News Story Clustering from Both What and How Aspects: Using Bag of Word Model and Affinity Propagation

Wei-Ta Chu[1], Chao-Chin Huang[1], and Wen-Fang Cheng[2]
[1]National Chung Cheng University, Chiayi, Taiwan
wtchu@cs.ccu.edu.tw, jackjack200142@gmail.com
[2]Industrial Technology Research Institute, Hsinchu, Taiwan
wfcheng@itri.org.tw

ABSTRACT
The 24-hour news TV channels repeat the same news stories again and again. In this paper we cluster hundreds of news stories broadcasted in a day into dozens of clusters according to topics, and thus facilitate efficient browsing and summarization. The proposed system automatically removes commercial breaks and detects anchorpersons, and then determines boundaries of news stories. Semantic concepts, the bag of visual word model and the bag of trajectory model are used to describe what and how objects present in news stories. After measuring similarity between stories by the earth mover's distance, the affinity propagation algorithm is utilized to cluster stories of the same topic together. The experimental results show that with the proposed methods sophisticated news stories can be effectively clustered.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering, information filtering*. I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding – *video anlaysis*.

General Terms
Algorithms, Performance, Experimentation.

Keywords
News story clustering, bag of visual word, bag of trajectory, semantic features, earth mover's distance, affinity propagation.

1. INTRODUCTION
Nowadays large amounts of TV channels broadcast news 24 hours a day. We can easily receive news by turning on the TV. However, everyday only dozens of stories are new to the audience, and in most time the same stories repeat again and again, while interleaving with variations of old and fresh content. Therefore, clustering topic-related news stories is interesting and urgently demanding as it is the fundamental step for news browsing, retrieval, and summarization.

It's worth clarifying some related terminologies. A *news shot* is a video shot containing part of news content, which may be shots with the anchorperson, interview, or events like parade or car accident. A *news story* may include several news shots to completely convey a message. It often contains a sequence of shots with the anchorperson and the event itself, and ends until the anchorperson reports the next message. A *news topic* may contain several news stories describing evolution of an event over time, such as "Congressman shooting" or "Steve Jobs' sick leave".

The goal of this work is to continuously monitor news broadcast and cluster topic-related stories into a number of groups that is much fewer than the total number of news stories. A few challenges should be addressed. Firstly, a lot of irrelevant materials, such as commercial breaks, should be eliminated to not only reduce computation complexity but also increase clustering accuracy. Secondly, semantics of news stories should be appropriately represented to well calculate similarity between stories. Rich information and significant visual variations in news videos make this issue more troublesome. Thirdly, the appropriate number of news clusters for a day is not known in advance, and thus an elegant clustering scheme is needed.

Generic concept detectors have been proposed to detect object/scene/event in TRECVID news corpus. However, most studies focus on detection in keyframes extracted from shots, rather than thinking a news story in a whole. In addition, imperfect text-based tags can hardly well represent significant visual variations. One recent work on news shots can be seen in [10]. Wu et al. [11] propose a multimodal news story clustering framework. They treat news stories as the basic analysis unit and exploit co-clustering scheme with near-duplicate constraints to link stories. However, in their work boundaries of stories are known in advance, which is not a trivial task in a real news monitoring system. Moreover, they detect near-duplicate keyframes as a matching constraint, without considering motion information in stories. To address the issues described above, we propose a system to automatically cluster topic-related news stories from a continuously-captured lengthy news video.

Figure 1 shows the system flowchart. We keep recording video broadcasted from TV news channels for eight to ten hours. A lengthy video is first segmented into shots, and the parts with commercials are removed. For the remaining shots, anchorpersons are detected, and thus boundaries between news stories are determined. What objects and how objects move in a news story is described by the bag of visual word model and the bag of trajectory model, and each news story is then represented by a sequence of feature vectors. Earth mover's distance is used to evaluate similarity between news stories, and then the affinity propagation algorithm is adopted to conduct clustering. Finally, news stories conveying the same topic are clustered together.

AIEMPro'11, Nov. 28–Dec. 2, 2011, Scottsdale, Arizona, USA.

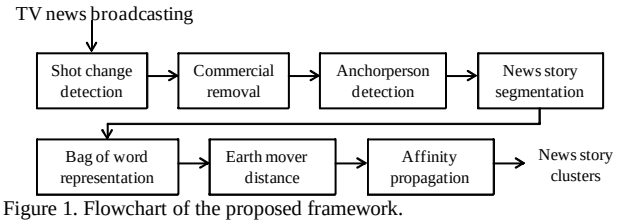

Figure 1. Flowchart of the proposed framework.

