ADAPTIVE VIDEO NEWS STORY TRACKING BASED ON EARTH MOVER'S DISTANCE

Mats Uddenfeldt

Uppsala University, Signals & Systems
Box 528, 751 20 Uppsala, Sweden
mats@uddenfeldt.se

Keiichiro Hoashi, Kazunori Matsumoto, Fumiaki Sugaya

KDDI R&D Laboratories, Inc.
2-1-15 Ohara Fujimino, Saitama, Japan
{hoashi, matsu, fsugaya}@kddilabs.jp

ABSTRACT

This paper proposes an adaptive system for video news story tracking based on the Earth Mover's Distance (EMD). When an interesting story appears in the news, it is flagged manually as a topic for tracking. Our system then tracks the events as they unfold over time and present accumulated results to the user for feedback. This feedback is used to adapt the topic model to changes in the tracked story. EMD provides the system with a robust way of performing many-to-many matching of news stories independent of the temporal order of their contents. This is particularly suitable in the news genre as stories often are subjected to video editing between shows. Experiments have been run with a range of topics and show promising results.

1. INTRODUCTION

Every day there is an abundance of information broadcasted by all the news networks of the world. The reports are instant and cover an impressive range of topics, including natural disasters, politics, economics, sports and other events. Due to the range of topics and various target audiences, the nature of the stories are very different. In addition to this, reports may be biased and differ depending on the individual network. This makes watching news stories from multiple sources necessary to provide us with a more meaningful view. However, the total number of stories broadcasted every day makes it a daunting task to watch everything.

In this paper, we investigate a novel approach to news story tracking, in which we adapt our query based on accumulated results and user feedback. We use visual features extracted from keyframes and the Earth Mover's Distance [1] (EMD) to provide a similarity measure between stories.

2. BACKGROUND

There are several ways to achieve semantic linking between news stories. An important starting point is how a viewer perceives the broadcasts. Given two news stories our first impression is their visual information. Besides the visual information we also acquire language information from the video in the form of audio and closed captions on screen. Semantic linkage is accomplished using either visual or language information, or a combination of the two. Ide et al. [2] presented a database management system for the purpose of structuring a news video archive. The news programs are first segmented into topics by applying morphological and semantic analysis of closed-caption text. The topics are then threaded into the video database in a chronological order, based on their semantic linkage. In their approach the user can manually track the story by following relevant links in the threaded news archive. Zhai et al. [3] proposed an advanced framework using a fusion of visual and language information. They used two methods for visual data, one for facial and one for non-facial keyframes, and used automatic speech recognition (ASR) to gain access to additional textual information. This information is used to compute a similarity mapping to match a given query story with other stories in the video archive.

While the text-based approach has proved to be effective for story tracking, such methods are dependent on the accuracy of the text extraction process. Therefore, text-based methods are not expected to be effective for video contents with spontaneous speech, on which erroneous ASR results are expected. Furthermore, existing video-based methods such as Zhai et al. utilize sophisticated video processing technologies, which are computationally expensive, and thus are difficult to implement in a practical application. Due to these problems, we choose to evaluate a video story tracking system based solely on basic visual features of news stories.

3. PROPOSED TRACKING SYSTEM

In this section we will present our adaptive news story tracking system. We will assume that the news feed has already been segmented into shots and stories. This segmentation can be achieved using many different techniques [4, 5], neither of which is favored by our system.

3.1. Feature extraction

When a new story is observed by the system, we need to create a representation of it in order to compare it to the previous ones. To extract features from a shot, a keyframe is

selected for every shot and the color histogram is computed from that keyframe. We use Haar encoded color histograms in the HSV color space, as defined for the Scalable Color Descriptor (SCD) in the MPEG-7 Visual Standard [6].

3.2. Story-based similarity measures

Since our system relies on visual features to describe the contents of a story, we needed to explore possible methods of achieving a story-based similarity measure. Existing ways to provide similarity ranking between stories, or video clips as they are more generally referred to, have been built on top of shot-based retrieval in [7, 8]. In addition to using shot similarity, these methods also takes other features like temporal order, granularity (level of one-to-one matching), and interference (percentage of unmatched shots) into account to calculate clip similarity. In the news genre the individual stories are often subjected to video editing, to change the shot order or to segment long shots into multiple shots. This makes methods which enforce a one-to-one mapping between shots inappropriate. However, the method presented in [9] by Peng et al. suggests that one can successfully use EMD to provide a many-to-many mapping between the shots of two stories.

