Structured Exploration of Who, What, When, and Where in Heterogeneous Multimedia News Sources

Brendan Jou*, Hongzhi Li*, Joseph G. Ellis*,
Daniel Morozoff-Abegauz and Shih-Fu Chang

Digital Video & Multimedia Lab, Columbia University

ABSTRACT

We present a fully automatic system from raw data gathering to navigation over heterogeneous news sources, including over 18k hours of broadcast video news, 3.58M online articles, and 430M public Twitter messages. Our system addresses the challenge of extracting "who," "what," "when," and "where" from a truly multimodal perspective, leveraging audiovisual information in broadcast news and those embedded in articles, as well as textual cues in both closed captions and raw document content in articles and social media. Performed over time, we are able to extract and study the trend of topics in the news and detect interesting peaks in news coverage over the life of the topic. We visualize these peaks in trending news topics using automatically extracted keywords and iconic images, and introduce a novel multimodal algorithm for naming speakers in the news. We also present several intuitive navigation interfaces for interacting with these complex topic structures over different news sources.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems; H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia

Keywords

Multimedia Analysis; Topic Linking; Speaker Diarization

1. INTRODUCTION

News today is arguably foundationally driven by just four of the five "W's" – *Who*, *What*, *When*, and *Where*. What happened? When did it happen? Where did it happen? Who was involved? Who said what? This natural progression of questions is a classic example of what one might ask about an event. Without any one of these questions, the story would fall flat and quickly become less captivating. For example, what is a story about a bombing without a mention of location and time? Each of the four W's are necessary to tell a full story.

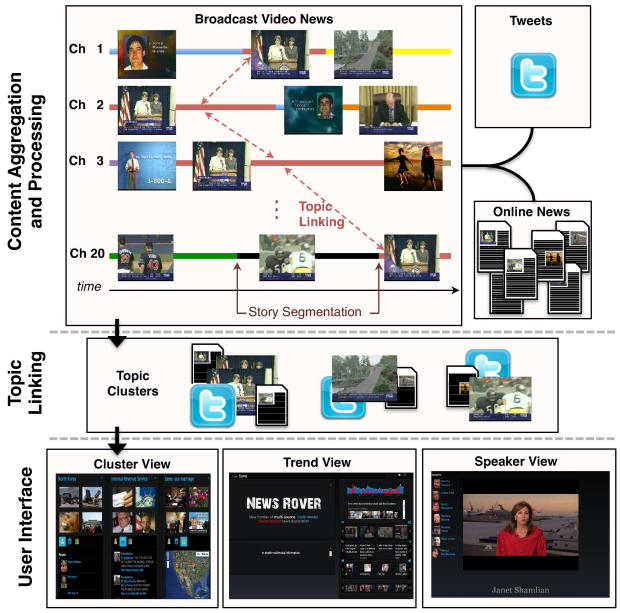

Figure 1: System Overview.

What about "Why" though – the fifth W? The goal of most news agencies and journalists is to bring as many facts as possible surrounding an event to the surface. As a result, *the why* tends to be relegated to opinion commentaries and post-event interview coverage. Further, the question of who, what, when and where an event happened precedes the question of why it happened. And so, our key motivating question was whether we could capture this principle quartet of W's, using each as a pillar toward a foundational platform to study and explore news in all its multimodal mediums.

Several works in academia and industry have been proposed to address similar challenges. Perhaps one of the most notable is the NIST TRECVID challenges, where from 2003 to 2006 there was a large focus on analyzing news productions and from 2007 to 2009 on more unconstrained news rushes. The goal in all these was to extract "name of the person/place/thing/event in the image/video" [12], that is, the four W's. However, the be-all-end-all in these experiments was exclusively video, and the breadth of topics is limited to those that get chosen for video broadcast. Instead, we expand the heterogeneity of our news sources in our system. Similarly, in the information extraction (IE)

*Denotes equal contribution.

