TOPIC TRACKING ACROSS BROADCAST NEWS VIDEOS WITH VISUAL DUPLICATES AND SEMANTIC CONCEPTS

Winston H. Hsu and Shih-Fu Chang

Department of Electrical Engineering, Columbia University, New York
{winston, sfchang}@ee.columbia.edu

ABSTRACT

Videos from distributed sources (e.g., broadcasts, podcasts, blogs, etc.) have grown exponentially. Topic threading is very useful for organizing such large-volume information sources. Current solutions primarily rely on text features only but encounter difficulty when text is noisy or unavailable. In this paper, we propose new representations and similarity measures for news videos based on low-level features, visual near-duplicates, and high-level semantic concepts automatically detected from videos. We develop a multimodal fusion framework for estimating relevance of a new story to a known topic. Our extensive experiments using TRECVID 2005 data set (171 hours, 6 channels, 3 languages) confirm that near-duplicates consistently and significantly boost the tracking performance by up to 25%. In addition, we present information-theoretic analysis to assess the complexity of each semantic topic and determine the best subset of concepts for tracking each topic.

1. INTRODUCTION

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

Current topic threading approaches primarily exploit text information from speech transcripts, closed captions, or web documents. The use of multimodal information such as visual duplicates [1] or semantic visual concepts [2], though not explored before, are very helpful. There are usually recurrent visual patterns in video stories across sources and can help topic threading. For example, Fig. 1 has three example stories[1] from Chinese, Arabic, and English news sources, which cover the same topic. The stories from different channels share a few near-duplicates such as those showing Bush and Blair, press conference location, and audiences. Such duplicates, confirmed by our analysis later, are actually effective for news threading across languages (cf. Section 2.1.3).

A major research group for topic threading based on text has been conducted under NIST Topic Detection and Tracking (TDT) event [3], which includes three tasks: (1) story link detection, determining whether two stories discuss the same topic; (2) topic tracking, associating incoming stories with topics that are known to the system; (3) topic detection, detecting and tracking topics that are not previously known to the system. In this paper, we mainly focus on topic tracking across international broadcast news videos. One representative work of text-based approach can be found in [4], where authors represent documents as vectors of words, weighted by term-frequency inverse-document-frequency (TF-IDF). The cosine angle is used for measuring document-pair similarity. A modified k nearest neighbor (kNN) approach is then used for classification.

Recently some work started to study new techniques using multimodal information for story topic tracking. Xie *et al*. [5] applied Hierarchical-HMM models over the low-level audio-visual features to discover spatio-temporal patterns, latent semantic analysis to find text clusters, and then fused these multimodal tokens to discover potential story topics. In [6], authors studied the correlation between manually annotated visual concepts (e.g., sites, people, and objects) and topic annotations, and used graph cut techniques in story clustering. In [7], authors addressed the problem of linking news stories across two English channels on the same day, using global affine matching of key-frames as visual similarity. In all of these prior works, neither visual duplicates nor automatically detected visual concepts were used. In addition, comparisons with text-based approaches were not clear.

In this work, we develop novel approaches for story topic tracking using multimodal information, including text, visual duplicates, and semantic visual concepts. We propose a general fusion framework for combining diverse cues and analyze the performance impact by each component. Evaluating on TRECVID 2005 data set [2], fusion of visual duplicates improves the state-of-the-art text-based approach consistently by up to 25%. For certain topics, visual duplicate alone even outperforms the text-based approach. In addition, we propose an information-theoretic method for selecting subsets of semantic visual concepts that are most relevant to topic tracking.

We describe the new multimodal topic tracking framework and story representations in Section 2. Similarity measures and fusion schemes are also discussed. In Section 3, evaluations of the proposed techniques are shown on the TRECVID 2005 benchmark. We present conclusions and future work in Section 4.

2. MULTIMODAL TOPIC TRACKING

Underlying any topic tracking method, two fundamental issues need to be addressed – (1) representation of each story and (2) measurement of similarity between story pairs. In this section, we first describe the text processing techniques to extract cue word clusters, and then present visual features at multiple levels, ranging from the low level visual features (color and texture), through parts-based near-duplicate similarity, to the high-level visual concepts. Finally, we propose a simple but effective method for fusing features of different modalities.

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

Fig. 1. Key-frames from 3 example stories of topic "Bush and Blair meet to discuss Mideast peace;" (a) in Chinese from NTDTV channel, (b) in English from MSNBC channel, and (c) in Arabic from LBC channel. Different near-duplicate groups are indicated in different colors.

