as [87] tracked the amounts of visual memes over time through YouTube. However, their analysis is based on video volumes and video relationship graphs, where the nodes were videos, and the edges indicate a shared visual meme. Videos are clustered using visual memes. Their primary concern was to determine the author's and the reposters' roles in visual meme diffusion.

In contrast, our analysis is based on a novel visual meme co-occurrence graph, which is a kind of dual to the previous graph. In our graph, the nodes are the visual memes themselves, and the edges indicate visual meme co-occurrence within the same video. Visual memes are clustered using videos. From this graph we can compute new measures of visual meme dynamics, which include, but are not limited to, the rate that a visual meme is re-posted, the relationship of visual meme influence on video composition versus the view counts of those videos, and a number of quantifications and visualizations of cross-cultural visual meme propagation.

5.2 Cross-cultural Visual Meme Influence and Tracking

Algorithm

In this section, we present the pipeline for calculating cross-cultural visual meme influence.

(Figure 5.2)

5.2.1 Visual Meme Clustering

We define a visual meme as a frequently re-posted video segment that starts and ends at shot boundaries. In order to find visual memes from our collected video repository, first we extract high-entropy I-frames from each video, perform shot detection over this reduced frame set, then select one keyframe to represent each shot. Then, we cluster these selected keyframes by near-duplicate matching, using SIFT-BoF and color histograms. Each visual meme is defined as consisting of a cluster with two or more keyframes.

5.2.2 Visual Meme Graph Construction and Influence Calculation

We track the influence of visual memes as they change across two discrete dimensions: culture and time. We represent a culture as $c \in C = \{US, Europe, China\}$, and a time as $t = 1, 2, \dots, T$. At each c and t we construct a visual meme influence graph, where the nodes are visual memes and the edges represent the degree to which the memes co-occur within the same video.

We first note that the total set of visual memes that occur across all cultures and all times can be represented by $M = \bigcup_i m_i$, where m_i is a visual meme. Each m_i carries a weight,

w_i , defined as the number of videos that contain a near-duplicate shot of this visual meme; more viral visual memes have higher weights. We also note that any pair of visual memes m_i and m_j can be related by a weighted undirected edge s_{ij} , which counts the number of videos which share both memes; we let the set $S = \{s_{ij}\}$.

Now, at any given t , and any culture c , we can form a graph that represents the sharing of visual memes at that time and in that culture, $G_{ct} = (M_{ct}, S_{ct})$, where the nodes and edges are selected into G_{ct} from M and S in the obvious way. For example, the memes are m_{cti} and the edges are s_{ctij} . We can create $|C| * T$ such graphs.

We can apply the PageRank algorithm [62] to each G_{ct} , computing for each visual meme $r(m_{cti})$, in the context of all visual memes in M , with $\epsilon = 0.15$:

$$r(m_{cti}) = (1 - \epsilon) \sum_{(m_{cti}, m_{ctj}) \in G_{ct}} \frac{r(m_{ctj})}{\text{out}(m_{ctj})} + \frac{\epsilon}{|M|} \quad (5.1)$$

However, because we have many such graphs G_{ct} , in some of these graphs there will be “dangling” visual memes, that is, visual memes of M that do not properly occur in M_{ct} . Let this set be $D_{ct} = \bigcup_k d_{ctk}$. Since this set varies over c and t , and since PageRank considers their effect in each graph G_{ct} , the values of individual PageRank scores, $r(m_{cti})$, are generally not comparable across different graphs. To normalize these scores, we borrow the idea in [10], which divides each score by a value l_{ct} that removes the varying effects of the D_{ct} as follows:

$$l_{ct} = \frac{1}{|M|} (\epsilon + (1 - \epsilon) \sum_{D_{ct}} r(d_{ctk})) \quad (5.2)$$

These normalized PageRank scores convey how much more likely a node is to be visited

in a random walk, compared to a node having the least possible importance. Thus, we can compare these normalized PageRank scores across different graphs, and we are able to track the influence of a visual meme across cultures and along time.

5.3 Discussions

In this chapter, we examine two events. We collected about 2400 videos and their metadata in an approximate 3:1 (US:Europe) ratio for the Ebola news event, in a date range from 8/21/14 to 10/30/14. We collected about 1000 videos and their metadata in an approximate 1:1 (China:US) ratio for the AirAsia Flight 8501 event, in a date range from 12/28/14 to 1/15/15. Videos sourced from US and Europe were collected from YouTube, and we verified their posted location in the metadata. Videos sourced from China were collected from Baidu¹, the biggest Chinese video search engine in the world, which aggregates videos from Chinese online news channels and from Chinese video sharing websites.

We successfully extracted cross-cultural visual memes from our data set: 3969 distinct ones for Ebola and 553 distinct ones for AirAsia. (The number for Ebola is greater than that for AirAsia because the date range for Ebola is longer.) We found that less than 1% of the visual memes were isolated, that is, they did not co-occur with some other visual memes in at least one video; most re-postings involve multiple visual memes at once. In this chapter, our analysis uses U.S., Europe, China, and “Both” cultures, where Both for Ebola means Europe and U.S. visual memes considered together, but for AirAsia it is China and U.S. together.

¹<http://v.baidu.com/>

	Ebola	AirAsia
Europe only	88	
US-Europe both	65	
US only	273	96
US-China both		37
China only		48

Table 5.1: The average number of visual memes per day for Ebola and AirAsia news events.

5.3.1 Visual Meme Coverage in Different Cultures

The average number of visual memes per day in each culture is shown in Table 5.1. For both events, US has more culturally-specific visual memes than the other two cultures. Aside from the fact that U.S. media dominate world media, this imbalance is partly caused by the large amount of videos remixed and uploaded by users from U.S. sources. The data also reflects the prevalence of more personal viewpoints in visual memes selection in the U.S. In general, there are very few personally edited videos from China, and even in those, the visual meme selection is not very different from those of the Chinese news video mainstream.

5.3.2 Visual Meme Influence Correlation

We calculate Pearson and Spearman correlation coefficients between the normalized PageRank scores, for cross-cultural visual meme coverage along time:

$$\rho_{Pearson} = Pearson(r(t, c_1), r(t, c_2)) \quad (5.3)$$

$$\rho_{Spearman} = Spearman(r(t, c_1), r(t, c_2)) \quad (5.4)$$

where r refers to normalized PageRank score, t refers to timestamp, c_1 is U.S., and c_2 is Europe for Ebola but China for AirAsia. We display these correlation coefficients

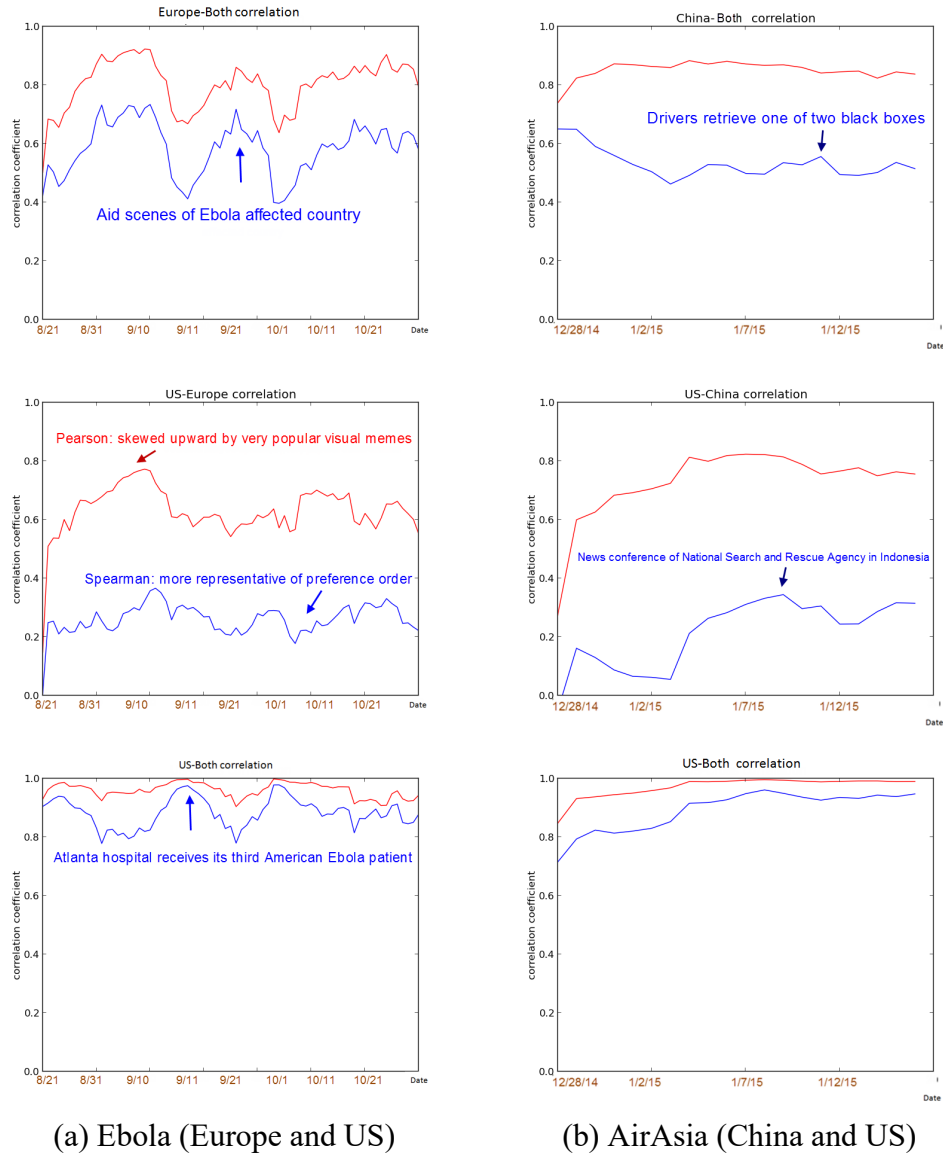


Figure 5.3: Cross-cultural visual meme influence correlation coefficients for different countries, for different news events, along time. Red curves are Pearson, blue curves are Spearman.

in Figure 5.3. We note that in general Pearson and Spearman are similar in shape, but the Spearman coefficient is always smaller. This is because visual meme influence is not uniformly distributed; most influence values are small, and are not shared across cultures. For example, one visual meme for Ebola scored a 8.34 PageRank in the U.S., but only scored 1.12 in Europe, which is near the normalized lower bound of 1. But we also note that a few visual memes scored very high in both cultures, showing to some degree a universal interest in specific visual memes.

We can interpret some cultural preferences from the ups and downs in the correlation curves. In the Europe-Both correlation curve for Ebola, we note that there are three peaks. The first peak (from 9/2 to 9/12) is due to news related to the Ebola survivor William Pooley (a British nurse), and to a WHO conference discussing aid to Africa. During the second peak (from 9/22 to 9/30), European news focused on the reporting of the spread of Ebola in certain countries. Europe re-posted many aid scenes sourced from Reuters, and those memes were also re-posted by the U.S. The third peak (from 10/18 to 10/28) contains many small visual memes, some of them sourced from U.S. news. For example, they include the CDC announcement for Ebola care and the fourth U.S. patient's arrival at Atlanta. Other visual memes in this peak are again related to aid scenes in Ebola-affected countries.

In addition, we note that the China-Both Spearman correlation coefficient is much lower than its Pearson correlation coefficient. This is due to the large inconsistency of the ranks of visual memes between China and Both. Many China videos contained visual memes of a conference held by the National Search and Rescue Agency of Indonesia, which were not stressed in U.S. videos. So this particular visual meme ranks lower in Both than in China. Additionally, there were visual memes reporting on the Chinese navy joining the

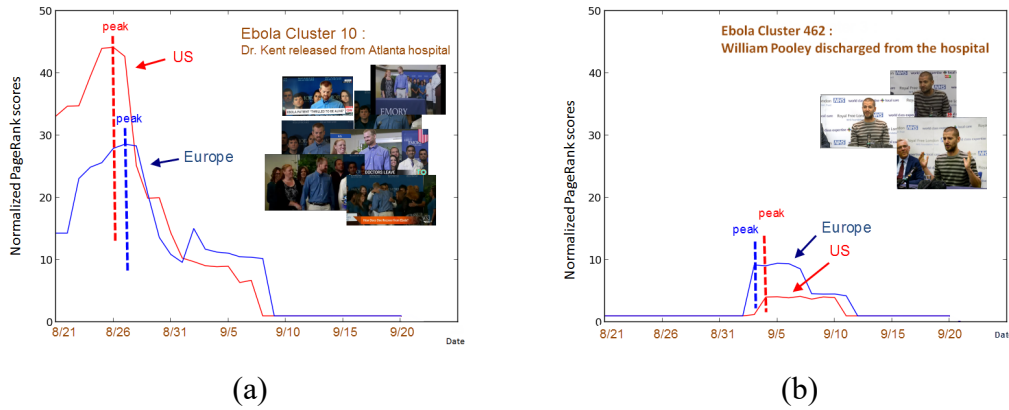


Figure 5.4: Timing of visual meme propagation. (a) Coverage of American Dr. Kent Brantly: the peak of the U.S. curve is ahead of Europe. (b) Coverage of British nurse William Pooley: the peak of the European curve is ahead of the U.S.

