Identifying News Broadcasters’ Ideological Perspectives Using a Large-Scale Video Ontology

Wei-Hao Lin and Alexander Hauptmann

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 U.S.A.
{whlin,alex+}@cs.cmu.edu

Abstract

Television news has been the predominant way of understanding the world around us, but individual news broadcasters can frame or mislead audience’s understanding about political and social issues. We aim to develop a computer system that can automatically identify highly biased television news, which may prompt audience to seek news stories from contrasting viewpoints. But can computers determine if news videos were produced by broadcasters holding differing ideological beliefs? We developed a method of identifying differing ideological perspectives based on a large-scale visual concept ontology, and the experimental results were promising.

1. Introduction

Television news has been the predominant way of understanding the world around us. Individual news broadcasters, however, can frame, even mislead, audience’s understanding about political and social issues. A recent study shows that people’s main news sources are highly correlated with their misconceptions about the Iraq War (Kull, 2003). 80% of the respondents whose primary news source is FOX have one or more misconceptions, while among people whose primary source is CNN, 50% have misconceptions.

The difference in framing news events is clearer when we compare news broadcasters across national and language boundaries. For example, Figure 1 shows how an American broadcaster (NBC) and an Arabic broadcaster (LBC) portray Yasser Arafat’s death in 2004. The two broadcasters’ footage looks very different: NBC shows stock footage of Arafat, while LBC shows the actual funeral and interviews with general public.

We consider a broadcaster’s bias in portraying a news event “ideological.” We take the definition of ideology as “a set of general beliefs socially shared by a group of people” (van Dijk, 1998). Television news production involves a large number of people who share similar social and professional beliefs. A news broadcaster may consistently exhibit bias in reporting political and social issues partly because producers, editors, and reporters collectively make similar decisions (e.g., what to cover, who to interview, and what to show on a screen) based on shared value judgments and beliefs.

We aim to develop a computer system that can automatically identify highly biased television news. Such system may increase audience’s awareness about individual news broadcasters’ bias and prompt them to seek news stories from contrasting viewpoints. However, can computer automatically understand differing ideological perspectives expressed in television news footage?

• In this paper we proposed a method of identifying differing ideological perspectives in news video based on the imagery chosen to show on the screen. We motivated our method based on visual concepts in Section 2. We described how to represent a video in terms of visual concepts (e.g., outdoor, car, and people walking) in Section 3.1., and then how to quantify the similarity between two news video footage in terms of visual concepts in Section 3.2..

• We evaluated the proposed method on a large broadcast news video archive (Section 4.1.). To determine if two videos portray the same news event from differing ideological perspectives, we trained a classifier to make a binary decision (i.e., same perspective or different perspectives). The classifier was shown to achieve high accuracy in Section 4.3.. We applied the same idea to determine if two videos covered the same news event in Section 4.2..

• So far we conducted the experiments using manual concept annotation to avoid concept classifiers’ poor performance being a confounding factor. In Section 4.4. we repeated the above experiments and replaced manual annotations with empirically trained concept classifiers.

2. Motivation

We were inspired by the recent work on developing large-scale concept ontology for video retrieval (Hauptmann, 2004), and considered a specific kind of visual grammar that may exhibit ideological perspective: composition (Efron, 1972). Here visual concepts are generic objects, scenes, and activities (e.g., outdoor, car, and people walking). Visual concepts can represent a video’s visual content more closely than low-level features (e.g., color, texture, and shape) can. Many researchers have actively developed concept classifiers to automatically detect concepts’ presence in video. A concept classifier reads an image and outputs the likelihood that a visual concept is present on the screen. Therefore, if computers can automatically identify the visual concepts, computers may be able to learn the
difference between broadcasters holding differing ideological perspectives based on what are chosen to show in news footage.