To be able to use EMD as a story-based similarity measure we implemented a story representation along the lines presented in [9]. Each story is represented as a weighted graph composed of shots. Each shot is represented as a tuple composed of the color histogram feature of the keyframe and the duration (number of frames) of the shot. In analogy with the original transportation problem, the shots of story A are considered to be the suppliers and the shots of story B the consumers. The cost of transporting a single frame between two shots A_i and B_j is calculated using a histogram distance measure. Here we take a different approach from [9], and use the normalized L1-distance as our cost function.

We can now use EMD to find the minimum expensive flow to transport frames between the supplier shots and the consumer shots. Since we are using a normalized cost function, EMD will also return a normalized value between 0 and 1. This value is used to measure the distance between stories.

3.3. Adaptive tracking of news stories

We propose an adaptive news story tracking system in order to be able to follow a story as it develops in the news over time. A flow-chart overview of the system can be seen in Figure 1. The system is initiated when a user flags an interesting story for tracking. This story is used as the initial topic model. When a new story enters the system, its features will be extracted and compared against all the stories of the model. If it is similar to one of the stories, it is considered similar to the entire model, and will be added to the result list. After a given interval, the list of results is presented to the user for

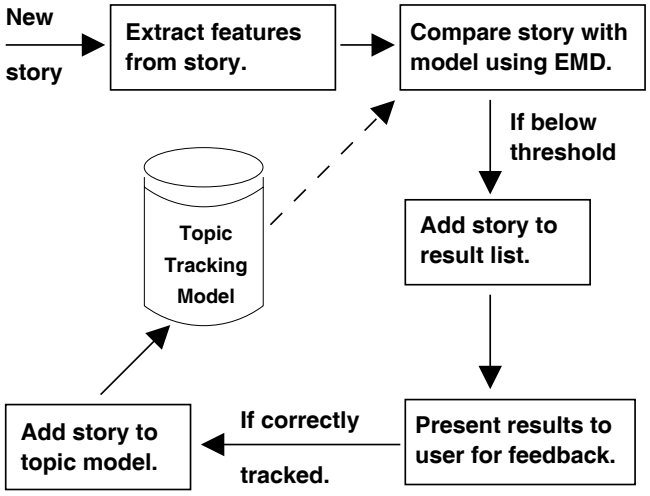

Fig. 1. A flow-chart overview of the adaptive tracking system.

feedback. This is the adaptive step of our tracking system. The user will go through the list and decide which stories have been correctly tracked. The correct stories are used to update the topic model to reflect changes in the tracked story. Tracking will cease when a given time-out period has passed without a single correct result.

The topic model is initially composed only of the manually flagged story, but as time progresses the model is updated with more recent stories. The model is operated as a FIFO (First In, First Out) list of the latest correctly tracked stories. Major parameters of the system are: the size of the model N, which expresses the number of stories stored in the topic model, the distance threshold S, which controls the number of observed stories on the result list, and the time-out period T, which defines when the system stops tracking a story.

4. EXPERIMENTS

This section provides the details of the experiments we have conducted to measure the effectiveness of our system.

4.1. Experiment data

For evaluation of the proposed method, we have prepared a video database which consists of approximately 100 hours of Japanese news shows recorded for a two-month period (March to May 2005). Story boundaries of all individual stories in the video database, as well as specific story labels, which indicate major stories that occurred within the period, were applied manually. These specific story labels are used as reference data for the following experiments, and can be seen in Table 1. Furthermore, we also applied labels for the studio shots and non-studio shots, i.e., video footage such as reports from the site of the story. The non-studio shots are used to represent each story, since the studio shots do not provide any

Table 1. Summary of labeled stories in experiment data (March to May 2005): *Duration* is the time between first and last appearance. *Shows* is the number of shows in which the story appears and *stories* is the total number of appearances.