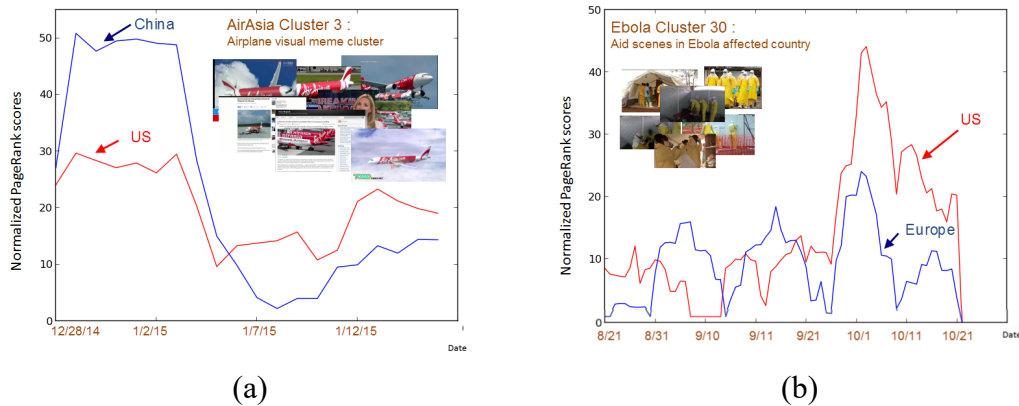


Figure 5.5: Long-lasting visual meme clusters. (a) Airplane visual meme from AirAsia. (b) One of aid scenes from Ebola.

rescue, and some other videos with interviews of Chinese and Indonesian officials, also not covered in the U.S. And in general, U.S. visual meme influence is more consistent with Both visual meme influence, partly since the U.S. media dominates the world media, and partly since the U.S. media is more effective in reporting influential news events in other countries.

5.3.3 Single Visual Meme Tracking

(A) U.S. ahead of Europe

We note that for some visual memes the peak of the U.S. curve of normalized PageRank score is ahead of Europe. For example, we note a particular Ebola subevent: the American Dr. Kent Brantly is released from an Atlanta hospital. Dr. Kent's recovery was important because it represented the success of the experimental drug, ZMapp. Therefore, this news became worldwide, but it originated in the U.S. Following U.S. media, the European media also covered this news story. (Figure 5.4 (a))

(B) Europe ahead of U.S.

In other cases, the peak of the Europe curve is ahead of the U.S. We note another particular Ebola subevent: the British nurse William Pooley is discharged from the Royal Free hospital. However, we also notice that this visual meme was less influential than other memes in U.S., probably because at this same time there was also another Ebola case in U.S. (Rick Sacra, an American missionary doctor) and the U.S. media gave precedence to it. (Figure 5.4 (b))

(C) Visual memes with long lifespan

Some visual memes last much longer than others. They were continually re-posted throughout the entire news event lifetime, and became a type of signature for the event: for example, aid scenes for Ebola, and airplane images and water scenes for AirAsia. In Figure 5.5 (a), we note that the visual meme influence scores of airplane scenes in China were much higher than those in U.S., even from the very beginning. Nevertheless, these airplane visual memes were still highly ranked in U.S. news coverage, and they persisted

over weeks in both. Figure 5.5 (b) shows one of visual meme clusters of aid scenes. Although its influence fluctuated over time for many weeks in both cultures (and at times was notably out of phase across them), they were always present during our tracking period.

5.3.4 Visual Meme and View Count Correlation

We hypothesized that visual meme influence may be related to video view counts obtained from video metadata. Figure 5.6 plots, on a log-log scale, influence versus view count, for each visual meme in our system. Influence here is defined as the average influence during a visual meme’s lifespan, and view count is defined as the average view count of videos containing this visual meme. The graph failed to support this hypothesis, as did its derived statistics.

We used the standard EM algorithm to fit a two-dimensional Gaussian distribution for the plot, starting with a random seed. The resulting μ and Σ are:

$$\mu = [8.4728, 1.6328], \quad \Sigma = \begin{bmatrix} 4.0746 & -1.6245 \\ -1.6245 & 6.4940 \end{bmatrix} \quad (5.5)$$

We believe the explanation is as follows. We note that influence is the result of a producer process, whereas view count is the result of a consumer process. In the real world, they do not appear to be strongly coupled to each other: a repeated visual meme is not necessarily viewed more because of the repetition. We note that not only is the covariance matrix nearly diagonal, suggesting independence, but also both processes are independently log-normal. Such distributions are typically the results of “multiplicative” processes, where several independent considerations must all occur together in order to achieve a maximum

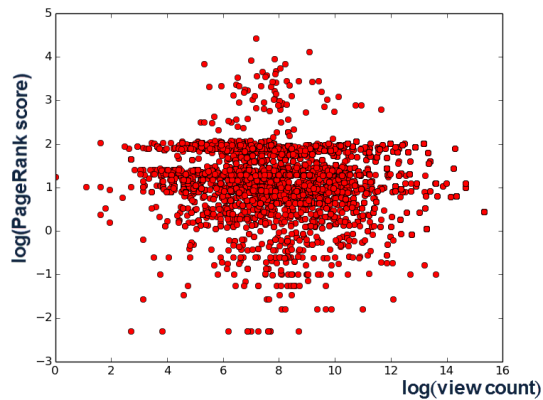


Figure 5.6: Log-log plot of visual meme influence (vertical) versus view count (horizontal), showing independence. Both influence and view count also follow log-normal distributions marginally.

effect. Since the probabilities of these considerations multiply, their logs add, and approach a normal central limit. Visual memes appear to become influential when the several causes of virality (for example, emotional impact, good photographic properties, regional interest, etc.) all co-occur. Likewise, videos can appeal to viewers for many reasons (political alignment, novelty, the resolution of fear, etc.), but these reasons are not necessarily the same as the ones for the re-posting of individual visual memes.

5.4 Future Work

We observe that many visual memes tend to be re-posted only in videos on specific topics. For example, visual memes related to experimental drugs are re-posted most often in videos about different Ebola-infected people. However, it is hard to automatically analyze these relationships without additional semantic (textual) information. We therefore plan to fuse visual memes with their related tag information for further analysis.

We showcased cross-cultural visual meme tracking by using two specific cultures for

each of two specific news events (Ebola and AirAsia), over a relatively short expanse of time. However, our proposed model is easily extended to multiple cultures and longer time ranges. We anticipate establishing additional interesting cultural-specific phenomena.

Cross-cultural Video Annotation

6.1 Culture-specific Tag Detection

There are many video archives on the Web (microblogs, news video archives and media-sharing websites such as YouTube [31, 80, 87]) that cover the same international human-interest events like health epidemics, elections, terrorism, financial crises, transportation disasters, or international sports. However, they are created, remixed, and maintained in different countries with differing cultural or political points of view. IR researchers have long been interested in summarizing similarities and differences among related documents [55]. In particular, Nakasaki *et al.* [60] analyzed the text portion of multimedia pages, and cross-culturally compared their expressed facts and opinions. Others have observed that both general users and news agencies tend to create “curated selections” based on what they like or think important [73]. For example, the re-posting of a visual meme [87] has been shown to be an implicit statement of the relevance of a video object. But no comprehensive study of culture-specific tags for news videos has yet emerged.

As a specific example from our work, we have been tracking the event, “Ebola”. One of its sub-events was the Sept. 12, 2014, news conference by the World Health Organization (WHO), which reported on world-wide support of the nations affected by the virus. We

have automatically noted some differences amongst the multimedia coverage of this sub-event. Most U.S. media mentioned “Cuba”, “WHO” and “Ebola” in their titles of video archives of this event, and had additional video clips illustrating Cuba’s support for Sierra Leone. However, nearly no Chinese media mentioned “Cuba”; instead, they focused on Margaret Chan, the Director-General of WHO, and on China’s aid to Africa. Since there was a single sub-event, the unabridged media source of the conference was the same, but the two countries extracted differing short video segments to illustrate their reports, augmenting them with additional segments and culture-specific texts.

We propose to detect culture-specific tags (textual) for news videos (visual) of human-interest international events, which can show culture-specific points of view. By representing a video into keyframes (images), our task is similar to regular image-text retrieval tasks, where one core problem is how to measure the semantic similarity between visual data (e.g., an input image or region) and text data (a sentence or phrase). A popular solution is to learn a joint embedding for visual and textual features into a shared latent space, where vectors from the two different modalities can be compared directly. This space is usually of low dimension and is very convenient for visual-textual retrieval. However, none of these algorithms extend tag retrieval in a cultural settings.

Several recent embedding methods learn a joint embedding space using Stochastic Gradient Descent with a ranking loss. WSABIE [83] and DeViSE [28] learn linear transformations of visual and textual features to the shared space. They use a single-directional ranking loss that applies a margin-based penalty to any incorrect annotations that get ranked higher than correct ones for each training image. A few other works have proposed a bi-directional ranking loss. In addition to ensuring that correct sentences for each training

image get ranked above incorrect ones, this also ensures that for each sentence, the image described by that sentence gets ranked above images described by other sentences [41] [42] [46] [74] [81]. However, these methods require sampling negative matches in addition to positive matches from data.

An alternative method to ranking loss is CCA-based embedding, which finds linear projections that maximize the correlation between projected vectors from the two views. In our cases, since it's not reasonable to define positive and negative samples for a culture-specific tag, we propose to learn a cross-cultural joint embedding via CCA-based approaches.

In this chapter, we learn “two-view pair-pair” embeddings and “three-view” embeddings, to detect culture-specific tags for news videos (sequences of images). In the former, we assume that we have initial image-text embeddings (“pairs”) for each of two different cultures (“views”). In both image-text embedding spaces, the relationship of a (image, text) pair is similar to that of a bigram in a monolingual embedding. After deriving these pairs in these two image-text embedding spaces, we further project relevant pairs in the two cultures to new joint embedding space via (deep) CCA for detecting culture-specific tags. In the latter, we assume that we have three initial embeddings (“views”): a textual embedding in culture 1 (say, U.S.), a textual embedding in culture 2 (say, Europe), and an images embedding (the union of the images in culture 1 and 2). we then learn a shared representation, G , for these three views via Generalized CCA [33], which generalizes embeddings to more than two sets of random variables, whereas (plain) CCA is limited to two sets.

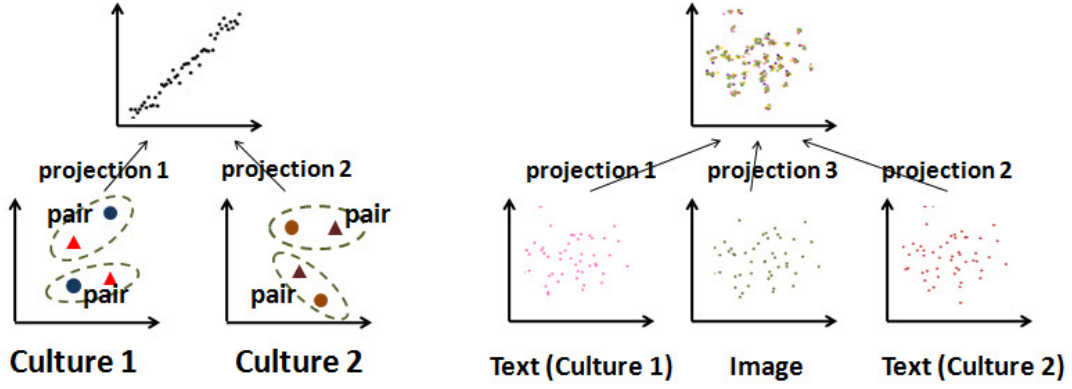


Figure 6.1: Left: Two-view Pair-Pair embedding via (deep) CCA. Right: Three-view embedding via GCCA. In the Two-view Pair-Pair embedding, each circle represents an image, and each triangle represents a text description, as a pair is a (image, text) pairing. Here we map pairs from two different cultures (culture 1, culture 2) into a joint embedding space via CCA. In Three-view embedding, we project data of three views (texts in culture 1, texts in culture 2, all images) into a joint embedding space.

6.1.1 Embedding Methods

In this section, we formalize the culture-specific tag detection task, and demonstrate two approaches to it: Two-view Pair-Pair embeddings, and Three-view embeddings. We represent a video as a sequence of keyframes, where each keyframe is an image. We define an image v in culture c as v_{ci} , with textual description t_{ci} , where i is an index. Let m and n denote two different cultures. Then, V_m, T_m, V_n, T_n denote the collections of images and textual descriptions in culture m and n separately. Our goal is then to detect the set of culture-specific descriptions $\{t_{nk}\}$ in culture n that is most accurate for an image v_{mj} in culture m .