We illustrate the idea in Figure 2. We counted the visual concepts in the television news footage about the Iraq War from two different broadcasters (an American broadcaster CNN vs. an Arabic broadcaster LBC), and displayed them in text clouds (see Section 4.1. for more details about the data). Due to the nature of broadcast news, it is not surprising to see many people-related visual concepts (e.g., “Adult”, “Face”, and “Person”). Because the news stories are about the Iraq War, it is also not surprising to see many war-related concepts (e.g., “Weapons”, “Military Personnel”, and “Daytime Outdoor”). The surprising differences, however, lie in the subtle emphasis on some concepts. “Weapons” and “Machine Guns” are shown more often in CNN (relative to other visual concepts in CNN) than in LBC. On the contrary, “Civilian Person” and “Crowd” are shown more often in LBC than in CNN. How frequently some visual concepts are chosen seems to reflect a broadcaster’s ideological perspective on a particular news event.

3. Measuring Semantic Similarity in Visual Content

To develop a computer program that can identify videos conveying differing ideological perspectives on a news event, we need to address the following two questions:

1. Can computers determine if two television news stories are about the same news event?
2. Given two television news stories on the same news event, can computers determine if they portray the event from differing ideological perspectives?

We could identify news stories’ topic using textual clues (e.g., words in automatic speech recognition transcripts), but here we attack a more challenging question: grouping television news stories on the same event using only visual clues. More and more videos are produced and consumed by users on the Internet. Contrary to news videos, web videos do not usually come with clear voice-over that describes what a video is about. An imagery-based topic tracking approach is more likely to be applicable for web videos than a text-based approach.

The two research questions can be boiled down to the same question:

How well can we measure the similarity in visual content between two television news videos?

News videos on the same news event are likely to have similar visual content, while news videos on different news events are less likely to have similar visual content. Similarly, given two news videos on the same news event, broadcasters holding similar ideological beliefs are likely to portray the new event in a similar manner, while news broadcasters holding different ideological views are less likely to display similar visual content. Therefore, the key research question becomes measuring the “semantic” similarity in visual content.

3.1. Representing Video As Visual Concepts

We proposed a method of measuring semantic similarity between two news stories using a large-scale visual concept...
ontology. Our method consists of four steps, as illustrated in Figure 3. In Step 1 we first run a shot detector to detect shot boundaries in a news story, and select the middle frame of a shot as its key frame. In Step 2 we check if any concepts in a visual concept ontology are present in the key frames. A concept’s presence can be manually labeled by human annotators, but can be also automatically but less accurately labeled using machine learning classifiers. An example key frame and its visual concepts are shown in Figure 4.

![Figure 3: Our method of measuring similarity in visual content consisted of four steps. Step 1: extract videos’ key frames. Step 2: determine what visual concepts are present in key frames. Step 3: model the occurrences of visual concepts using a multinomial distribution. Step 4: measure “distance” between two multinomial distributions using Kullback-Leibler divergence.](image)

Figure 3: Our method of measuring similarity in visual content consisted of four steps. Step 1: extract videos’ key frames. Step 2: determine what visual concepts are present in key frames. Step 3: model the occurrences of visual concepts using a multinomial distribution. Step 4: measure “distance” between two multinomial distributions using Kullback-Leibler divergence.

In Step 3 we model the occurrences of visual concepts in key frames using a statistical distribution. A natural choice is all visual concepts. The value of KL divergence is asymmetric, and we take the average of $D(P||Q)$ and $D(Q||P)$ as the (symmetric) distance between $P$ and $Q$.

We measure the similarity between two videos’ multinomial distributions in terms of Kullback-Leibler (KL) divergence (Cover and Thomas, 1991). KL divergence is commonly used to measure the “distance” between two statistical distributions. The KL divergence between two multinomial distributions $P$ and $Q$ is defined as follows:

$$D(P||Q) = \sum_c P(c) \log \frac{P(c)}{Q(c)},$$

where $c$ is all visual concepts. The value of KL divergence quantifies the similarity between two news videos in terms of visual concepts chosen by individual broadcasters. The smaller the value of KL divergence, the more similar two news videos. KL divergence is asymmetric, and we take the average of $D(P||Q)$ and $D(Q||P)$ as the (symmetric) distance between $P$ and $Q$.

---

Table 1: The major categories and sample LSCOM concepts in each category.