Contributions of this work are summarized as follows:

- Automatic story segmentation: this system automatically removes commercial breaks and detects anchorpersons based on face information to segment news stories.
- Bag of word model: we describe each news story from both *what* and *how* aspects, which is verified to be more robust to characterize news content.
- Clustering: the affinity propagation algorithm is exploited to cluster topic-related stories without any prior knowledge.

The rest of this paper is organized as follows. Section 2 provides literature survey. Section 3 describes how we segment and represent news stories. In Section 4, similarity between news stories is calculated and thus clustering is conducted. Experimental results are reported in Section 5, followed by the concluding remarks in Section 6.

2. RELATED WORK

2.1 News Story Clustering

Clustering similar video clips to facilitate browsing and annotation has a long history. Zhong et al. [13] propose an earlier method to conduct video shot clustering based on color, motion, and temporal variance. Specifically for news videos, topic tracking or news story clustering facilitate us to access large volume of news video corpus. Ide et al. [14] recognize closed caption into text, and then use text information to segment news topics and conduct topic tracking. Also based on text information, Nallapati et al. [15] investigate methods for modeling the structure of a topic in terms of its events. They propose an event model to cluster stories and describe dependency between them. Zhai and Shah [18] present a semantic linking method to find similar news stories across sources. They consider both facial and non-facial keyframes, and language correlation based on automatic speech recognition. Hsu and Chang [16] propose representation and similarity measure for news videos based on visual features, visual near-duplicates, and semantic concepts. Given a news story, they measure its topic relevance score by fusing the aforementioned features. Wu et al. [11] propose a co-clustering algorithm, with the constraint derived from near-duplicate detection, to mine topic-related news stories. Similar to [16], they find that near-duplicate detection provides important clues to detect topic-related stories. Following the work in [14], Wu et al. [17] first search stories related to the query video based only on textual information, and then near-duplicate detection is adopted to further polish the detection results. A query-expansion algorithm is further proposed to rank stories in the same group.

In [14], although news threads can be elaborately discovered, closed caption does not exist for all news broadcasting. The work in [15] only works for text news. Zhai and Shah [18] mainly rely on face information and spoken content to link topic-related topics. In our work, we would like to exploit more robust features recently proposed. The work in [16] reported promising performance. However, the topics they experimented are too rough, such as the topic *bush_blair*. Furthermore, given a targeted topic, they evaluate the relevance score of a news story to this topic. In [11], number of topic in the news collection is needed before clustering. In our work, number of topic and characteristics of these topics are not known in advance. Given an eight-hour long news videos, we would like to cluster stories into appropriate number of groups without any prior knowledge. Comparing with [17], our system segments a news video into appropriate number of groups, without any initial query video as the template.

2.2 Video Copy Detection vs. Near-Duplicate Detection

Video copy detection refers to determine whether some videos in a database contain content similar to the query video. If we take a news story as the query video, and detect video copies from a collection of news stories, news story clustering can be achieved.

Near-duplicate detection is another technique highly related to news video clustering [11][17]. There may be a large number of near-duplicate frames in topic-related news stories. Comparing with video copy detection, near-duplicate detection for videos tends to consider perceptual similarity. In general, video copies often differ from each other in visual editing, such as gamma/contrast change, resolution change and re-encoding. On the other hand, near duplicates often differ from each other in not only visual editing, but also camera parameter changes, photometric changes, and scene changes. The last factor may cause significant content variations. Overall, the definition of near-duplicate is more extensive than video copy.

For clustering topic-related news stories in the same TV channel, techniques of video copy detection may be useful because topic-related news stories have highly similar visual content. For clustering topic-related news stories across TV channels, techniques of near-duplicate detection may be useful because topic-related news stories have relatively varied visual content but convey the same semantics. However, because different channels have different broadcasting styles, or even different news resources, in many cases neither video copy detection nor near-duplicate detection techniques are sufficient to handle news story clustering. This problem becomes more severe nowadays because of fancy special effects (e.g. marquee) on screen and unconventional broadcasting styles (e.g. dual anchor or multi-party conversation).