6.1.1.1 Two-view Pair-Pair Embedding via (Deep) Canonical Correlation Analysis

We define E^c as image-text embedding space of culture c , and each image and its relevant text description is a pair (v_{ci}, t_{ci}) in c . We assume that we have initial image-text

embeddings for two cultures separately, denoted by E^m and E^n , and that we have the set of near-duplicate keyframe pairs $\{(v_{mj}, v_{nk})\}$ across cultures. We then derive a pairing of the image-text pairs, $\{((v_{mj}, t_{mj}), (v_{nk}, t_{nk}))\}$. Our goal is to find pairs of projections that maximize the correlation of relevant image-text pairs from the two cultures (two views); see Figure 6.1 Left. One popular method for two-view representation learning is canonical correlation analysis [34]. Given two sets of random vectors, the CCA objective is to find the linear combinations of the two views which have maximum correlation with each other. When establishing input vectors to CCA, each pair of pairs $((v_{mj}, t_{mj}), (v_{nk}, t_{nk}))$ can be further represented as the three ordinary pairs that exploit already known image-image and image-text matchings, leaving the cross-cultural text-text pairings to be discovered. Therefore, we use $(v_{mj}, v_{nk}), (v_{mj}, t_{nk}), (t_{mj}, v_{nk})$ as input to CCA.

Let $X \in \mathbb{R}^{D_x}$ be the collection of left elements of all pairs and $Y \in \mathbb{R}^{D_y}$ be the collection of right elements of all pairs. The object function of CCA is to find $u_x \in \mathbb{R}^{D_x}$ and $u_y \in \mathbb{R}^{D_y}$ such that projections of X, Y onto u_x, u_y are maximally correlated:

$$\begin{aligned} (u_x^*, u_y^*) &= \underset{u_x, u_y}{\operatorname{argmax}} \operatorname{corr}(u_x^T X, u_y^T Y) \\ &= \underset{u_x, u_y}{\operatorname{argmax}} \frac{u_x^T \sum_{xy} u_y}{\sqrt{u_x^T \sum_{xx} u_x u_y^T \sum_{yy} u_y}} \end{aligned} \quad (6.1)$$

where \sum_{xy} is cross-view covariance and \sum_{xx} and \sum_{yy} are within-view covariances.

The optimal k -dimensional projection mappings are given in closed form via the rank- k singular value decomposition (SVD) of the $D_x \times D_y$ matrix $\sum_{xx}^{-1/2} \sum_{xy} \sum_{yy}^{-1/2}$ [56].

Deep Canonical Correlation Analysis

Since a linear feature mapping is often not powerful enough to capture hidden non-linear relationships within data, Deep CCA was proposed to solve various problems such as image classification, image-text retrieval and speech recognition [4] [54] [82] [89]. In the DCCA model, two deep neural networks f and g extract features from view X and view Y , respectively, and are trained to maximize the correlations between the outputs of the two views. Applying f to x and g to y in Equation (1), then $F = f(X)$ and $G = f(Y)$, and the neural network weights and linear projections are optimized together using the objective:

$$(\mathbf{W}_f^*, \mathbf{W}_g^*, u_f^*, u_g^*) = \underset{u_f, u_g}{\operatorname{argmax}} \operatorname{corr}(u_f^T F, u_g^T G) \quad (6.2)$$

The weights, \mathbf{W}_f , \mathbf{W}_g , of the neural networks can be trained through standard back-propagation to maximize the CCA objective. In this work we follow the mini-batch stochastic gradient descent-like approach in [82] to solve for the optimal weights of neural networks and projection mappings.

6.1.1.2 Three-View Embedding via Generalized Canonical Correlation Analysis

Assume now we have three views instead: 1. T_m is the set of text description from culture m . 2. T_n is the set of text description from culture n . 3. $V_{mn} = V_m \cup V_n$ is the union of images in culture m and n . We then derive the set of triplets $\{(t_{mp}, v_p, t_{np})\}$ from the three views.

Generalized Canonical Correlation Analysis is an extension of CCA, which addresses

the limitation on the number of views. Its objective is to find a shared representation G of J ($J \geq 2$) different views; see Figure 6.1 Right:

$$\begin{aligned} \min_{U_j \in \mathbb{R}^{d_j \times r}, G \in \mathbb{R}^{r \times N}} \sum_{j=1}^J \|G - U_j^T X_j\|_F \\ \text{subject to } GG^T = I_r \end{aligned} \quad (6.3)$$

where N is the number of data points, d_j is the dimensionality of the j th view, r is the dimensionality of the learned representation, and $X_j \in \mathbb{R}^{d_j \times N}$ is the data matrix for j th view. Solving GCCA requires finding an eigen-decomposition of an $N \times N$ matrix, which scales quadratically with sample size and leads to memory constraints. Unlike CCA and DCCA, which only learn projections or transformations on each of the views, GCCA also learns a view-independent representation G that best reconstructs all of the view-specific representations simultaneously.

Therefore, by training a common representation for triplets

$\{(t_{mp}, v_p, t_{np})\}$ via GCCA, we are able to detect culture-specific tags, starting from an image in either culture.

6.1.2 Experiments

We collected datasets for three international news events: Ebola Virus, AirAsia Flight 8501, and Zika Virus. They are long-term news events lasting 2 months to 1 year. For the Ebola news event, we have collected about 3100 videos and their metadata, in an approximate 1:3 (Europe:U.S.) ratio, in a date range from 8/21/14 to 11/30/14. For AirAsia Flight 8501

events, we have collected about 1000 videos and their metadata, in an approximate 1:1 (China:U.S.) ratio, in a date range from 12/28/14 to 1/15/15. For Zika Virus, we have collected about 1700 videos and their metadata, in an approximate 7:10 (South America:U.S.) ratio, in a date range from 12/01/15 to 2/15/16. Videos sourced from US, Europe, and South American were collected from YouTube, and we verified their posted location in the metadata. Videos sourced from China were collected from Baidu, the biggest Chinese video search engine in the world, which aggregates videos from Chinese online news channels and from Chinese video-sharing websites.

We decomposed each video into a sequence of keyframes, and removed duplicate keyframes within the same video. For Ebola Virus, we then have about 27,000 keyframes in U.S., and 9,000 keyframes in Europe. For AirAsia Flight 8501, we have about 4,300 keyframes in U.S., and 2,000 keyframes in China. For Zika Virus, we have about 61,000 keyframes in U.S., and 44,000 keyframes in South America. For the text descriptions of each video, if it is not already in English, we first translate it into English by Google Translate. Then we follow the standard NLP pipeline, by removing stop words and pre-processing extracted words by WordNet’s lemmatizer. Then, after removing low-frequency tags, for Ebola Virus, we have 3,582 tags in U.S., 2,002 tags in Europe, and 1,530 tags in both cultures. For AirAsia Flight 8501, we have 3,391 tags in U.S., 828 tags in China, and 434 tags in both cultures. For Zika Virus, we have 7,846 tags in U.S., 3,081 tags in South America, and 2,732 in both cultures.

For visual features, to get the best global image representation for each keyframe, following [81], we use the ImageNet-trained 19-layer VGG network. Following standard procedure, the original 256×256 image is cropped in ten different ways into 224×224 images,



Figure 6.2: Near-duplicate keyframes across cultures with different texts. Pair (A) is from AirAsia Flight. Description of US image give more detailed information. Pair (B) is from Zika virus. Description of South America image takes Zika more seriously.

and we average the image features over the ten crops. For textual features, we train culture- and event-specific word2vec embeddings, and we represent each tag as a 256-dimension word2vec feature.

6.1.2.1 Significant Differences in Tags from Different Cultures

We note that for Ebola Virus, we find 2,052 U.S.-specific tags and 472 Europe-specific tags; for AirAsia, 2,957 U.S.-specific tags and 394 China-specific tags; for Zika Virus, 5,114 U.S.-specific tags and 349 South American-specific tags. The tags from different cultures are quiet different, at least in number.

Chi-square Test of Homogeneity.

However, by intersecting the tags from different cultures, for Ebola Virus, we still have 1,530 tags appearing in both U.S. and Europe; for AirAsia, 434 tags in both U.S. and China; for Zika Virus, 2,732 in both US and South America. We therefore tested if the populations of the set of even the intersecting tags from different cultures have significant differences.

For each event, we conducted a chi-square test of homogeneity for the populations of the intersecting tags.

Assume that we have n tags, and 2 populations (cultures) c_1 and c_2 . The null hypothesis would state that each population would have the same proportion of observations of every tag. Thus,

$$H_0 : P_{t_1, c_1} = P_{t_1, c_2}$$

$$H_0 : P_{t_2, c_1} = P_{t_2, c_2}$$

...

$$H_0 : P_{t_n, c_1} = P_{t_n, c_2}$$

H_a : at least one of the null hypothesis statements is false.

The test statistic is a chi-square random variable χ^2 defined by:

$$\chi^2 = \sum_{c,t} \left[\frac{(O_{c,t} - E_{c,t})^2}{E_{c,t}} \right]$$

where $O_{c,t}$ is the observed frequency count in population (culture) c for tag t , and $E_{c,t}$ is the expected frequency count in population (culture) c for tag t .

For Ebola Virus the df is 15, 29, and $\chi^2 \approx 9,000$. The p-value is < 0.00001 . Since results are significant at $p < 0.01$, we reject the null hypothesis. Thus at least the proportion of one of tags has a significant difference across cultures. By looking at single tags, we note

that the tag “msf” which refers to “Medecins Sans Frontieres” has a proportion of 0.000139 in U.S., while it is 0.001250 (10 times higher) in Europe. The tag “cdc”, which refers to “Centers for Disease Control” has proportion 0.003217 in U.S., while it is 0.001811 (half) in Europe. “Vaccine” has proportion 0.001538 in U.S., while it is 0.003622 (twice) in Europe.

We also conducted chi-square tests of homogeneity for AirAsia and Zika Virus. For AirAsia Flight 8501, the df is 433 and $\chi^2 \approx 4,000$. For Zika Virus event, the df is 2,731, and $\chi^2 \approx 43,000$. Their p-value are both < 0.00001 , so we reject the null hypothesis here also. Figure 6.2 shows that even near-duplicate video keyframes can have different description in different cultures.

We note that, since we had translated everything into English, our result could be biased depending on the accuracy of translation. However, we found that Google Translate does quite well on news data. For example, the following description is one that has been translated from Chinese: “The aircraft lost contact; AirAsia aircraft lost; AirAsia Indonesia to Singapore; the flight lost.” Converting these English words into tags, they are both reasonable and accurate.

6.1.2.2 Experiments on Two-view Pair-Pair Embedding

In order to conduct experiments on the Two-view Pair-Pair embedding, we need to have initial image-text embeddings for two cultures separately, and a set of near-duplicate keyframe pairs across both cultures. We demonstrate how we generate these requirements.

Culture-specific Image-Text Embedding Spaces.

There are many state-of-the-art algorithms which can map visual and textual features into a joint embedding space [81, 41, 42, 46]. The inputs to those algorithms are images and texts within the same culture. We mostly follow the implementation in [81]. In their model, there are two branches in their deep network, one for images (V) and the other for text (T). Each branch consists of fully connected layers with ReLU nonlinearities between them, followed by L2 normalization at the end. Given an image, we extract the 4096-dimensional activations from the 19-layer VGG model. For each tag, we use the 256-dimensional word2vec feature. To pair each image with its description, we first randomly selected N ($N \leq 7$) tags from its description as inputs to the neural network. On the image side, the output dimensions of the two hidden layers we use are [2048, 512]; on the tag side, we found that using only one hidden layer is best, and its output dimension is [256]. In our implementation, our embedding dimension is 256, we tuned $\lambda_1 = 1.5, \lambda_2 = 0, \lambda_3 = 0.05$, and we usually observed convergence within 20 epochs.

For each news event, we generated two culture-specific image-text embedding spaces: for Ebola Virus: one for U.S., one for Europe; for AirAsia Flight 8501, one for U.S., one for China; for Zika Virus, one for U.S., one for South America. For each culture-specific embedding we trained, we randomly selected 1000 images with their tags, or 10% of dataset, whichever is smaller, as the test dataset. The average $Recall@10$ of our trained embeddings is 67%. This recall rate is somewhat lower than the state of the art because both the images and the tags are polysemic in our dataset.

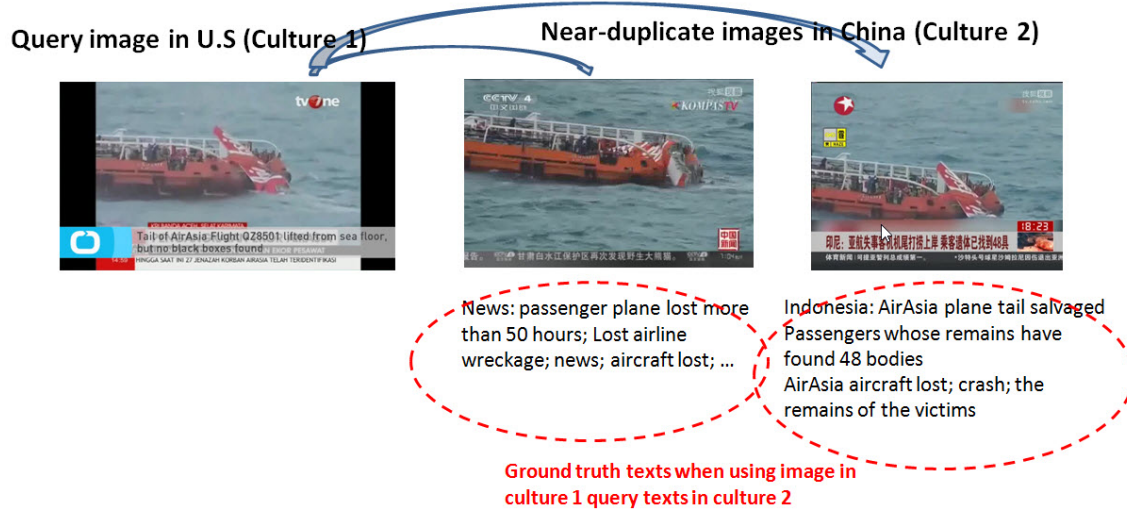


Figure 6.3: We use an image in culture 1 (say, U.S.) to query texts in culture 2 (say, China). We use the texts of near-duplicate images in culture 2 as ground truth to calculate the recall.