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program</td>
<td>advertisement, baseball, weather news</td>
</tr>
<tr>
<td>Scene</td>
<td>indoors, outdoors, road, mountain</td>
</tr>
<tr>
<td>People</td>
<td>NBA players, officer, Pope</td>
</tr>
<tr>
<td>Objects</td>
<td>rabbit, car, airplane, bus, boat</td>
</tr>
<tr>
<td>Activities</td>
<td>walking, women dancing, cheering</td>
</tr>
<tr>
<td>Events</td>
<td>crash, explosion, gun shot</td>
</tr>
<tr>
<td>Graphics</td>
<td>weather map, NBA scores, schedule</td>
</tr>
</tbody>
</table>

---

1The complete list of visual concepts is available at [http://www.lscom.org/concept.htm](http://www.lscom.org/concept.htm)

---

In this paper we chose the Large-Scale Concept Ontology for Multimedia (LSCOM) (Kennedy and Hauptmann, 2006) to represent television video’s visual content. LSCOM, initially developed for improving video retrieval, contains hundreds of generic activities, objects, and scenes. LSCOM started from more than ten thousands of concepts collected from various sources such as TGM, Time Life, TV Anytime, Comstock, and WordNet. Later around one thousand concepts were chosen based on video retrieval utility, machine-learning feasibility, and observability. The LSCOM taxonomy was also mapped to Cyc to suggest new concepts. The major categories and example concepts in each category are listed in Table 3.1.
4. Experiments

4.1. Data

We evaluated the proposed method of identifying differing ideological perspectives on a broadcast news video archive from the 2005 TREC Video Evaluation (TRECVID) (Over et al., 2005). The TRECVID 2005 video archive consisted of television news videos recorded in late 2004. The news programs came from multiple news broadcasters in three languages: Arabic, Chinese, and English, as shown in Table 2.

<table>
<thead>
<tr>
<th>Language</th>
<th>Hours</th>
<th>News Broadcasters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>33</td>
<td>LBC</td>
</tr>
<tr>
<td>Chinese</td>
<td>52</td>
<td>CCTV, NTDTV</td>
</tr>
<tr>
<td>English</td>
<td>73</td>
<td>CNN, NBC, MSNBC</td>
</tr>
</tbody>
</table>

Table 2: The news broadcasters and the total length of news videos in each language in the TRECVID’05 video archive.

We used the official shot boundaries that the TRECVID organizer, NIST, provided for the TRECVID 2005 participants. We ran an in-house story segmentation program to detect news story boundaries (Hauptmann et al., 2005), resulting in 4436 news stories. The story segmentation program detected a news story’s boundary using cues such as an anchor’s presence, commercials, color coherence, and average story length. We removed anchor and commercial shots because they contained mostly talking heads and conveyed little ideological perspective.

We collected ten news events in late 2004 and news videos covering these news events. We made sure the news events in Table 3 were covered by broadcasters in more than one language. A news story covered a news event if a news event’s keywords were mentioned in the video’s English automatic speech recognition (ASR) transcripts. NIST provided English translation for non-English news programs. Note that ASR transcripts were used only for linking stories on the same news event. LSCOM annotators did not use ASR transcripts and made judgments solely based on visual content.

<table>
<thead>
<tr>
<th>News Event</th>
<th>Stories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iraq War</td>
<td>231</td>
</tr>
<tr>
<td>United States presidential election</td>
<td>114</td>
</tr>
<tr>
<td>Arafat’s health</td>
<td>308</td>
</tr>
<tr>
<td>Ukrainian presidential election</td>
<td>11</td>
</tr>
<tr>
<td>AIDS</td>
<td>21</td>
</tr>
<tr>
<td>Afghanistan situation</td>
<td>42</td>
</tr>
<tr>
<td>Tel Aviv suicide bomb</td>
<td>2</td>
</tr>
<tr>
<td>Powell’s resignation</td>
<td>45</td>
</tr>
<tr>
<td>Iranian nuclear weapon</td>
<td>46</td>
</tr>
<tr>
<td>North Korea nuclear issue</td>
<td>51</td>
</tr>
</tbody>
</table>

Table 3: The number of television news stories on the ten news events in late 2004

We used visual concepts annotation from the Large-Scale Concept Ontology for Multimedia (LSCOM) v1.0 (Kennedy and Hauptmann, 2006). The LSCOM annotations consisted of the presence of each of the 449 LSCOM visual concepts in every video shot of the TRECVID 2005 videos. There are a total of 689064 annotations for the 61901 shots, and the median number of annotations per shot is 10.