Topic	Duration	Shows	Stories	Description
Anti Japan Demonstration	21 days	28	131	Anti-Japan demonstrations and riots in China.
Derailment Accident	6 days	11	79	Derailment accident of Japanese train in Amagasaki (4/25).
Earthquake, Fukuoka	30 days	14	29	Major earthquake at Fukuoka, Japan (3/21).
Earthquake, Sumatra	40 days	15	30	Major earthquake at Sumatra (3/28).
Expo 2005	54 days	24	29	World Exposition 2005 held in Aichi, Japan.
Kokudo	39 days	11	20	Arrest of Kokudo president Yoshiaki Tsutsumi.
Livedoor	48 days	52	123	The M&A of Livedoor, a Japanese IT company, and Fuji TV.
North Korea	59 days	15	22	Stories about North Korea related incidents.
Pirate	8 days	13	32	Pirate assault of Japanese tug boat in Malaysia.
Pontiff	22 days	14	40	Death of Pope John Paul II, election of Pope Benedict XVI.
Non-labeled stories	62 days	103	1272	

meaningful visual information of the current news story.

4.2. Method

We have implemented a simulation of our adaptive topic tracking system. For the purpose of feature extraction, we quantized the HSV color space into 256 distinct color sets, with hue quantized into 16 bins, saturation and value into 4 bins respectively. The first chronological appearance of a labeled story was chosen to initiate the topic model. User feedback was simulated using the provided topic story labels, and given after every story processed by the system. The tracking system was tested with the following parameters: $N = \{1, 3, 5, 7\}$, $S = \{0.50, 0.55, 0.60, 0.65\}$, and $T = \{1, 3, 5, 7\}$ days.

4.3. Evaluation measures

The results of the experiments are evaluated by conventional measures of information retrieval: precision and recall. A result is every story of which distance is below the threshold S for EMD comparison. The ground truth for a given topic consists of all the stories, after the original story, with an identical topic label. This is a rather strict limitation, but we do not agree that a story can be "somehow relevant" as defined in [3]. If a potential result has a label which does not exactly match the given topic, it is considered a false alarm. Based on these measures we can calculate precision (P) and recall (R) after the system has finished running. Furthermore, we calculated the F-measure based on the following formula: $F = \frac{2PR}{P+R}$.

4.4. Results

We evaluated our system to discover which combination of parameters lead to the best results. The similarity values of our data set are normally distributed with a mean of 0.654 and $\sigma = 0.088$. A distance threshold of 0.50 causes the system to fail tracking 4 out of the 10 stories of Table 1, due to being too restrictive. On the other hand, a threshold of 0.65 proved to be too inclusive. We therefore exclude the thresholds 0.50 and 0.65 from our results. The mean F-measures for $S = \{0.55, 0.60\}$, $N = 7$ and $T = 1$ are 0.23 and 0.22 respectively, with increasingly higher recall (0.39, 0.50) and lower precision (0.18, 0.15). The trend is the same for all N and all T in our experiments. Since our system relies on user feedback, we want to put emphasis on precision. Hence, $S = 0.55$ provides an optimum threshold for our data set. Similarly, increasing time-out values, $T = \{1, 3, 5, 7\}$ with $S = 0.55$ and $N = 3$ results in the following F-measure means: 0.23, 0.19, 0.17 and 0.16 with decreasing precision (0.21, 0.17, 0.16, 0.15). Using a long time-out period will reduce the chance of ending tracking prematurely, but extend the time we continue to track a canceled story. Major news stories, like the ones we are interested in, often appear on a daily basis, which is why $T = 1$ show the best results.

The most interesting parameter turns out to be the size of the topic model, N. Using the optimum parameters $S = 0.55$, and $T = 1$, we investigated the results for varying the size of the model. Basing tracking on a single ($N = 1$) story, discards all references to older footage in the topic model. In this case, the mean F-measure is only 0.10, with 2 stories failing to be tracked. The results for $N = \{3, 5, 7\}$ can be seen in Table 2. Different trends in the precision, recall, and F-measure can be observed with the change of parameter N. Some of the stories give a better result when tracked with a larger model, and some with a smaller one. However, $N = 5$ appears to provide a stable middle ground. Comparing the results of Table 2, where *Derailment Accident* is in the top, with the Shows-column of Table 1, we can see that the more frequently a story appears, the better results we get from a high N. With less frequent stories, the opposite appears to be true. This trend will be interesting to test with data from multiple networks. Four stories were omitted from Table 2, due to their poor results (F-measure < 0.10). Closer inspection shows that these stories are not individual stories, but rather groups of stories, e.g., the *Pontiff* topic is split up into the the

Table 2. Results with $N = \{3, 5, 7\}$, $S = 0.55$ and $T = 1$.