Near-duplicate Keyframe Pairs.

We established the set of near-duplicate keyframe pairs across cultures by first selecting an image in one culture and then calculating its euclidean distance with all images in the other culture. The image feature we used is the 4096-dimensional 19-layer VGG features. We kept only those cross-cultural pairs whose distance was below a threshold (20 ~ 45). Near-duplicate detection on VGG-19 features is quiet accurate, the accuracy is at least 90%, depending on the threshold we selected. We then manually remove any remaining incorrect pairs. For Ebola Virus, we then have 4,445 near-duplicate keyframe pairs across cultures; For AirAsia Flight 8501, 644 pairs; for Zika, 8,798 pairs.

For each near-duplicate keyframe pair across cultures (v_{mj}, v_{nk}) , we incorporate their textual features, and then derive the pairing of image-text pairs $((v_{mj}, t_{mj}), (v_{nk}, t_{nk}))$ across cultures.

Table 6.1: Tag detection performance of two-view pair-pair embedding. EU: Europe, CH: China, SA: South America.

		tag detection recall		
		R@1	R@5	R@10
Ebola Virus	US image query EU tags	8.2	29.8	40.3
	EU image query US tags	9.5	29.5	44.1
AirAsia Flight	US image query CH tags	7.6	18.8	29.1
	CH image query US tags	9.3	23.3	36.4
Zika Virus	US image query SA tags	11.8	31.2	54.1
	SA image query US tags	9.6	32.9	52.7

We then use these pairs of pairs as input to train the Two-view Pair-Pair embedding. We noted during our experiments on the CCA-based methods that pooling tags for each image into one single representational tag gives better results. Thus we average the 256-dimensional word2vec vectors of randomly selected N ($N \leq 7$) tags for each image.

We train the embedding via linear CCA first. We randomly select 10% pairings of image-text pairs as testing data, and the remaining 90% as training data. We evaluated our culture-specific tag detection performance by treating an image in culture 1 (say, U.S.) as a query to retrieve the texts in culture 2 (say, Europe). We defined the ground truth texts in culture 2 as the set of texts belonging to the images in culture 2 that were near-duplicates of the query image in culture 1. (See Figure 6.3.) Finally, we report $Recall@K$ ($K = 1, 5, 10$), or the percentage of image queries for which a correct match has rank at most K (ranked by cosine distance). We note that the recall via linear CCA embedding is low (around 15% $recall@10$), and this recall rate is similar to those of the linear CCA experiments conducted in [30]. That implies that our dataset could benefit from a nonlinear transform.

Pursuing this, we then trained the embedding via DCCA [82]. We selected 10% pairings

of image-text pairs as tuning data, 10% pairs as testing data, and the remaining 80% as training data. We used two hidden layers in the neural network, and tuned the layer widths using $\{128, 256, 512, 1024\}$. We observed that the result is better when using 256 and 512. We also tuned the output dimensionality, using $\{10, 20, 30, 40, 50, 70, 100\}$. We observed that using $30 \sim 50$ dimensions gave the highest recall. Table 6.1 shows tag retrieval performance. The average $recall@10$ for the detection of culture-specific tags is 40%. We note that AirAsia performance is lower than the other two events. This appears to be because we have too few near-duplicate image pairs (644 only).

6.1.2.3 Experiments on Three-view Embedding

For each news event, we have two cultures, $M = \{US\}$ and $N = \{Europe, China, SouthAmerica\}$. For the Three-view embedding, we generated our experiment dataset from near-duplicate image pairs across both cultures (see Section 4.2). For each of these near-duplicate image pairs, we then derived 2 triplets (t_m, v_m, t_n) and (t_m, v_n, t_n) , resulting in 8,890 triplets in Ebola; 1,288 triplets in AirAsia; and 17,596 triplets in Zika. For each event, let View 1 be the collection of all left elements from triplets, View 2 the collection of all middle elements, and View 3 the collection of all right elements. We then applied generalized canonical correlation analysis to solve for the embedding. To increase the performance, we used Principal Component Analysis on the 4096-dimensional VGG features to improve on the data sparsity, by compressing VGG features to 1000-dimensional vectors. We kept the text features, which are only 256-dimensional even after averaging, as is.

When applying GCCA, we again use $recall@K$ (the same as section 4.2) to mea-

Table 6.2: Tag detection performance of three-view embedding. EU: Europe, CH: China, SA: South America.

		tag detection recall		
		R@1	R@5	R@10
Ebola Virus	US image query EU tags	9.2	18.9	27.1
	EU image query US tags	7.9	11.6	19.3
AirAsia Flight	US image query CH tags	4.2	10.1	19.1
	CH image query US tags	5.3	13.3	21.4
Zika Virus	US image query SA tags	9.1	17.2	23.0
	SA image query US tags	8.6	16.3	24.8

sure the performance of tag detection. We sweep over several embedding widths using $\{10, 20, 30, 40, 50, 70, 100, 200\}$, and we note that using 50 or 70 can have somewhat higher recall. Table 6.2 shows the recall of the GCCA embedding. We note that, for Ebola, using U.S. images to query Europe texts has higher recall than using Europe images to query U.S. texts. We also observed that the performance depends on how diverse is the set of selected N tags taken from the descriptions of images. Two different images can have exactly the same set of tags if their descriptions are too short or too general. Notably for Ebola virus, there are more videos with short and general descriptions in U.S than in Europe. We find that the recall of AirAsia is much lower because its dataset is small and its tags are fewer than the other two events; the descriptions can not be easily distinguished from each other. (We partially avoided some of this problem in section 4.2, by using DCCA transform, whose non-linear transform is more flexible).

The recalls of GCCA are in general 10% higher than linear CCA, which is about (15%@10), but lower than DCCA. Since GCCA is also a linear transform, this suggests the relationship of image features and text features (basically, the relationship of VGG features to word2vec features) is strongly non-linear, at least in our datasets.

6.1.3 Conclusion

In this chapter, we introduced a new task—detecting culture-specific tags for news videos. We collected for our experiments the videos of three common real world news events, Ebola Virus, AirAsia Flight 8501, Zika virus, as covered in different cultures. We first showed by statistics that, for the same news event, tags of videos in different cultures have significant differences. We then demonstrated embedding algorithms based on CCA variants to detect culture-specific tags. Our recall is at most 54% $recall@K$, and we note that the relationship of image features and text features is non-linear, by comparing the detection results of DCCA against the linear CCA and GCCA. It therefore remains to future work to develop non-linear multiview ($views \geq 3$) embedding algorithms

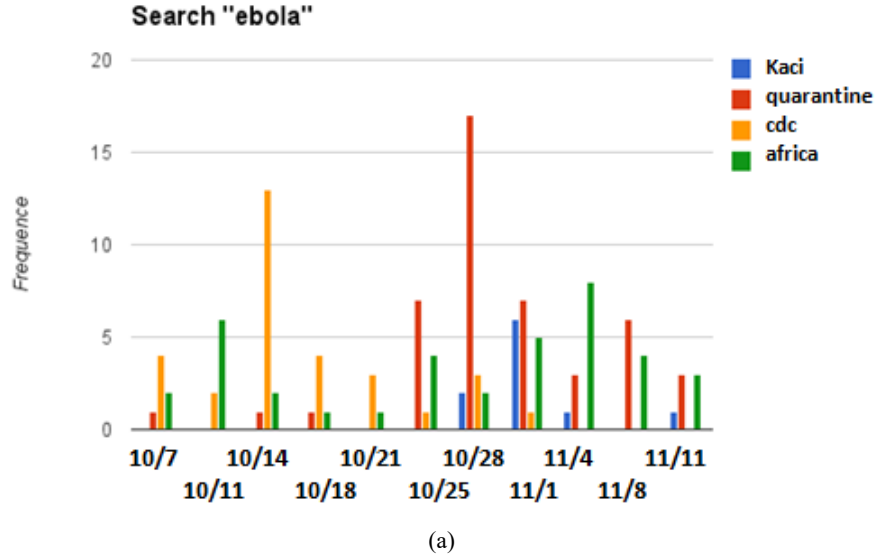
6.2 Cross-cultural Video Annotation

In section 6.1, we detected culture-specific tags for near-duplicate visual memes (segments) across cultures. However, for video segments only in one culture (say, China), we can not effectively map their tags into another culture (say, US). In this section, we demonstrate an algorithm which annotates the entire news video from different cultural points of view. By doing this, we can allow people in different lingual/cultural settings to retrieve the video archives posted by other countries or in other languages more precisely.

We build a computer system that begins with a news video in one culture (say, U.S.) that has no text or tags, and then locates through computer vision techniques those similar images in the image archives of another culture (say, China) that are annotated. The visual similarity allows us to derive from these annotations an understanding of the (Chinese) text and tags that would more properly summarize and tag the original (U.S.) video from the viewpoint of the other (Chinese) culture. If the original (U.S.) video already does have text and tags, our system can provide and contrast the cross-cultural differences in tags. Although we present a system using English and Chinese data, our algorithm can apply to any languages that have language-specific NLP algorithms and web search engines.

6.2.1 Data Collection

The data source we collected for this experiment are news videos archives for international events, consisting of news webpages (video-plus-text documents), which have richer text descriptions compared to YouTube videos. We hypothesize that text similarity often implies video similarity, so our initial query is verbal. But since a full event (“Ebola”) often has



(a)

Subevent cycle	Subevent title	Keyword set
8/24~9/8	William Pooley fighting ebola at London hospital	William, London, British, Pooley
8/25~9/2	New Ebola outbreak emerges in Congo	Congo, Democratic, Republic
8/30~9/4	Ebola experimental drug ZMapp cures monkeys	ZMapp, Experimental, Drug, monkey, hope
9/2~9/6	UN Ebola out of control	Ebola, Chan, WHO's, Health, Margaret
9/16~9/18	Obama to send 3000 military forces to fight Ebola	U.S., 3000, military, troops

(b)

Figure 6.4: (a) Keywords co-occurring with “Ebola”, which is an event (E). “Kaci”, “quarantine”, “cdc”, “africa”, are each a possible keyword for a subevent (e_i). (b) Collocated keywords detected for five time-limited subevents.

multiple subevents (“Ebola CDC”), we use the concept of news cycle as documented by Leskovec *et al.* [49], to locate time-limited subevent descriptions that are sets of distinctive additional words and phrases. Both that work and [87] noted that reposting probabilities tended to follow a power-law distribution, and that the majority of posting is within 2 to 3 days, with an occasional “echo” on weekends.

We mine the titles and text of the 3800 returns from Google search for the full event, for four events so far. Using “Ebola” as our example, Figure 6.4 (a) is a partial illustration of the distribution of four of these keywords over time; two in fact follow a power-law. Those keywords that do not, like “Africa”, tend to be nearly synonymous with the main event. Empirically we have found that a “good” subevent query usually consists of several

time-limited keywords, often together with the original event name, as discovered through the community analysis techniques described by Sayyadi *et al.* [68]. Some examples are shown in Figure 6.4 (b) also.

We measured the hypothesis that text similarity implies video similarity, using measures of graph connectivity amongst multimedia pages, computed separately for the sharing of (stemmed) textual words computed by thresholded cosine similarity, and of (near-)duplicate single frames computed by thresholded feature point similarity. These measures tended to correlate moderately well, with $\rho \sim 0.6$.

6.2.2 Document Clusters by Visual Similarity

Since our goal is to select candidate phrases for annotating the videos in the documents, we then explored the degree to which multimedia pages of an event (E) and its subevents (e_i) clustered visually. We consider each multimedia page to be a video-plus-text document ($d_j = \langle v_j, t_j \rangle$), where d is the webpage document, v is video and t is text. We define the news archives of subevent e_i naively as the set $a_i = \{d_1, d_2, d_3, \dots, d_j\}$, and the news archives of an event E as the set $A = \{a_1, a_2, a_3, \dots, a_i\}$.

We establish the visual correspondences between multimedia documents by using a variant of near-duplicate detection between the frames of one video against the frames of another. To avoid the cost of the full $\mathcal{O}(N^2)$ matches, we restrict the visual data to high-entropy I-frames. (Most such videos are in mp4 format, which makes this easy.) We do shot-detection on this reduced set within each video, keeping no more than three I-frames per shot.