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2508118.

community, a significant effort has been focused on detecting events and topics in newswire documents [11] and social media [2]. Another recent challenge called REPERE [10] is also focused on answering the questions, "Who is speaking? Who is present in the video? What names are cited? What names are displayed?" for European broadcast news. Our focus here will be not only on the persons or names, but also on the larger scope of the events and topics themselves, utilizing the full set of information in U.S. broadcast news which naturally differs programmatically from European news. We propose a unified framework where *both* news articles and social media posts, namely from Twitter, *and* broadcast video intersect, where events and entities are indexed regardless of media type by extracted topic structures. An overview of our system can be seen in Fig. 1

Our key contributions include (1) a recording architecture that archives and processes over 700 hours of video and 72k online articles per week, (2) a hierachically organized topic structure indexing who, what, when, and where, (3) a topic trend summarization method, (4) a novel approach to named multimodal speaker diarization and transcription, and (5) several novel user interfaces to navigate heterogeneous multimodal news.

2. SYSTEM ARCHITECTURE

2.1 Data Collection

Recording and Crawling. A major component of our system is recording and processing a variety of news content available through broadcast and cable news outlets. We use 12 cable tuners with on-chip encoding to simultaneously record. We currently have the capability to record news from about 100 U.S. English-speaking channels. To maximize our news recording coverage, we developed an automatic scheduling program that finds news programs using an online electronic program guide, and automatically queues them for recording by available tuners.

Along with our broadcast video recording capabilities, we also developed crawlers for Google News and Twitter feeds. Our Google News crawler searches every five minutes for new articles and indexes each document by the topic ontology defined by Google News. By retaining the topics from Google News and not just the articles, we are able to incorporate current trending topics as well as define a separate set of long-term topics we wish to track. In particular, there are 2,000 popular long-term news topics, such as the "Syrian Revolution" and "North Korea," that we always search for regardless of if they make the front page of Google News. Our Twitter crawler takes a similar approach, finding and retaining Twitter trends and popular hashtags. We index these Twitter trends and their associated "tweets", while also maintaining a list of 2,000 long-term popular topics on Twitter. Weekly recording statistics are in Table 1.

Table 1: Average Estimates of Data Per Week.

Programs recorded	700	Online articles	72,000
Hours of video	700	Google topics	4,000
Stories segmented	7,000	Twitter trends	3,500

Story Segmentation. The news programs we record are often 30 minutes to one-hour long and contain news commentary or reporting on a variety of different stories as well as commercials. Our news recording and processing system automatically finds the boundaries between these story segments and cuts up the original program into individual story segments that can be linked to topics and more easily consumed by an end user.

Closed captions (CC) are required by law to be made available for all U.S. news programs and many CC transcripts include special markers that denote the end of a story segment, >>>, and speaker changes, >>. Closed captions often lag the actual audiovisual content of the program by 10-15 seconds. And so, to acquire an accurate boundary detection for story segments we perform Automatic Speech Recognition (ASR) on the entire news program and then temporally align the CC transcript with the output of ASR using dynamic time warping. We further refine the boundary precision by aligning the story marker to the closest detected shot boundary. In cases where special markers do not exist in CC, we will apply our previous work using multimodal segmentation [7] that proved a F1 score performance of 0.76.

2.2 Topic Structures

News media is hierarchical in nature and generally characterized by a group of documents, often from different sources and media types, surrounding a single event. A news topic can be these single, one-off events, but can also be a concept or a collection of events. For example, one might imagine the 2012 Olympics as a topic of news-worthy discussion, but the Olympics itself, as a collection of events and concept, could be a topic as well. Topics form the "what" of news.

To discover and mine topics in the news, many previous works have employed Latent Dirichlet Allocation (LDA) and its variants. However, today, there are also a number of commercial platforms that apply proprietary topic discovery algorithms using implicit, but strong signals like user traffic. Google News extracts trending topics from the online articles largely using search logs and click-through traffic [4]. Twitter mines hot trends from tweets by potentially leveraging content sharing (i.e. re-tweets) and publish-throughput via hashtags. Our system collects trending topics from these sources, including Google News, Google Search Statistics and Twitter, and combines them into a generic pool. We keep an active list of these trending topics, continuously assigning incoming tweets, articles, and video to topics.