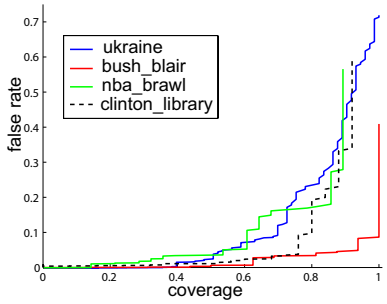

Fig. 2. The coverage and false rate of visual duplicates among 4 topics by varying duplicate thresholds (cf. Section 2.1.3).

2.1. Story-level feature representations

2.1.1. Cue word clusters

We represent the text modality of each story by compact "cue word clusters" or "pseudo-words" [8]. The text transcripts for each story are from the automatic speech recognition (ASR) and machine translation (MT) transcripts included in the TRECVID 2005 data set [2]. The first text processing step involves stemming the ASR and MT tokens and removing the stop words, resulting in a total of 11562 unique words. Then a mutual information (MI) approach is used to select the top 4000 informative words based on the MI between the stories and words. Specifically, given a set of stories D over the word set O, the MI between stories and words can be computed as $I(O;D) = \sum_{o \in O} I(o)$, where $I(o)$ represents the contribution of word o to the total MI $I(O;D)$.

$$I(o) \equiv p(o) \sum_{d \in D} p(d|o) \log \frac{p(d|o)}{p(d)}, \qquad (1)$$

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

These words are further grouped into 120 cue word clusters (or pseudo-words) using the Information Bottleneck (IB) principle [8]. Words in the same cue word cluster are associated with the same semantic – for example, {*insurgent*, *insurgents*, *iraq*, *iraqis*, *iraqi*, *marine*, *marines*, *troops*} or {*budget*, *capitol*, *politics*, *lawmakers*, *legislation*, *legislative*, *reform*}. Later each story is represented as a 120-dimensional pseudo-word frequency vector.

The story-level similarity $\psi_t(s_i, s_j)$ is computed using the cosine similarity between the pseudo-word vectors of story s_i and s_j. Note that these pseudo-word vectors are also weighted by TF-IDF [4] and normalized into unit vectors. Besides, in the word selection and grouping processes the topic labels are not used.

2.1.2. Low-level visual features

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

The low-level visual feature similarity between two stories s_i and s_j is $\psi_l(s_i, s_j) = \max_{i' \in s_i, j' \in s_j} \{\exp - \frac{|x_{i'} - x_{j'}|}{\sigma}\}$, where $\sigma = 2^5$ is empirically determined through cross-validation evaluation. The measure takes the highest low-level similarities between story key-frames since there are usually multiple key-frames within a story (cf. Fig. 1-(a)). Based on that, two stories are considered likely relevant if sharing at least one pair of visually similar shots. In our implementation, all the feature elements are normalized by dividing each dimension with its standard deviation.

2.1.3. Visual duplicates

As shown in Fig. 1, near-duplicates often occur in stories of the same topic. Detection of near-duplicates provides great potential for story linking. In our TRECVID 2005 evaluation work [9], we have found near-duplicate linking to be the most effective tool in the interactive video search task.

For topic tracking, one interesting question arises: how many stories from the same topic actually have near-duplicates. To answer this, we address the following two issues: (1) *coverage* – the percentage of stories that share near-duplicates with other stories in the same topic; (2) *false rate* – the percentage of out-of-topic stories which have duplicates with those within-topic stories. In the ideal case, the coverage is 1 and the false rate is 0. Evaluating over the 4 topics in the data set (cf. Section 3.1), we plot the coverage vs. false rate curves by varying the near-duplicate detection thresholds in Fig. 2. When a higher threshold value is used, we will get a lower coverage and at the same time a lower false rate. It is very impressive to see that we can achieve a moderate coverage (40%-65%) even at almost zero false rate. It strongly supports that story-level duplicate similarity is effective for topic threading.

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

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

2.1.4. Semantic concepts

Besides low-level visual features, detection of high-level semantic concepts has gained significant interest from researchers. NIST TRECVID video retrieval evaluation has included high-level feature detection in the last few years [2]. Research has shown the power of using such concepts in improving video search [9].

For concept detection, we adopted the SVM-based method over two low-level visual features mentioned in Section 2.1.2. Such detection method has been shown to be general and effective [9, 10]. We apply the same detection framework on the whole set of 39 semantic concepts included in the TRECVID 2005 annotations.