We conducted the experiments first using the LSCOM annotations, and later replaced manual annotations with predictions from empirically trained concept classifiers. Using manual annotations is equal to using very accurate concept classifiers. Given the state-of-the-art classifiers for most visual concepts are far from perfect, why would we start from assuming perfect concept classifiers? It is because manual annotations allow us to test the idea of measuring similarity in visual concept using concepts without being confounded by the poor accuracy of the concept classifiers.

4.2. Identifying News Videos on the Same News Event

Because we are interested in how the same news event is portrayed by different broadcasters, we need to find the television news stories on the same news event in a video archive. As we argued in Section 3, this task boils down to comparing similarity between two videos’ visual content. News videos on the same news event are likely to show similar visual content. Given two news videos, we could measure their similarity in terms of visual concepts as proposed in Section 3.

We developed a classification task to evaluate the proposed method of identifying news videos on the same event. Each time the classifier is presented with a pair of television news videos, and is asked to make a binary decision between two categories: Different News Events (DNE) vs. Same News Event (SNE). DNE contains news video pairs that are from the same broadcaster but on different news events (e.g., two videos from CNN: one is about the “Iraq War” and the other is about “Powell’s resignation”). SNE contains news video pairs from the same broadcaster and on the same news event (e.g., two videos from CCTV about the same event “Tel Aviv bomb”). The predictor for the classification task is the value of KL divergence between two videos. Our method is effective if such classifier achieves high accuracy.

Among all possible video pairs that satisfy the conditions of Different News Event (DNE) and Same News Event (SNE), we randomly sampled 1000 video pairs for each category. We looked up their LSCOM concept annotations (Section 3.1.), estimated multinomial distributions’ parameters, and trained classifiers based on the values of (symmetric) KL divergence (see Section 3.2.). We varied the training data from 10% to 90%, and reported the accuracy on the held-out 10% of video pairs. Accuracy is defined as the number of video pairs that are correctly classified divided by the total number of video pairs in the held-out set. Because there were an equivalent number of video pairs in each category, a random guessing baseline would have 50% accuracy. We repeated the experiments 100 times by sampling different video pairs, and reported the average accuracy. The choice of classifier did not change the results much, and we reported only the results using Linear Discriminant Analysis and omitted the results using Support Vector Machines.

The experimental results in Figure 5 showed that our method based on visual concepts can effectively tell news videos on the same news event from news videos on dif-
different news events. The classification accuracy was significantly better than the random baseline (t-test, $p < 0.01$), and reached a plateau around 70%. Our concept-based method of identifying television news stories on the same event could thus well complement other methods based on text (Allan, 2002; Zhang et al., 2004), color (Zhai and Shah, 2005), and near-duplicates images (Wu et al., 2007). Although LSCOM was initially developed for supporting video retrieval, the results also suggested that LSCOM contained large and rich enough concepts to differentiate news videos on a variety of news events.

4.3. Identifying News Videos of Differing Ideological Perspectives

Given two news videos on the same news event, how can computers tell if they portray the event from different ideological perspectives? As we hypothesized in Section 2., given a news event, broadcasters holding similar ideological beliefs (i.e., the same broadcaster) are likely to choose similar visual concepts to compose news footage, while broadcasters holding different ideological beliefs (i.e., different broadcasters) are likely to choose different visual concepts. The task of identifying if two news videos convey differing ideological perspectives boils down to measuring if two videos are similar in terms of visual concepts (Section 3).

We developed a classification task to evaluate the proposed method of identifying news videos from differing ideological perspectives. There were two categories in the classification task: Different Ideological Perspectives (DIP) vs. Same Ideological Perspectives (SIP). DIP contains news video pairs that are about the same news event and from different broadcasters (e.g., two videos about “Arafat’s death”: one from LBC and one from NBC). SIP contains news video pairs that are about the same event but from the same broadcaster (e.g., two videos both from NTDTV and about “Powell’s resignation”). We trained a binary classifier to predict if a news video pair belongs to DIP or SIP. We followed the classification training and testing procedure in Section 4.2.