Given our gathered topics and their associated articles and tweets, we then need to link in the video news stories. We use a multi-modal topic linking method to assign each video news story to a topic. As shown in Fig. 2, we extract keyframes from each video and perform near-duplicate image matching to topic-organized images crawled from the online news articles and tweets. We also use the CC transcript from the story segments for text matching using a term-frequency inverse-document-frequency (TF-IDF) model.

Entity Extraction. From each story segment's CC transcript we perform named entity extraction [6], including persons, locations, and organizations. This naturally gives us the "who" and "where" that are associated with stories and topics. As our CC transcript is time-aligned with ASR, accurate time stamps are known for each word, and therefore we know when each entity is mentioned in each story segment. Since CC names are not always accurate, and often times people are referred to by only their first name, we use community-edited databases like DBpedia and Wikipedia to perform name co-referencing over all news stories and topics.

Topic Trend Visualization & Representation. Hav-

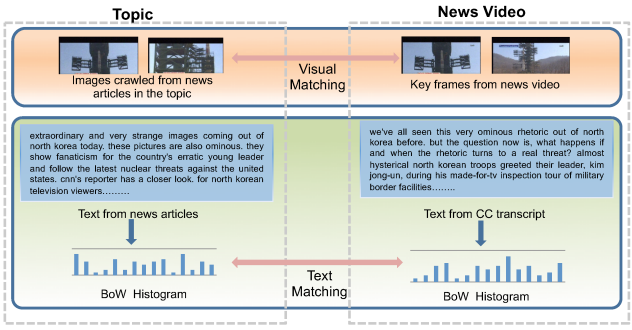

Figure 2: Topic Matching Process.

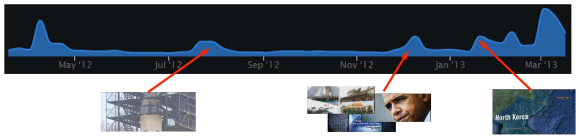

Figure 3: Example Trend for "North Korea" Topic.

ing collected data for the "long-term" topics over the past three months, we are able to track the news coverage a single topic receives over time. A "traffic" curve for a topic is visualized as in Fig. 3, where the y-axis represents the volume of news content we have linked to this topic versus publish time. In the trend, we observe visible peaks for times when there is heightened news coverage on this topic.

The question then is how we can quickly understand what happened "when," for each of these peaks or (sub-)events. We approach this summarization problem in two ways. In one approach, we extracted keywords and phrases from articles appearing often within each peak [3] and simply used the most commonly occurring words or phrases. However, we noted that certain pictures or scenes in the news are often repeated by multiple news programs during important events. These iconic images are highly descriptive, conveying information in a succinct and complete way that text cannot. To quickly detect these reoccurring images we have implemented a near-duplicate detection algorithm to find these reoccurring scenes or images. Example results of extracted representative images are shown in Fig. 3.

3. WHO SAID WHAT

"Talking heads" are a critical facet of news reporting. As we have addressed the challenge of extracting the four W's, deeper and interesting structures can be explored in the relationship between two or more of these W's. We begin to study these relationships by tackling the combination of *who* and *what* in broadcast news. In particular, we seek to address the question of "who said what," that is, naming speaking persons and determining their respective quotes.

The problem of speaker diarization is the closest to our setting. Speaker diarization seeks to answer the question of "who spoke when," often by clustering detected speech and mapping clusters to names [1]. Recently, [13] explored multimodal speaker diarization using a Dynamic Bayesian Network in both the business meeting and broadcast news videos. Several works extending from [5], have tried to tackle a similar problem using multimodal information for television shows but rely on the a priori presence of fully annotated transcripts that have names mapped to spoken text.