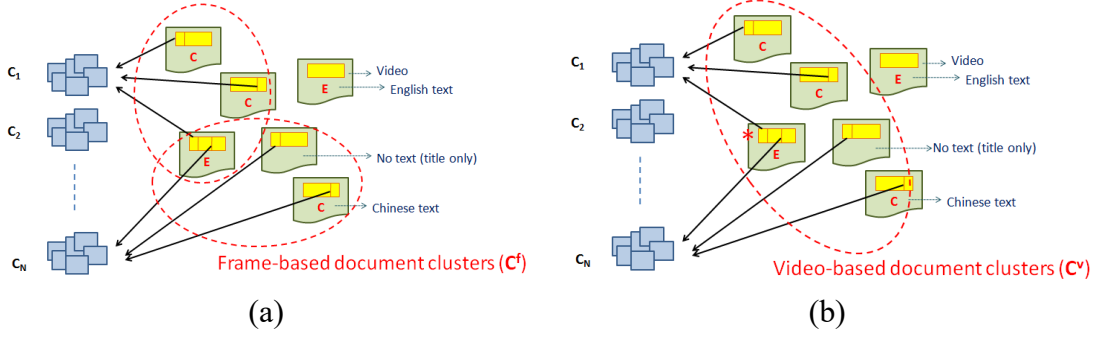


Figure 6.5: (a) Shows a frame-based document cluster: it contains all $d_j = \langle v_j, t_j \rangle$ containing keyframes in the same keyframe cluster. (b) Shows a video-based document cluster: it contains all frame-based document clusters which have near-duplicate keyframes from the video noted by “*”.

As shown in Figure 1.1, we then normalize the size of these keyframes, extract visual features (SIFT-BOF), and store each feature vector, from each of these keyframes, from each video in e_i , as a row in a descriptor matrix, which accumulates a total of m_i such rows. Using the FLANN library [59], we find the K nearest feature vectors to each feature vector, where $K = \sqrt{m_i}$, further limiting time complexity without compromising accuracy.

We now record these distances to these nearest neighbors in a keyframe-to-keyframe similarity matrix, which is then binarized via thresholding to yield a keyframe similarity graph. Its transitive closure is computed via a union-find algorithm [29] to find rough equivalence classes of near-duplicate keyframes.

Abstracting back up to the video level, we define a *frame-based document cluster* (Figure 6.5 (a)) to be the class of those multimedia documents whose videos contain one or more near-duplicate keyframes from the *specific* keyframe equivalence class associated with a given keyframe. Any text t_j of any $(d_j = \langle v_j, t_j \rangle)$, where d is a webpage of this cluster, therefore can be used as a source of potential annotations for any other document in the same cluster, irrespective of the language of t_j . We can similarly define a *video-based*

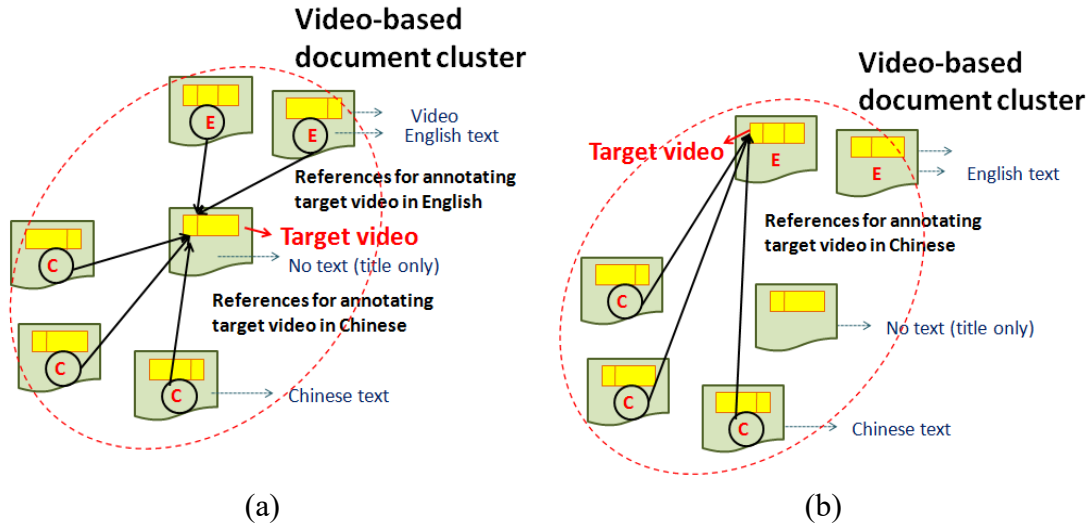


Figure 6.6: (a) Text from Chinese or from English reference video-plus-text documents used to annotate a target video with no tags or text. (b) Text from Chinese video-plus-text used to annotate a target video already annotated with English text.

document cluster (Figure 6.5 (b)) to be the class of those multimedia documents whose videos contain one or more near-duplicate keyframes from *any* keyframe equivalence class that is associated with a given video. Clearly this is a superset of frame-based document clusters, and therefore can be used as a source of potential annotations, also.

We have run a number of tests against these methods, comparing them to ground truth, and these clusters appear to be reasonably well-formed, but in need of some further refinement. We use them as the basis for our cross-cultural tagging method which provides additional semantic checks.

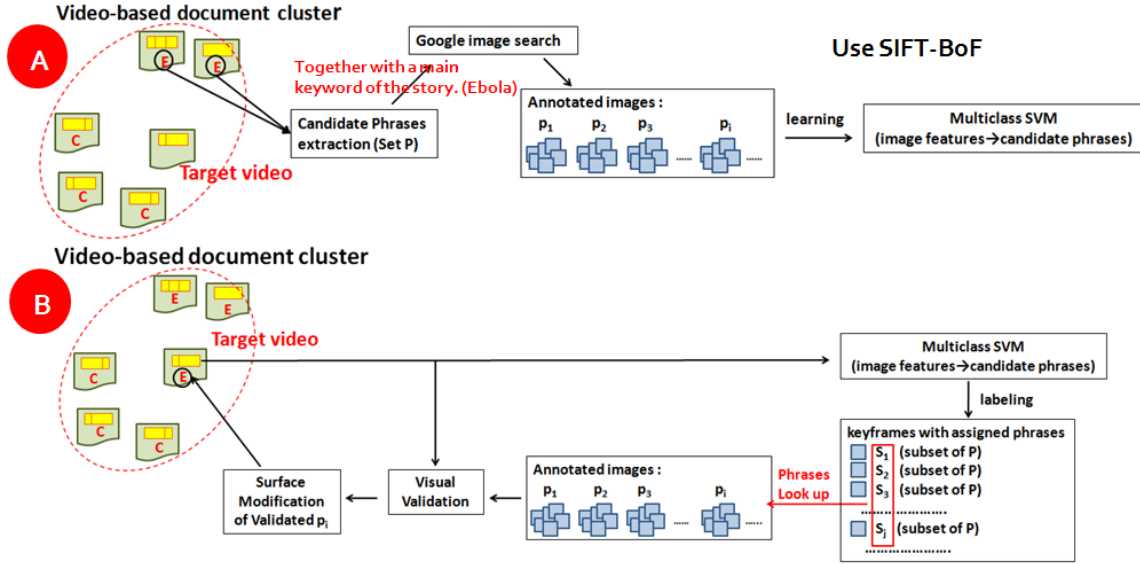


Figure 6.7: The processing pipeline of cross-cultural tag annotation. (A) Candidate phrases are extracted, and used as queries into an annotated image store. A multiclass SVM is trained to map visual features of these retrieved verification images into these phrases. (B) The SVM associates phrases with each keyframe of the target video. A keyframe’s phrase is verified if that target keyframe finds a visual match to a retrieved verification image that also carries the phrase.

6.2.3 Annotation Algorithm

We illustrate our annotation algorithm, which can be used to annotate webpage documents that have no annotations, or are annotated in a language unknown to the user (Figure 6.6). We define the video we want to annotate as the target video, and the references we use to annotate it as the reference archive.

Figure 6.7 shows the processing pipeline. We collect the reference archive from the video-based document cluster of the target video. Then, we use keyword extraction algorithms to select candidate annotation phrases from the texts of this cluster. We validate individual candidate phrases by collecting further evidence: we use the candidate phrases to retrieve additional images using a search engine, and match the visual content of the target video with those additional validation images. Finally, we produce final tags by some

filtering and surface text modification.

6.2.3.1 Candidate Phrase Extraction

We select candidate phrases $P = \{p_1, p_2, p_3 \dots\}$ from the reference archive in a language-dependent manner, where p_i represents a *phrase* in English, but a *phrase set* in Chinese, as determined by examining the explanatory structure of each language. If the target video has no text and tags, we select texts in the language we want to use to annotate it. Otherwise, we can use English text to produce English tags for Chinese source videos that have text, and vice versa. Although we use English and Chinese as examples, we can apply extraction algorithms in any language based on the languages and types of text sources. The extraction result usually contains phrases with strong cultural points of view if those phrases are recognized as important by the algorithms.

Candidate Phrase Extraction from English Texts We use RAKE [12], Rapid Automatic Keyword Extraction, for extracting candidate phrases from English texts. According to its patent application—and our experience—RAKE is particularly useful for (English) documents, like webpage texts, that do not necessarily follow grammatical conventions. We then compute, over the phrases, word frequency (number of occurrences) and word degree (sum of number of co-occurrences with other words in phrases) by standard NLP processing. We rank each phrase by average degree-to-frequency ratio, $\sum_{w \in p_i} \frac{\text{deg}(w)}{\text{freq}(w)}$, which favors words that predominantly occur in longer phrases. One sample extraction result is shown in Table 6.3.

English Phrases	RAKE score
amazing william pooley ebola patient estimates suggest	45.83
every day helping sick people	20.73
world class “care”	19.44
world health organization doctors	16.50
charity medecins sans frontieres	16.00
main story start quote	16.00
special isolation unit	9.00
experimental drug zmapp	9.00
first british person	9.00

Table 6.3: English candidate phrases extracted from news video archives by RAKE algorithm, from “Ebola” news on Sept. 3, 2014, using query: *British Ebola survivor William Pooley*.

Chinese Phrase sets	Google Translation
埃博拉、病患、威廉、普利、痊愈	ebola, patients, william, pooley, heal
威廉、埃博拉	william, ebola
治療、埃博拉、病患、隔離	therapy, ebola, patients, isolation
威廉、普利、ZMapp	william, pooley, zmapp

Table 6.4: Left: Chinese candidate phrase sets extracted from news video archives by Jieba system, from “Ebola” news on Sept. 3, 2014, which is related to *British Ebola survivor William Pooley*. Right: Google translation into English, except “Plymouth” has been corrected to “Pooley”.

Candidate Phrase Extraction from Chinese Texts Chinese uses a logographic system for its written language. We use Jieba Participle¹ to parse Chinese texts into phrases, by recognizing n-grams patterns, which are then ranked by the TextRank algorithm [58]. But because we observe that single Chinese phrases generally are not long enough to capture the semantics of the text, we bundle together the extracted candidate phrases from the same paragraph as a candidate phrase set instead. One sample extraction result is shown in Table 6.4.

¹[HTTPS://pypi.python.org/pypi/jieba/](https://pypi.python.org/pypi/jieba/)


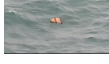




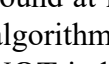
Keyframe	Coarse-level labeling → Fine-level verification (HIT or NOT)
	world class care ebola → NOT 3000 people infected ebola → NOT
	亚航 (AirAsia), 印尼 (Indonesia), 发现 (found), 物体 (objects), 确认 (confirmed), 失联 (lost contact), 航班 (flight), QZ8501 → HIT(exact) 亚航 (AirAsia), 搜救 (Search and Rescue), 发现 (found), 漂浮物 (floating debris), 疑似 (suspected) → HIT(exact)
	main story start quote ebola → NOT amazing william pooley ebola patient estimates suggest → HIT(exact) world class care ebola → NOT
	亚航 (AirAsia), 失事 (accident), 班机 (flight), 打捞 (salvage), 遗体 (remains), 当局 (authorities), 飞机 (aircraft), 残骸 (wreckage) → HIT(exact) 亚航 (AirAsia), 印尼 (Indonesia), 当局 (authorities), 黑匣子 (black box) → HIT(exact)
	royal free hospital ebola → HIT(exact) special isolation unit ebola → HIT(exact)
	亚航 (AirAsia), 黑匣子 (black boxes) → HIT(inferred)
	amazing william pooley ebola patient estimates suggest → HIT(exact) every day helping sick people ebola → NOT experimental drug zmapp ebola → HIT(exact)
	世界杯 (World Cup), 诺伊尔 (Neuver), 出色 (oustanding) → HIT(exact)
	charity medecins sans frontieres ebola → HIT(inferred)
	世界杯 (World Cup), 德国队 (German team), 决赛 (finals), 击败 (beat), 阿根廷 (Argentina), 夺冠 (win), 替补 (substituted), 格策 (Goetze), 大力神杯 (World Cup) → HIT(inferred)
	airstrikes, ISIS → HIT(inferred)
	美军 (US), 一箱 (a box), 落入 (fall into), 伊斯兰 (Islam), 武装 (armed), 物资 (supplies) → HIT(exact)

Figure 6.8: Keyframes with assigned verifying phrases. **HIT(exact)** indicates algorithm found at least one exact image match among annotated images. **HIT(inferred)** indicates algorithm failed to find an exact image match, but recognized similar image semantics. **NOT** indicates neither was found. Google Translate result is provided in parens for each Chinese word.