A concept is present in a key-frame if its detection confidence score is larger than a threshold. Counting the present concepts across key-frames of the story results in a representing concept vector. Once we have the frequency vector of the visual concepts, we apply the same set of tools for text (in Section 2.1.1) to derive TF-IDF weighting and unit vector normalization.

The story-level semantic concept similarity $\psi_c(s_i, s_j)$ is defined as the cosine similarity between the concept vectors of two stories. Authors of [6] proposed to use mid-frequency concepts for better story representation. In our work, we have observed that TF-IDF weighting on semantic concepts is able to achieve the same effect since such weighting typically suppresses frequent concepts across stories. From our experiment, the cosine similarity on TF-IDF weighted semantic concepts shows $\sim$30% improvement over the "dice" measure used in [6].

In order to assess the influence of individual concepts on topic tracking, we applied the same information-theoretic approach, as described in Eq. 1, to measure the relative MI between each detected concept and stories, and to select the most informative concepts. Tab. 1 shows the ranking of 39 TRECVID 2005 concepts based on this criteria. Later in Section 3.2, we will analyze the effect of concept selection on the accuracy of topic racking. Note that different from [6] the concept tokens used here are from the automatic detectors rather than manual annotations.

2.2. Topic relevance

Using the story-level representations and similarity measures described above, we propose an approach to estimate the relevance score of a new story with respect to a topic. Motivated by [4], we adopt a modified kNN approach to measure the topic relevance score $R_m(s_i)$ of story s_i using modality m . It is defined as follows.

$$R_m(s_i) = \frac{1}{K} \sum_{s_j \in N_K(s_i)} y_{s_j} \cdot \psi_m(s_i, s_j), \qquad (2)$$

where $y_{s_j} \in \{-1, +1\}$ means relevant or irrelevant to the topic and $N_K(s_i)$ are the K nearest sample stories of s_i, measured with the story-level similarity metric $\psi_m(\cdot, \cdot)$ in modality $m \in \{t$: text, l : low-level visual, d : duplicate, c : concept$\}$.

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

We use a linear weighted fusion method for combining the relevance scores from different modalities. Such linear fusion model, though simple, has been shown to be one of the most effective approaches to fuse visual and text modalities in video retrieval and concept detection [9, 10]. For story s_i, the fused topic relevance score $R(s_i) = \sum_m w_m \cdot R_m(s_i)$, where $\sum_m w_m = 1$. The linear weights w_m among modalities are determined empirically based on cross-validation evaluation.

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

Table 1. The 39 TRECVID concepts ordered by MI (Eq. 1).

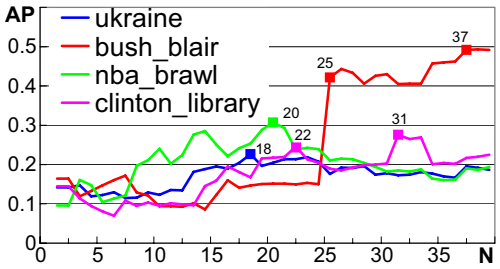

Fig. 3. Topic tracking performance at variant concept dimensions.

3. EXPERIMENTS

3.1. Data set

The data set contains 277 international broadcast news videos from TRECVID 2005 [2], which includes 171 hours of videos from 6 channels in 3 languages (Arabic, English, and Chinese). The time span is from October 30 to December 1, 2004. The story boundaries are from manual annotation, except 42 Chinese news videos are replaced with automatically detected story boundaries, around 0.84 accuracy (See [9] for more explanations), due to unavailability of manual annotations. There are a total of 4247 stories after commercials are excluded from the original total of 5538. The ASR and MT transcripts are provided by NIST [2]. Anchor shots are automatically detected and removed from key-frame sets.

Without official topic annotations, we conducted our own pooling and annotation processes to obtain some topic group truth. First, an unsupervised IB clustering approach [8] was applied on the ASR and MT transcripts to discover the candidate topics in the corpus. Among them, 4 cross-channel topics are manually selected and then annotated following the guidelines in [3]. The topics are: (1) *ukraine*: Ukrainian presidential election, 74 stories; (2) *bush_blair*: Bush and Blair meet to discuss Mideast peace, 16 stories; (3) *nba_brawl*: NBA players fighting with some viewers in the audience, 29 stories; (4) *clinton_library*: Clinton Presidential Library opens, 25 stories.