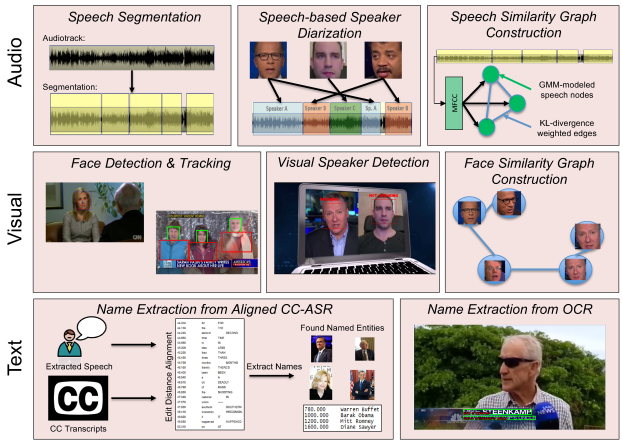

Figure 4: "Who Said What" Features.

Such transcripts are not available in broadcast news and make it far more challenging without this supervision.

Our approach differs from these previous approaches in one fundamental way: we have a more stringent definition of "who." For a useful system, we require that "who" does not simply mean a class or cluster number but rather an actual name, and to do so without manual intervention. We also exploit the multimodal aspect of the "who said what" problem by using audio, visual, and textual cues.

3.1 Speaker Representation

Visual/Facial. We perform face detection on the news stories using OpenCV's Haar-like cascade and then extract SIFT features from detected landmark points within each detected face [15]. We generate face tracks by exploiting temporal and spatial continuity within the detected faces. Finally, we compute the similarity between face tracks using a Gaussian Kernel distance between every pair of faces in the tracks, and then average the distance in the top face pairs.

We also perform text detection on sampled frames from the stories to extract on-screen person names. These names from optical character recognition (OCR), along with those from CC, are combined and normalized to form the set of candidate name labels during the prediction stage.

Audio/Speech. As our basic visual unit is the *face track*, our basic audio unit is a *speech segment*. Speech segments denote contiguous speech by one or more persons without extended silence. We extract segments by applying the segmentation tool in [8]. On average, detected speech segments are 2.3 seconds long. To measure the similarity between segments, we extract MFCCs from each audio segment and model each as a multivariate Gaussian. The distance between speech segments is measured by the symmetric Kullback-Leibler (KL) divergence over the multivariate Gaussian distributions [14], and then a Gaussian kernel over the KL-distances normalized by the standard deviation of all the distances in a story segment.

Multimodal. In order to link the audio and visual features, we use a variant of the visual speaker detection in [5]. This knowledge of who is visually speaking allows us to disambiguate from whom speech is coming from when there is more than one person on-screen. Instead of a pure visual speaker detection, we take a hybrid multimodal approach to detecting visual speakers. Using the facial land-

marks, we affine-align the face, determine the mouth region and perform template matching to detect whether the face is speaking. Repeating this over the entire face track, we get a series of best template matches which correspond to the smallest sum of square differences for "inter-frames" or frame pairs. Following [5], two thresholds are used to predict if the "inter-frame" is non-speaking, reject and speaking. We use majority voting within face tracks that overlap speech segments to predict if the face is visually speaking.

3.2 Named Multimodal Speaker Diarization

Given our multimodal speaker identity cues, we position our problem in a transductive learning setting and use label inference over a heterogeneous graph with weak labels, which correspond to the names automatically extracted from the news video. We create a visual subgraph consisting of face tracks and an audio subgraph consisting of speech segments, both constructed using b-matching [9]. Cross-modal edges are formed between vertices of the two subgraphs if a face track temporally overlaps a speech segment and is detected as speaking via our multimodal speaker detection algorithm.

We apply the extracted names from CC and OCR as weak labels on our constructed graph. Two approaches to weak label assignment have proven effective in our experiments. First, if a face track temporally overlaps with an occurrence of an OCR name on-screen we assign the name to that face node. Second, if a new face track appears on screen up to 10 seconds after a CC name appears in the transcript, we assign the CC name to this face node.

These weak labels are then propagated on the graph using local and global consistency [9], enforcing smoothness using the normalized Laplacian and softly constraining the solution to the labels since they are weak. We set the highest scoring label and its corresponding name as the prediction for each node in the graph. Given the predicted names for speakers, getting the "what," or quote, related to them is now trivial because we have their associated speech segments and simply extract the portion of closed captions that is time-locked with the speech.