6.2.3.2 Candidate Phrase Verification

Both the clusters of frames and the extraction of candidate phrases are approximate. We use the following verification method in order to examine the suitability of each candidate phrase (or phrase set) p_i in $P = \{p_1, p_2, p_3 \dots\}$. We collect a new verification set of images, retrieved by using each p_i as a query into an annotated image store (of about 100 images for each p_i). Then we evaluate the visual similarity of these new verification images to the keyframes $F = \{f_1, f_2, f_3 \dots\}$ of the target video. If we observe a similarity between f_i

(one of the target keyframes) and one of the images retrieved by p_i (one of the candidate phrases), we recognize p_i as an appropriate annotation for the target video.

The verification therefore proceeds via a series of image and text interactions. First, candidate phrases retrieve a set of verification images. Then, a machine learning method finds the reverse mapping, from these verification images back to these phrases. Next, this reverse mapping is applied to the target keyframes, so they too now carry candidate phrases. Lastly, verification occurs when there is both an image match and a phrase match between a verification image and a target keyframe. The overall algorithm is summarized in Algorithm 2, but this verification method requires a bit more detail and care.

Algorithm 2 Fine-level verification

```

for each keyframe  $k \in$  target video do
  for each candidate phrase  $p \in S(k)$  do
    if FindNearDuplicateImage( $k, \text{AnnotatedImages}(p)$ ) then
       $\text{Tags} = \text{Tags} \cup \{p\}$ 
    else
      for each visual topic  $vt \in \{\text{VisualTopics}\}$  do
        if  $\text{prob}(vt|k) > T$  then
          for each word  $w \in p$  do
            if  $w \in \text{KeyPhrases}(vt)$  then
               $\text{Tags} = \text{Tags} \cup \{w\}$ 

```

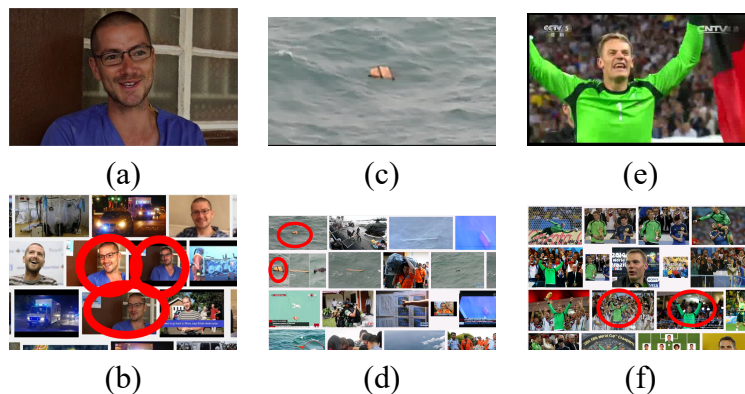


Figure 6.9: Image pairs recognized as matching by our algorithm. First row: Keyframes. Second row: Annotated images returned by web search engine. Keyframes (a)(c)(e) and images shown by red circles in (b)(d)(f) are respective pairings.

Candidate Phrase Verification Details: Coarse For the machine learning method we use a libsvm multiclass SVM classifier² trained on SIFT-BOF features of the verification image set. Thus, each verification image is represented as a histogram of visual words. To ensure that each target keyframe receives at least some phrases, we adjust the machine learning method to allow it to coarsely classify each keyframe by S_j , a subset of P , even though we know that these verification images are mixed across different topics and events [27, 38]. We also know that SIFT-BOF abstracts away much global spatial information. Candidate coarse-level labels derived for these and other matching are shown in Figure 6.8.

Candidate Phrase Verification Details: Fine The visual match of target keyframe to verification image is again through near-duplicate detection using SIFT image features. But it is tuned more strictly than that used in keyframe clustering, and it verifies the match by inspecting the homography between them. For efficiency, it prefilters image pairs by color features. Examples of these visual matches are shown in image Figure 6.9.

²<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

We have found, however, that for some candidate phrases there are no near-duplicate visual matches found between target keyframes and verification images, even though the candidate phrase appears appropriately descriptive, and even though it retrieves verification images strongly suggestive of its semantics. For an example, see Figure 6.10.

Analysis suggests that this is partly a result of the mixture of topics that characterize verification image sets. We therefore apply methods of Latent Dirichlet Allocation (LDA) [13] to tease out subsets of verification images whose SIFT-BOF visual word representations are locally more similar. Basically, we use a less supervised learning method instead.

LDA states that each “document” can be viewed as a mixture of a small number of “topics”, and that each “word” is attributable to one of the document’s topics. We apply LDA strictly visually, using the verification images (I) as documents, the as-yet-undetermined visual topics (VT) as topics, and the SIFT-BOF visual words (VW) as words. Then, for each candidate phrase p_i , we find the set of images I_i retrieved by p_i , and create the document-word matrix $M_{I_i, VW}$. Factoring this matrix into $M_{I_i, VT} \times M_{VT, VW}$ exposes the visual topics. This allows us to collect into a subset, for each visual topic VT_j , those images most strongly associated with the topic, whether they are target keyframes or verification images. If both of these associations are strong enough (i.e., greater than a fixed threshold), then we consider those individual candidate words associated with this collected subset to have been verified for the target keyframe.

As an example, we performed LDA on the verification images retrieved by “charity medecins sans frontieres ebola” and “launched 11 airstrikes overnight ISIS” shown in Figure 6.10. We found strong visual topics, and between them they verified some of the candidate words in each phrase.



Figure 6.10: Keyframe verified by semantic match. (a) Keyframe in Ebola video. (b) Image set returned by query term, “charity medecins sans frontieres ebola”. There are no exact images matches, but the semantics (group of people in masks, etc.) are similar. (c) Keyframe in ISIS video. (d) Image set returned by query term, “launched 11 airstrikes overnight ISIS”. There are no exact image matches, but more than 50% of retrieved image are about airstrikes of the same place from different perspectives and with different lighting.

6.2.3.3 Surface Modification

After verifying candidate phrases, we post-process them depending on their language. For Chinese results, we use the verified phrases directly because they are extracted via proper lexical patterns. For English, we note that Barr *et al.* [8] found that proper nouns constitute 40% of query terms, and proper nouns and other nouns together constitute over 70% of query terms. We therefore generate tags by using the NLTK³ package to select proper nouns first, noun phrases second, and any remaining parts of speech third.

³<http://www.nltk.org/>


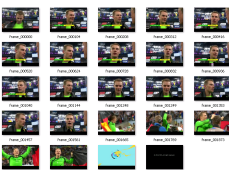

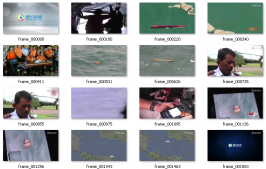
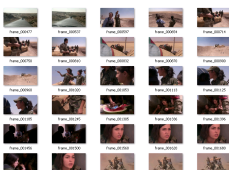
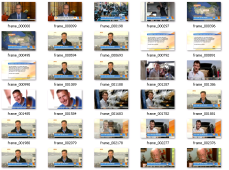
Key frames in the video	English tags	Chinese tags
	Cuba health minister Roberto chief Margaret Chan West Africa countries foreign professionals	埃博拉 (Ebola) 世卫 (WHO) 组织 (organization) 总干事 (Director-General) 冯富珍 (Margaret) 感染 (infection) 疫情 (epidemic) 赖比瑞亚 (Liberia) 治疗 (treatment)
Key frames in the video	English tags	Chinese tags
	World Cup FIFA Manuel Neuer golden glove ceremony	世界杯 (World Cup) 决赛 (finals) 德国队 (German team) 替补 (substituted) 格策 (Gotze) 击败 (beat) 门神 (keeper) 诺伊尔 (Neuer) 手套 (gloves) 本届 (current)
Key frames in the video	English tags	Chinese tags
	AirAsia Indonesian navy divers cockpit voice recorder passenger jet traffic controllers investigators black boxes flight	亚航 (AirAsia) 印尼 (Indonesia) QZ8501 失事 (wreck) 资料 (data) 记录仪 (logger) 黑匣子 (black box) 录音器 (recorder)
Key frames in the video	English tags	Chinese tags
	rescue director sb supriyadi child six weeks bodies AirAsia crashes	失联 (lost) QZ8501 碎片 (fragment) 印尼 (Indonesia) 搜救 (search rescue) 疑似 (suspected) 物体 (object) 发现 (discovery) 航班 (flight) 确认 (confirm)
Key frames in the video	English tags	Chinese tags
	female kurdish fighter Ceylan Ozalp protection Iraqi ammunition Ozalp goodbye northern syrian town	库尔德 (Kurdish) 科班 (Kobane) 伊斯兰 (Islam) 武装 (armed) 极端 (extreme) 女战士 (female soldiers) 叙利亚 (Syria)
Key frames in the video	English tags	Chinese tags
	Peter Kassig hostage Abdul-Rahman ISIS British militant worker 26-year-old jihadi fighters	伊斯兰 (Islam) 卡西格 (Kassig) 美国 (USA) 人质 (hostage) 处死 (killed) 蒙面 (masked) 刽子手 (Executioner) 口音 (Accent)

Figure 6.11: Examples of video annotation produced by our algorithm for four events: Ebola (top left), World Cup (top right), AirAsia (middle both) and ISIS (bottom both). Google Translate is provided for each Chinese tag, with a few corrections for proper names (e.g., “Ka Xige” to “Kassig”, etc.)

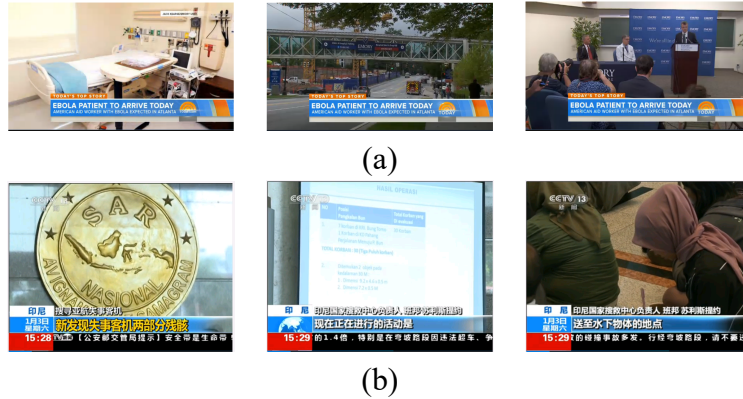


Figure 6.12: Keyframes from videos presented from the perspective of a different culture. The results are retrieved by our cross-cultural tags, from top 20 returns from Google video search. (a) Emory University Hospital from US media source. (b) News conference of National Search and Rescue Agency in Indonesia from Chinese media source. Neither segment appeared in the other country’s news coverage.

6.2.4 Results and Evaluation Methods

We collected from multimedia webpages about 1500 video-plus-text documents approximately in a 1:1 (US:China) ratio, from Google search⁴, for four human-interest international events: “Ebola”, “AirAsia”, “ISIS” and “World Cup”. For searching, we also specified a date range for each event: 8/15/14 ~ 12/15/14 for Ebola, 12/18/14 ~ 1/15/15 for AirAsia, 8/19/14 ~ 12/15/14 for ISIS, and 7/11/14 ~ 7/16/14 for World Cup. And, to increase the amount of near-duplicate video segments, we used queries that specified subevents, as detected using the results in Section 3.

We archived at most three video-plus-text documents per day for each event, after disabling all personalized search customizations. Each video-plus-text document contained one embedded video plus adjacent plain text descriptions. (We did not use closed captions since we found that they were neither accurate nor organized well; they were often in all up-

⁴Example US sources are Fox News, CNN, CBS, NBC news, YouTube, etc. Example Chinese sources are 央视网 (CCTV), 腾讯新闻 (News QQ), 中国新闻网 (China News), 新浪新闻 (Sina News), 凤凰卫视 (ifeng), etc.

per case, making proper name detection difficult; and they were offset by a varying amounts of time from their corresponding images⁵). We found 87 video-based document clusters, and successfully produced many cross-cultural tags; see Figure 6.11.

Our hypothesis is that cross-cultural tags can facilitate general searches for videos of an international event, by associating English and Chinese tags, and thus allowing the user to retrieve other imagery of the same news story from videos presented from the perspective of a different culture. For example, in the case of Ebola, Chinese users can use our system's tags to retrieve a video segment for introducing the isolation unit of Emory University Hospital in Atlanta, which Chinese media never focused on or presented (Figure 6.12 (a)); in the case of AirAsia, US users can use our system's tags to retrieve a more detailed video segment of news conference held by National Search and Rescue Agency in Indonesia, which US media never focused on or presented. (Figure 6.12 (b)).

We evaluated our hypothesis by examining our annotation results in four ways: by tag precision, tag cross-cultural frequency, user satisfaction, and retrieval performance.