Topic	N	P	R	F
Anti Japan Demonstration	3	0.38	0.52	0.44
	5	0.35	0.62	0.45
	7	0.35	0.71	0.47
Derailment Accident	3	0.78	0.44	0.56
	5	0.63	0.56	0.59
	7	0.60	0.66	0.63
Earthquake, Fukuoka	3	0.18	0.50	0.26
	5	0.17	0.54	0.25
	7	0.15	0.54	0.24
Kokudo	3	0.21	0.44	0.29
	5	0.15	0.50	0.23
	7	0.14	0.56	0.22
Livedoor	3	0.19	0.44	0.26
	5	0.16	0.48	0.24
	7	0.17	0.59	0.27
Pirate	3	0.14	0.31	0.20
	5	0.18	0.54	0.27
	7	0.17	0.54	0.26

death of Pope John Paul II and the election of Pope Benedict XVI, and *Expo 2005* features many different stories.

5. DISCUSSION

While various tracking systems using visual and language information have previously been designed, our method is unique because the nature of the query is adaptive. Instead of a simple query based system for retrieval out of a database, we provide an online system which can be set to watch any news stream and track interesting stories. The advantages of our system are that it requires no training and that it relies on low-level video features. Therefore, it is highly suitable for use in an online environment, where requests to track incoming video are expected to be submitted continuously.

We have conducted our evaluation with a very strict measure of truth, i.e., the semantic context of a story has to match the tracked topic exactly to be considered correct. Despite this very strict definition, we are able to track the semantic content of news stories, using only visual features. However, it should be noted that the system currently takes a naive approach to keyframe extraction, by extracting the middle frame of every shot. Implementing more advanced keyframe selection and comparison methods should lead to better results.

The experiments in this paper simulate a situation where the user actively sends feedback information to the system, which applies a heavy burden to system users. While such conditions are difficult to be accepted in a practical application, user interaction can be minimized by various methods, such as utilizing "implicit" feedback information, e.g., user logs of viewed stories, to update the topic model.

6. CONCLUSIONS

In this paper we have described a method to provide adaptive tracking of news stories based on EMD. We have shown that this is possible and investigated which parameters give the best results. The future direction of our work will be to investigate how the system can be further improved, and to incorporate the story segmentation methods presented in [5] in order to achieve an even more autonomous system.

7. ACKNOWLEDGMENTS

The work performed by Mats Uddenfeldt of Uppsala University was partially funded by the Sweden-Japan Foundation. We also extend thanks to Yossi Rubner for his EMD code.

8. REFERENCES

[1] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[2] Ichiro Ide, Hiroshi Mo, Norio Katayama, and Shin'ichi Satoh, "Topic threading for structuring a large-scale news video archive," in *CIVR*, 2004, pp. 123–131.

[3] Yun Zhai and Mubarak Shah, "Tracking news stories across different sources," in *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*. 2005, pp. 2–10, ACM Press.

[4] Chong-Wah Ngo, Ting-Chuen Pong, Roland T. Chin, and HongJiang Zhang, "Motion-based video representation for scene change detection," in *ICPR*, 2000, pp. 1827–1830.

[5] Keiichiro Hoashi et al., "Video story segmentation and its application to personal video recorders," in *CIVR*, 2005, pp. 39–48.

[6] Thomas Sikora, "The mpeg-7 visual standard for content description-an overview," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 11, no. 6, pp. 696–702, 2001.

[7] Liping Chen and Tat-Seng Chua, "A match and tiling approach to content-based video retrieval," in *ICME*, 2001.

[8] Yuxin Peng and Chong-Wah Ngo, "Clip-based similarity measure for hierarchical video retrieval," in *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, New York, NY, USA, 2004, pp. 53–60, ACM Press.

[9] Yuxin Peng and Chong-Wah Ngo, "Emd-based video clip retrieval by many-to-many matching," in *CIVR*, 2005, pp. 71–81.