To evaluate our performance, we collected annotations using Amazon's Mechanical Turk over 225 detected face tracks from NBC Nightly News and News 4 New York over an average of 5 to 7 unique names per story. We limited our face tracks to a subset of mostly frontal tracks by performing a second pass of face detection using a higher threshold. When considering all identities, including those who never spoke, we correctly labeled 105 face tracks for a total accuracy of 0.475, using the extracted CC and OCR names as weak labels. This represents a significant improvement over simply using CC or OCR names as weak labels alone, which give accuracies of 0.284 and 0.40, respectively. Additionally, as our goal is to name speakers, if we limit to speaking non-anchors appearing on-screen, we achieve an accuracy of 0.619.

4. NAVIGATION INTERFACE

We have developed three novel interfaces that showcase the hierarchy of our extracted topics from the story linking capability to individual topics, and who-said-what speaker identification. First, our "topic cluster" interface shows a user the major news topics or stories of the day as well as a quick summary of each topic via "who," "what," and "where." Second, our "topic exploration" interface allows exploration of news videos, articles, tweets, and people involved in a particular topic. The content shuffles and rearranges dynamically based on user clicks over the trend indicating the "when" of a topic, and we present our visual summarization of the topic as well. Last, our "who said what" interface showcases a video and dynamically changing name display, including biographical information like a Wikipedia page and image. Screenshots of these interfaces are in Fig 1.

5. CONCLUSION

We have presented a unique and novel system integrating heterogeneous multimodal news sources including broadcast news, online articles, and Twitter tweets. Extracting topics over time we are able to generate trends, observe peaks in each topic, and summarize each (sub-)event per trend by keywords and iconic images. In addition, we presented a novel multimodal speaker diarization framework using label propagation that is able to directly infer names in broadcast video automatically without manual intervention for class-name assignment. We also showed several intuitive interfaces that streamline the navigation of these modalities and gives a fresh look at news past and present. In the future, we plan to pursue a user study to evaluate the utility of our system for understanding and structuring the news.

6. REFERENCES

[1] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *ASLP*, 2012.

[2] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on Twitter. In *ICWSM*, 2011.

[3] E. L. Bird, Steven and E. Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[4] A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google News personalization: Scalable online collaborative filtering. In *WWW*, 2007.

[5] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automated naming of characters in TV video. *Image Vision Computing*, 2009.

[6] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, 2005.

[7] W. Hsu, L. Kennedy, C.-W. Huang, S.-F. Chang, C.-Y. Lin, and G. Iyengar. News video story segmentation using fusion of multi-level multi-modal features in TRECVID 2003. In *ICASSP*, 2004.

[8] M. Huijbregts. *Segmentation, Diarization, and Speech Transcription: Suprise Data Unraveled*. PhD thesis, University of Twente, 2008.

[9] T. Jebara, J. Wang, and S.-F. Chang. Graph construction and b-matching for semi-supervised learning. In *ICML*, 2009.

[10] J. Kahn, O. Galibert, L. Quintard, M. Carre, A. Giraudel, and P. Joly. A presentation of the REPERE challenge. In *CBMI*, 2012.

[11] Q. Li, S. Anzaroot, W.-P. Lin, X. Li, and H. Ji. Joint inference for cross-document information extraction. In *CIKM*, 2011.

[12] National Institute of Standards and Technology. Text REtrieval Conference (TREC): VIDeo track (TRECVID).

[13] A. Noulas, G. Englebienne, and B. J. A. Kröse. Multimodal speaker diarization. *PAMI*, 2012.

[14] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern. Automatic segmentation, classification and clustering of broadcast news audio. In *DARPA Speech Recogn. Workshop*, 1997.

[15] M. Uřičář, V. Franc, and V. Hlaváč. Detector of facial landmarks learned by the structured output SVM. In *VISAPP*. SciTePress, 2012.