⁵<http://www.bbc.co.uk/rd/publications>

	Average tag frequencies (normalized)					
	$\frac{ E }{ A_E }$	$\frac{ C^t }{ A_E }$	$Loss$	$\frac{ C }{ A_C }$	$\frac{ E^t }{ A_C }$	$Loss$
Ebola	2.35	1.82	23%	2.71	0.63	76%
AirAsia	1.91	1.52	21%	3.78	0.43	89%
ISIS	2.29	1.53	33%	2.43	0.24	90%
World Cup	1.95	0.87	55%	2.46	0.86	66%
Overall	2.00	1.3	35%	2.98	0.64	79%

Table 6.5: Average tag frequency per document for each event. Loss is calculated as $|origin - translated| / |origin|$.

6.2.4.1 Tag Precision

Given ground truth, the precision of the suggested tags can be computed in a straightforward way. We measured separately the precision of those tags derived from the reversed SVM mapping $HIT(exact)$, and those derived from the LDA clustering $HIT(inferred)$, as previously defined in Figure 6.8. The precision of $HIT(exact)$ was about 0.89 and that of $HIT(inferred)$ about 0.71, and most tags were the former. Many of the $HIT(inferred)$ were from aid scenes (Ebola), plane and black box discovery (AirAsia), airstrikes (ISIS), or single players (World Cup). In most cases, errors were due to violations of the assumption that the text of a webpage was related to their embedded videos or images. For example, there is a video in AirAsia that interviews travelers who were supposed to be on the flight. It is embedded in a webpage with the title, “Search for Missing AirAsia Plane”. The text of the webpage is not related to the embedded video, so our algorithm does not produce accurate tags.

6.2.4.2 Cross-Cultural Tag Frequencies

We quantified the cross-cultural frequencies of each set of tags we produced, and the differences among them. For a given video, we define E to be the set of English tags suggested

by our algorithm and C to be the set of Chinese tags suggested by our algorithm. Similarly, we define E^t to be the Chinese translation of E , and C^t the English translation of C . We denote by A_E an archive of English words and their frequencies, A_C a corresponding Chinese archive. We computed for both languages how similar the frequency of suggested tags—and their translations—are, relative to their archival frequencies. Specifically, for English, we compared $|C^t|/|A_E|$ to $|E|/|A_E|$ and $|E^t|/|A_C|$ to $|C|/|A_C|$, where $|\cdot|$ is a measure of usage. We noted that the first ratio, which measures cross-culture tagging, remains smaller than the second, which measures native tagging; in fact, the first ratio could be zero. We then define $Loss$ to measure the percentage loss of frequency of translated tags. Table 6.5 shows the results. The loss of frequency not only includes the loss caused by translation, but includes the loss caused by cross-cultural differences.

We noted two observations. Our first is that the $Loss$ of E^t is larger than that of C^t . In general, this was because English text provides more details on subjects and objects. For example, the English “Indonesian commander Gen Moeldoko” was usually “Indonesian commander” in Chinese; the English “voice recorder” was simply “recorder” in Chinese. In addition, the loss of E^t for AirAsia was severe, due to the many ways in which “AirAsia Flight QZ8501” was abbreviated in English: “AirAsia plane”, “flight”, “flight QZ8501”, “QZ8501”, “aircraft”, “jet”, “passenger jet”, etc. In contrast, the loss of C^t for AirAsia was much less, since in Chinese, “flight” is less likely to be selected as a keyword in Chinese due to the difference between language grammars. In fact, “亚航 (AirAsia)” can be used as a pronoun instead, and some tag sets did not include “flight” at all. Instead, about half of the text we collected put “失事 (crash)” between “亚航 (AirAsia)” and “客机 (Airliner)”, as “亚航失事客机 (AirAsia Crash Airliner)”, and this influenced keyword recognition. Our

second observation is that the *Loss* is often higher in events in which countries are in direct competition, and where the text reflects national interest or cultural viewpoint. As seen with World Cup, English descriptions were more generic and less involved, and translated less well than the Chinese tags did.

6.2.4.3 Tag Quality from User Studies

To explore the differences between the tag sets in different languages, we conducted two user studies, one to measure the subjective quality of the automatically generated tags, and the other to measure the interchangeability of tags from different cultures. These were designed to explore our initial intuitions derived from a study of our first event, Ebola. There, we found that: Chinese-language videos, even those originating in the U.S., used only Chinese names, not Americanized ones; both U.S. and Chinese videos tended to use only local technical terms (“ZMapp” versus “jk-05”); Chinese videos appeared to favor more “international” rather than local imagery; Chinese videos tended to focus on public impact, such as the danger of an escaped infected patient, whereas U.S. videos tended to focus on individuals; Chinese image libraries were not as richly annotated, as even the concept of “keyword” is not as well defined; Chinese Ebola video archives stressed their country’s aid to Africa, whereas U.S. ones stressed the problem of medical detection.

To detect these broader patterns, we selected 20 representative videos with their cross-cultural tags (Chinese and English), and created three test sets from them. Since the amount of data we had collected was approximately in an 8:4:4:4 ratio (Ebola:AirAsia:ISIS:WorldCup), our three sets were composed of 4:2:0:2, 4:0:4:0, and 0:2:0:2 videos. Each of 15 bilingual participants were given two sets, so that each set was seen by 10 participants.

	Selected from E source		Selected from C source	
	E (English)	E^t (Chinese)	C (Chinese)	C^t (English)
Ebola	3.50	3.70	2.96	3.05
AirAsia	2.68	1.98	4.10	3.95
ISIS	3.33	3.18	3.10	3.43
World Cup	3.68	4.00	2.33	3.68

Table 6.6: The average goodness of tags, which were ranked 1~5.

Goodness of automatically generated tags. For the first study, we asked: “How good do you think the following four sets of tags are, for the target video? (Please enter the score: 1~5.)”

We provided four sets of tags to the participants: English, English translated from Chinese, Chinese translated from English and Chinese. Table 6.6 shows the average of goodness of these four sets for each event. We note that if the tags that were generated from a source language were good, so were their translation. On average, the tags for Ebola and World Cup generated from English sources were better than those generated from Chinese source, but for AirAsia this was reversed. This appears to have been due, in part, to the richness of detail in Chinese AirAsia annotations. And for ISIS, both English and Chinese have culturally specific tags, so that the average of goodness of tags depends strongly on the preferences of participants.

We were surprised to notice that sometimes translated tags scored more highly than the original tags. We followed up with several of the participants, who gave a number of reasons. First, they tended to be forgiving of inelegant translations, such as “get well” instead of “recovery”. Second, some of them preferred Western-style proper names. This is seen in Ebola and World Cup, where there are a number of well-known named entities, and the score for C^t is higher than that of C . Third, some of the native Chinese were more comfortable with Chinese representations for less well-known proper names or more

Events	English	Chinese
Ebola	3.38	3.63
AirAsia	3.65	4.83
ISIS	2.70	3.25
World Cup	3.68	3.85
Overall	3.35	3.89

Table 6.7: Interchangeability of native tags with translated tags, for 20 videos of four international events. User evaluation of interchangeability was on a five-level likert scale, ranked 1~5.

specialized vocabularies, again as seen in Ebola and World Cup, where the score for E^t is higher than that of E . In contrast, AirAsia has few proper names and no specialized terms, and translations lose quality.

We also compared the result of this section with the result of section 6.4. We note that user preference can be used to tell in which language the visual content is more richly annotated, but this has little to do with tag retrieval performance.

Interchangeability of tags from different cultures. For the second study, we asked: “Are the two sets of tags interchangeable when tagging the same video? That is, can one set of tags be replaced with the other set without loss of precision?”

For each of the 20 videos, we provided the participants with two pairs of tag sets. The first pair were in English, tag sets E and C_T ; the second pair were in Chinese, tags sets C and E_T . The answers were given on a five-level Likert scale, from “Strongly disagree” to “Strongly agree”. Results are in Table 6.7. In general, the scores were 3~4, or in the range “Neither agree nor disagree” to “Agree”, showing partial interchangeability. It appeared on follow-up that there were differences in strategies among the participants. Some compared the two sets of tags by looking for commonalities between people, actions, and settings,

Events	E	E^T	C	C^T
Ebola	55%	38%	75%	3%
AirAsia	75%	40%	60%	13%
ISIS	65%	35%	74%	15%
World Cup	55%	55%	75%	40%
Overall	63%	42%	71%	18%

Table 6.8: The percentage of related videos retrieved by our cross-cultural tags that were in the top 10 returns from a Web search engine.

resulting in high scores. But others more stringently required that all given details needed to match. We notice that for ISIS, which suffered from more cross-cultural inconsistency, both interchangeability of English (2.70) and Chinese (3.25) tags are lower than the other three events. Overall, the interchangeability rating of Chinese tags, 3.89, is higher than that of English, 3.35, mostly because Chinese tags are more general and therefore suffer less in translation. In fact, in AirAsia, Chinese tags achieved the highest interchangeability, 4.83, meaning that Chinese native tags translated into English showed nearly no semantic differences.

6.2.4.4 Tag Retrieval Performance

Lastly, we measured how effective the generated tags were in retrieving videos within and across cultures. We measured (through visual inspection) the percentage of related videos retrieved within the top 10 returns from Web search engine using the tags as queries, after again disabling any customized search. We compared the retrieval performance to the translated tags (baseline), which can be directly translated from source language. The results are in Table 6.8. What stands out clearly is how poorly English translations of Chinese tags perform, again reflecting a relative lack of detail in Chinese annotations. Surprisingly, the tag

retrieval performance of E and E^T for World Cup are the same. This is because there are many more internal differences in the same video-based document cluster for World Cup from the U.S. source. For example, the video-based cluster of “closing ceremony song” contains some variations such as “Shakira song”, “Sharkira comparing to Pitbull”, “World Cup review” and “Shakira and her son”, so that the retrieved tags are more generic (so are their translation), giving few restrictions on selecting related videos from the retrieved list.

6.2.5 Discussion

Our algorithm successfully suggests cross-cultural tags for English and Chinese, through algorithms that appear to be applicable to other languages and cultures, and even to differences among American, Canadian, and British news coverage. We have shown that for human-interest international events, online annotated images and text can generate cross-cultural tags for unannotated videos, with a performance that exceeds the simple translations of similar videos. We established that near-duplicate frames and their tags are frequently re-used across videos in different cultures, but with detectably different frequencies.

We noted a number of similarities and contrasts. Because of the editing process, some visual images and textual tags are more frequent than others. For example, in Ebola, Obama’s speech; in AirAsia, rescue and search; in ISIS, propaganda videos released by ISIS; in World Cup, ceremonies and goals. On the other hand, we found that in Ebola, annotated U.S. videos were more numerous, and they tended to focus on individuals and specific spokespeople, whereas Chinese videos tended to focus on public impact. However, in AirAsia, there were more annotated Chinese videos, which focused more on rescue. In

further explorations, we have also noted similar asymmetries with coverage of the Chinese SARS epidemic, the Japanese 311 tsunami, but not as much with the world-wide H1N1 flu threat.

We also noted that popular search engines such as Google adopt their responses in a cultural way. In general, users need to type queries in English to get annotated English content. In the four events we analyzed, only the Chinese query, “夏奇拉世界杯 (Shakira World Cup)”, successfully retrieved better and longer videos annotated in English. This suggests that search engines would be more comprehensive and effective if they adopted methods such as ours, to establish tag correspondences between the language of the query and the language most associated with the event.

We anticipate that this work can be used to aid academic research in Journalism and in History, which currently focuses on the study of news differences in text, and, to a much lesser extent, in static imagery. Video objects are more difficult to analyze, but they are becoming more popular. Our methods, which we intend to improve and incorporate into a novel comparative cross-cultural video browser, should make this research more tractable.

6.2.6 Future work

In this chapter, we retrieved the links of multimedia webpages and evaluated tag retrieval performance on Google search, and used Google translated tags as baseline for comparison. Although the performance of cross-cultural tags would vary if using different search engines, we believe cross-cultural tags still allow more precise search than translated tags. Although the proposed algorithm is language independent, it may not be able to apply to

less influential international news events, or to a culture with fewer related news documents.

We are investigating these aspects.

Chapter 7

Future Work

We plan to further extend the work of this thesis in three main directions:

1. **Summarize the differences in news videos across cultures.**

Besides summarizing the commons, comparing and summarizing the differences in multilingual printed news is a popular research topic both in Journalism and in Information Retrieval. For example, Wan *et al.* [79] manually assign to each sentence a score that indicates how much the sentence contains differential information, compared to the other culture. They then use a ranking algorithm to find and cluster the most significant differences in order to produce a bias summary. Kwak *et al.* [47] reveal the structure of global news coverage of disasters and its determinants by using a large-scale news coverage dataset collected by the GDELT (Global Data on Events, Location, and Tone) project that monitors news media in over 100 languages from the whole world. However, these works are restricted to text analysis. We propose to compare and summarize the differences in multicultural news videos. This will allow user to know what important news videos might not be covered in your country, and to discover which news videos are being covered in certain locations.