The tracking is conducted per topic with 2-fold cross-validation. Negative data for each topic are randomly sampled from the negative pool and its size is controlled to be 16 times of that of the positive data. Each experiment is repeated 6 times and then the mean of the performance is calculated. We use the average precision (AP), as the official metric in TRECVID, to be the performance metric. AP corresponds to the area under an ideal (non-

interpolated) recall/precision curve, given a result list ranked based on the relevance scores (cf. Section 2.2). Note that the AP for the random-guess baseline is $1/16 \approx 0.063$.

3.2. Performance and discussions

As shown in Fig. 4, the most significant finding is that near-duplicate plays an important role in story tracking. When used alone, its tracking performance has been very impressive (AP from 0.33 to 0.71), compared to text-based approach (AP from 0.64 to 0.93). It even outperforms text approaches for certain topics, such as *bush_blair*, in which near-duplicates are frequent. When near-duplicate is combined with text, it consistently improves the text-only accuracy by up to 25%.

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

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

For concept-based story tracking, we also compare our TF-IDF weighted representation and cosine similarity with the dice measure used in [6]. The TF-IDF method was found to have a performance gain by about 30%.

In addition, we analyze the impact of using subsets of concepts on topic tracking performance. Fig. 3 shows the tracking performance when only N most informative concepts were included. The informativeness of each concept is computed using Eq. 1, independent of topics. It is interesting to note that different topics reached peak performance at different N values, e.g., 20 for *nba_brawl*, 18 for *ukraine*, and 31 for *clinton_library*. We hypothesize that such difference may be correlated to the diversity of visual content used in each topic and thus may be used to assess the "visual complexity" of each topic. We have found such analysis technique exciting – to the best of our knowledge, it has been the first work on visual complexity assessment of semantic topics. Finally, some concepts show large influence on specific topics, e.g., "sports" for *nba_brawl*, "flag-us" for *bush_blair*, and "walking-running" for *ukraine*. This confirms the finding mentioned earlier.

4. CONCLUSION AND FUTURE WORK

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

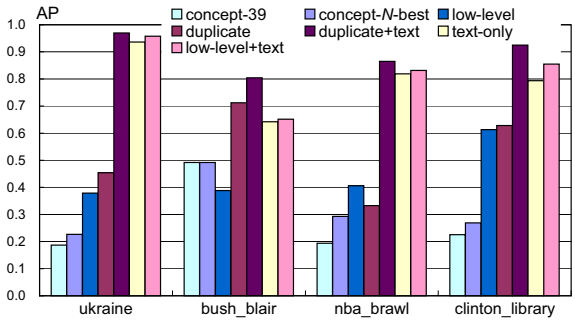

Fig. 4. Topic tracking performance among different modalities and fusion sets. See explanations in Section 3.2.

In future work, we will investigate other story-level similarity measures based on local objects, background, and faces. Furthermore, we are interested in developing techniques to summarize the stories in salient objects for each topic.

5. ACKNOWLEDGMENTS

We thank D.-Q. Zhang for sharing results in near-duplicate detection. This material is based upon work funded in whole by the U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

6. REFERENCES

[1] D.-Q. Zhang and S.-F. Chang, "Detecting image near-duplicate by stochastic attributed relational graph matching with learning," in *ACM Multimedia*, New York, 2004.

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

[3] LDC, "TDT3 evaluation specification version 2.7," 1999.

[4] Y. Yang *et al.*, "Learning approaches for detecting and tracking news events," *IEEE Intelligent Systems*, vol. 14, no. 4, 1999.

[5] L. Xie *et al.*, "Discover meaningful multimedia patterns with audio-visual concepts and associated text," in *ICIP*, Singapore, 2004.

[6] J. Kender *et al.*, "Visual concepts for news story tracking: Analyzing and exploiting the NIST TRECVID video annotation experiment," in *CVPR*, San Diego, 2005.

[7] Y. Zhai and M. Shah, "Tracking news stories across different sources," in *ACM Multimedia*, Singapore, 2005.

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

[9] S.-F. Chang *et al.*, "Columbia University TRECVID-2005 video search and high-level feature extraction," in *TRECVID Workshop*, Washington DC, 2005.

[10] A. Amir *et al.*, "IBM Research TRECVID-2004 video retrieval system," in *TRECVID Workshop*, Washington DC, 2004.