The experimental results in Figure 6 showed that our method based on visual concepts can effectively tell news videos produced by broadcasters holding similar ideological beliefs from those holding differing ideological beliefs.

The classification accuracy was significantly better than the random baseline (t-test, $p < 0.01$), and reached a plateau around 72%. Given two news videos are on the same news event, we can then use the propose method to test if they portray the news from differing ideological perspectives. Because we already knew a video’s broadcaster when the video was recorded, wasn’t the task of identifying if two news videos portray the news event from differing ideological perspectives as trivial as checking if they come from different broadcasters? Although we can accomplish the same task using metadata such as a news video’s broadcaster, this method is unlikely to be applicable to videos that contain little metadata (e.g., web videos on YouTube). We opted for a method of broader generalization, and developed our method solely based on visual content and generic visual concepts.

4.4. Concept Classifiers’ Accuracy

So far our experiments were based on manual annotations of visual concepts from LSCOM. Using manual annotation is equal to assuming that perfect concept classifiers are available, which is unrealistic given that the state-of-the-art classifiers are far from perfect for most visual concepts (Naphade and Smith, 2004). So how well can computers determine if two news videos convey a differently ideological perspective on a news event using empirically trained classifiers?

We obtained 449 LSCOM concept classifiers’ empirical accuracy by training Support Vector Machines on 90% of positive examples and testing on the held-out 10%. We first trained uni-modal concept classifiers using single low-level features (e.g., color histogram in various grid sizes and color spaces, texture, text, audio, etc), and built multi-modal classifiers that fused the outputs from best uni-modal classifiers (see (Hauptmann et al., 2005) for more details about the training procedure). We evaluated the performance of the best multi-modal classifiers on the held-out set in terms of average precisions (AP).

We varied concept classifiers’ accuracy by injecting noise into manual annotations. AP is a rank-based evaluation metric, but our experiments relied on set-based metrics. We thus approximated AP using recall-precision break-even points, which was highly correlated with AP (Manning et
al., 2008). We randomly flipped the positive and negative labels of visual concepts until we reached the desired break-even points. We varied the classifiers’ break-even points from APs obtained from empirically trained classifiers to 1.0 (i.e., perfect accuracy), and repeated the experiments in Section 4.2. and Section 4.2.

![Graph](image)

(a) Identifying news video pairs covering similar news events
(b) Identifying news video pairs of different ideological perspectives

Figure 7: We varied the classifiers’ accuracy and repeated the two experiments in Figure 5 and Figure 6. The x axis is the (simulated) classifiers’ accuracy in terms of precision-recall break-even points. The leftmost data point was based on the performance of the empirically trained classifiers. The y axis is the classification accuracy.

The experimental results showed that the empirically trained classifiers cannot satisfactorily identify news videos covering the same news event (Figure 7a) and news videos conveying differing perspectives (Figure 7b). Although the classification accuracy using empirically trained concept classifiers (i.e., the leftmost data point) was statistically significantly from random (t-test, p < 0.01), the difference was not practically significant. The median AP of the empirically trained classifiers was 0.0113 (i.e., the x coordinate of the leftmost data point in Figure 7). It was not surprising to see the classification accuracy improved as concept classifiers’ break-even points increased. To achieve reasonable performance we seemed to need concept classifiers of break-even points 0.6.

We should not be easily discouraged by current classifiers’ poor performance. With the advance of computation power and statistical learning algorithms, it is likely that concept classifiers’ accuracy will be continuously improved. Moreover, we may be able to compensate for poor accuracy by enlarging the number of concepts, as demonstrated recently in the study of improving video retrieval using more than three thousand of visual concepts (Hauptmann et al., 2005).

5. Conclusions

We proposed a method of measuring difference in visual content using a large-scale video concept ontology. The experiment results showed that by representing news footage in terms of visual concepts, we could start to learn news broadcasters’ patterns in composing news videos about different news topics and in portraying a news event from different ideological perspectives.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable suggestions for improving this paper. This research was supported in part by the National Science Foundation (NSF) under Grant No. IIS-0205219.

6. References