2. **Incorporate other modalities in analysis.**

In this thesis, we focus on four modalities: textual, visual, time, and culture. How-

ever, there are some other modalities such as sentiment and action, which has attracted wide attention in multimedia researches. Take sentiment as an example, multimedia contents are tools for social media users to convey their sentiment, emotion and opinions; conversely, the sentiment has influence on topics of multimedia contents. It's better to incorporate other modalities in multimedia analysis.

3. Identify culture origins or channel origins of news reports.

This thesis demonstrates algorithms for news video analysis based on unsupervised co-clustering, tracking, and visual-textual matching. However, no algorithm was developed to identify (predict, or classify) the culture origins of news videos, and this will be an interesting topic to investigate in the future. In addition, we propose to further identify the channel origin of each news report (text description, images and videos). This task is similar to authorship attribution and prediction [75] which has been taking advantage of research advances in areas such as machine learning, information retrieval, and natural language processing. We propose to identify which channel a specific news report comes from, for example, CNN or Fox news.

Bibliography

- [1] Microsoft translator text api. <https://www.microsoft.com/en-us/translator/translatorapi.aspx>.
- [2] TRECVID Multimedia Event Recounting Evaluation Track, 2012. <http://www.nist.gov/itl/iad/mig/mer.cfm>.
- [3] YouTube.com: Statistics – YouTube., 2016. <https://www.youtube.com/yt/press/statistics.html>.
- [4] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning(ICML)*, pages 1247–1255, 2013.
- [5] A. Banerjee, S. Basu, and S. Merugu. Multi-way clustering on relation graphs. In *Proceedings of the 2007 SIAM international conference on data mining*, pages 145–156. SIAM, 2007.
- [6] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha. A generalized maximum entropy approach to bregrman co-clustering and matrix approximation. *Journal of Machine Learning Research*, 8(Aug):1919–1986, 2007.
- [7] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, et al. Video in sentences out. *arXiv preprint arXiv:1204.2742*, 2012.
- [8] C. Barr, R. Jones, , and M. Regelson. The linguistic structure of english web-search queries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP)*, pages 1021–1030, 2008.
- [9] A. Benton, R. Arora, and M. Dredze. Learning multiview embeddings of twitter users. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

- [10] K. Berberich, S. Bedathur, G. Weikum, and M. Vazirgiannis. Comparing apples and oranges: normalized pagerank for evolving graphs. In *Proceedings of the 16th international conference on World Wide Web*, pages 1145–1146. ACM, 2007.
- [11] S. Bergsma and B. Van Durme. Learning bilingual lexicons using the visual similarity of labeled web images. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1764. Citeseer, 2011.
- [12] M. Berry and J. Kogan. *Text Mining: Applications and Theory*. Wiley InterScience, 2010.
- [13] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [14] R. Cai, L. Lu, and A. Hanjalic. Co-clustering for auditory scene categorization. *IEEE Transactions on multimedia*, 10(4):596–606, 2008.
- [15] L. Cao, S. Chang, N. Codella, et al. IBM Research and Columbia University TRECVID-2012 multimedia event detection (MED) multimedia event recounting (MER) and semantic indexing (SIN) systems. In *Proc. TRECVID 2012 workshop. Gaithersburg, MD, USA*, 2012.
- [16] W.-S. Chu, Y. Song, and A. Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *CVPR*, 2015.
- [17] P. Clough, H. Müller, T. Deselaers, et al. The CLEF 2005 cross-language image retrieval track. In *CLEF 2005 Workshop Working Notes*, 2005.
- [18] P. Das, C. Xu, R. F. Doell, and J. J. Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2634–2641, 2013.
- [19] J. Deng, J. Krause, A. Berg, and L. Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [20] I. Dhillon, S. Mallela, and Modha. Information-theoretic co-clustering. In *KDD*, 2003.
- [21] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274. ACM, 2001.

- [22] C. Ding, W. P. T. Li, and H. Park. Orthogonal nonnegative matrix trifactorizations for clustering. In *KDD*, 2006.
- [23] D. Ding, F. Metze, S. Rawat, P. F. Schulam, S. Burger, E. Younessian, L. Bao, M. G. Christel, and A. Hauptmann. Beyond audio and video retrieval: towards multimedia summarization. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 2. ACM, 2012.
- [24] W. Dong and W.-T. Fu. Cultural difference in image tagging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 981–984. ACM, 2010.
- [25] E. Dumont and G. Quénot. Automatic story segmentation for tv news video using multiple modalities. *International journal of digital multimedia broadcasting*, 2012, 2012.
- [26] S. Essid and C. Févotte. Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring. *IEEE Transactions on Multimedia*, 15(2):415–425, 2013.
- [27] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google’s image search. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1816–1823. IEEE, 2005.
- [28] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. DeViSE: A deep visual-semantic embedding model. In *Advances in neural information processing systems (NIPS 2013)*, pages 2121–2129. 2013.
- [29] B. A. Galler and M. J. Fisher. An improved equivalence algorithm. *Communications of the ACM*, 7(5):301–303, 1964.
- [30] Y. Gong, L. Wang, M. Hodosh, and J. Hockenmaier. Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision (ECCV)*, pages 529–545, 2014.
- [31] Google news. <https://news.google.com/>.
- [32] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

- [33] P. Horst. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17(4):331–347, 1961.
- [34] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.
- [35] G. Irie, D. Liu, Z. Li, and S.-F. Chang. A bayesian approach to multimodal visual dictionary learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 329–336, 2013.
- [36] M. Jiang, A. Beutel, P. Cui, B. Hooi, S. Yang, and C. Faloutsos. A general suspiciousness metric for dense blocks in multimodal data. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 781–786. IEEE, 2015.
- [37] W. Jiang, S.-F. Chang, and A. C. Loui. Active context-based concept fusion with partial user labels. In *Image Processing, 2006 IEEE International Conference on*, pages 2917–2920. IEEE, 2006.
- [38] F. Jing, C. Wang, Y. Yao, K. Deng, L. Zhang, and W.-Y. Ma. IGroup: web image search results clustering. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 377–384. ACM, 2006.
- [39] B. Jou, H. Li, J. G. Ellis, D. Morozoff-Abegauz, and S.-F. Chang. Structured exploration of who, what, when, and where in heterogeneous multimedia news sources. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 357–360. ACM, 2013.
- [40] K. Kampa. Hierarchical biclustering toolbox. <https://sites.google.com/site/kittipat/hierarchical-biclustering>, November 2008.
- [41] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, 2015.
- [42] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, pages 1889–1897, 2014.
- [43] J. R. Kender and B.-L. Yeo. Video scene segmentation via continuous video coherence. In *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, pages 367–373. IEEE, 1998.

- [44] M. U. G. Khan and Y. Gotoh. Describing video contents in natural language. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 27–35. Association for Computational Linguistics, 2012.
- [45] M. U. G. Khan, L. Zhang, and Y. Gotoh. Towards coherent natural language description of video streams. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 664–671. IEEE, 2011.
- [46] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.
- [47] H. Kwak and J. An. Understanding news geography and major determinants of global news coverage of disasters. *Computation+Journalism Symposium*, 2014.
- [48] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and G. J. Jones. Automatic tagging and geotagging in video collections and communities. In *Proceedings of the 1st ACM international conference on multimedia retrieval*, page 51. ACM, 2011.
- [49] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM, 2009.
- [50] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *IEEE transactions on pattern analysis and machine intelligence*, 30(6):985–1002, 2008.
- [51] X. Li, C. G. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 11(7):1310–1322, 2009.
- [52] Y.-R. Lin, H. Sundaram, M. De Choudhury, and A. Kelliher. Discovering multirelational structure in social media streams. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 8(1):4, 2012.
- [53] H. Liu, Z. Wu, X. Li, D. Cai, and T. S. Huang. Constrained nonnegative matrix factorization for image representation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.
- [54] A. Lu, W. Wang, M. Bansal, K. Gimpel, and K. Livescu. Deep multilingual correlation for improved word embeddings. In *HLT-NAACL*, pages 250–256, 2015.
- [55] I. Mani and E. Bloedorn. Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1-2):35–67, 1999.

- [56] N. Martin and H. Maes. *Multivariate analysis*. Academic press, 1979.
- [57] A. Messina, M. Montagnuolo, R. Di Massa, and R. Borgotallo. Hyper media news: a fully automated platform for large scale analysis, production and distribution of multimodal news content. *Multimedia tools and applications*, 63(2):427–460, 2013.
- [58] R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. Association for Computational Linguistics, 2004.
- [59] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 331–340, 2009.
- [60] H. Nakasaki, M. Kawaba, T. Utsuro, and T. Fukuhara. Mining cross-lingual/cross-cultural differences in concerns and opinions in blogs. In *International Conference on Computer Processing of Oriental Languages*, pages 213–224. Springer, 2009.
- [61] T. Opsahl, F. Agneessens, and J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social networks*, 32(3):245–251, 2010.
- [62] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.
- [63] E. E. Papalexakis, N. D. Sidiropoulos, and R. Bro. From k-means to higher-way co-clustering: Multilinear decomposition with sparse latent factors. *IEEE transactions on signal processing*, 61(2):493–506, 2013.
- [64] S. Phuvipadawat and T. Murata. Breaking news detection and tracking in Twitter. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 3, pages 120–123. IEEE, 2010.
- [65] A. Popescu and I. Kanellos. Multilingual and content based access to FLICKR images. In *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on*, pages 1–5. IEEE, 2008.
- [66] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative multi-label video annotation. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 17–26. ACM, 2007.
- [67] A. Sadovnik, Y.-I. Chiu, N. Snavely, S. Edelman, and T. Chen. Image description with a goal: Building efficient discriminating expressions for images. In *Computer Vision*

- and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2791–2798. IEEE, 2012.
- [68] H. Sayyadi and L. Raschid. A graph analytical approach for topic detection. *ACM Transactions on Internet Technology (TOIT)*, 13(2):4, 2013.
- [69] H. Shan and A. Banerjee. Bayesian co-clustering. In *ICDM*, 2008.
- [70] S. Siersdorfer, J. San Pedro, and M. Sanderson. Automatic video tagging using content redundancy. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 395–402. ACM, 2009.
- [71] B. Sigurbjörnsson and R. Van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web*, pages 327–336. ACM, 2008.
- [72] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
- [73] P. Snickars and P. Vonderau. The YouTube Reader. 2010.
- [74] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, pages 207–218, 2014.
- [75] E. Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.
- [76] A. Sun, S. S. Bhowmick, and J.-A. Chong. Social image tag recommendation by concept matching. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1181–1184. ACM, 2011.
- [77] C. C. Tan, Y.-G. Jiang, and C.-W. Ngo. Towards textually describing complex video contents with audio-visual concept classifiers. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 655–658. ACM, 2011.
- [78] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [79] X. Wan, H. Jia, S. Huang, and J. Xiao. Summarizing the differences in multilingual news. *SIG IR*, 2011.

- [80] D. Wang, M. Ogihara, and T. Li. Summarizing the differences from microblogs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1147–1148. ACM, 2012.
- [81] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [82] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes. Unsupervised learning of acoustic features via deep canonical correlation analysis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4590–4594. IEEE, 2015.
- [83] J. Weston, S. Bengio, and N. Usunier. WSABIE: Scaling up to large vocabulary image annotation. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence (IJCAI)*, pages 2764–2770, 2011.
- [84] X. Wu, A. G. Hauptmann, and C.-W. Ngo. Measuring novelty and redundancy with multiple modalities in cross-lingual broadcast news. *Computer Vision and Image Understanding*, 110(3):418–431, 2008.
- [85] X. Wu, C.-W. Ngo, and Q. Li. Threading and autodocumenting news videos: a promising solution to rapidly browse news topics. *IEEE Signal Processing Magazine*, 23(2): 59–68, 2006.
- [86] W. Xiaoxuan, X. Lei, L. Mimi, M. Bin, E. S. Chng, and L. Haizhou. Broadcast news story segmentation using conditional random fields and multimodal features. *IEICE TRANSACTIONS on Information and Systems*, 95(5):1206–1215, 2012.
- [87] L. Xie, A. Natsev, X. He, J. Kender, et al. Visual memes in social media. In *ACM Multimedia*, 2011.
- [88] R. Xu, C.-Y. Tsai, and J. R. Kender. An adaptive anchor frame detection algorithm based on background detection for news video analysis. In *Audio, Language and Image Processing (ICALIP), 2016 International Conference on*, pages 743–748. IEEE, 2016.
- [89] F. Yan and K. Mikolajczyk. Deep correlation for matching images and text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

- [90] J. Yang and J. Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 587–596. ACM, 2013.
- [91] T. Yao, T. Mei, C.-W. Ngo, and S. Li. Annotation for free: Video tagging by mining user search behavior. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 977–986. ACM, 2013.
- [92] Q. Yu, J. Liu, H. Cheng, A. Divakaran, and H. Sawhney. Multimedia event recounting with concept based representation. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1073–1076. ACM, 2012.
- [93] W.-L. Zhao, X. Wu, and C.-W. Ngo. On the annotation of web videos by efficient near-duplicate search. *IEEE Transactions on Multimedia*, 12(5):448–461, 2010.
- [94] Q. Zhou, G. Xu, and Y. Zong. Web co-clustering of usage network using tensor decomposition. In *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on*, volume 3, pages 311–314. IEEE, 2009.