3. NEWS STORY SEGMENTATION AND REPRESENTATION

3.1 News Story Segmentation

The recorded video is segmented into shots according to frame differences based on linear combination of YCbCr color histogram and edge change ratio [1]. Because we continuously monitor and capture video signals from TV news broadcasting, many commercial breaks would be captured. To remove commercial breaks, the idea proposed in [12] is adopted. We measure how likely a shot belongs to a commercial break by calculating shot change rate in its neighborhood. If a video shot starts at the ith second, its "commercial likeness" is the number of shot changes in the range $(i-30, i+30)$. The shots with commercial likeness higher than a threshold are determined as in

commercial breaks. The threshold can be set according to regulations governing advertisements in different countries. In our work, the shots that have the largest 20% commercial likeness are determined as in commercial breaks and are not considered in the following processes. A video section that contains consecutive non-commercial shots is viewed as a part of news program, and is called a *news section* in the following.

Most news stories start with a shot where the anchorperson reads the script, have some shots of interviews or events in the following, and end with a shot where the anchorperson reports the next news story. Therefore, appearance of the anchorperson provides important clues for news story segmentation. Because the anchorperson appears more frequently than others, we determine the "major face" [6] for each news section to detect the anchorperson.

For each shot in a news section, we first detect faces and connect temporally adjacent faces that are similar and spatially close as a face track. Relationship between face tracks in a news section is modeled as a weighted undirected graph $G = (V, E)$, where each node $v \in V$ denotes a face track, and an edge $e_{i,j} \in E$ connecting v_i and v_j contains a weighting defined as the similarity between these two face tracks. After this process, finding the most frequent face in the news section has been transformed into finding the largest subgraph in the graph G. Details of graph partition please refer to [6].

In our work, appearance of the major face indicates boundaries of news stories. Note that the major face is adaptively determined for each news section. Therefore, the proposed work can automatically segment news stories for an 8-hour-long news program where the anchorperson may change. This method does not need training data for face recognition. In the following text, the shot consisting of anchorperson is eliminated from each news story, and only the news content is described.

3.2 News Story Representation

We describe *what* and *how* [2] objects/events present based on the bag of visual word model and semantic concept detection, and the bag of trajectory model, respectively.

- Bag of visual word (BoW)

For shots in a news story, we sample one out of every five frames and use the TOP-SURF visual word toolkit [3] to transform every selected frame as a visual word histogram. Visual word histograms for sampled frames in the same shot are then averaged. Finally, a news story i consisting of a sequence of shots is transformed into a sequence of visual word histograms:

$$BoW_i = (\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_m), \quad (1)$$

where $\boldsymbol{h}_k$, $1 \le k \le m$, denotes the average visual word histogram for the kth shot in the ith news story, and m is the number of shots in this story. This representation describes weights indicating presence and absence of visual words, and characterize news stories from the *what* aspect.

- Bag of trajectory (BoT)

Similar to the model described above, we sample one out of every five frames from shots in a news story to reduce computation. From each selected frame we extract the SURF (Speech Up Robust Features) feature points [4]. The KLT (Kanade-Lucas-Tomasi) algorithm is then applied to conduct feature tracking in consecutive selected frames. Motion vectors of matched feature points are determined, and sequences of matched feature points in consecutive selected frames constructs motion trajectories.

For a shot, a large number of trajectories with different lengths may be extracted. To efficiently represent a trajectory, we collect statistics of moving directions in trajectories [5]. Moving direction is categorized into five classes: moving toward up-right (denoted by 1), moving toward up-left (denoted by 2), moving toward left-bottom (denoted by 3), moving toward right-bottom (denoted by 4), and no movement (denoted by 0). We calculate the probability of each moving direction and form a 5-dimensional vector to describe a trajectory. For example, if moving directions of a trajectory of five frames are (4, 1, 2, 2), they are transformed as the vector (0:0.0, 1:0.25, 2:0.5, 3:0.0, 4:0.25), in which (*m*:*n*) indicates the probability of moving toward direction *m* is *n*.

Motion trajectories are viewed as the basic elements to describe how object moves in videos [5]. We conceptually map a video into a document, and map trajectories into visual words for constituting the document. Feature vectors transformed from motion trajectories are clustered by the k-means algorithm. The ones that are grouped into the same cluster are claimed to represent the same bag of trajectory (BoT) word. A BoT word conceptually represents a set of trajectories that have similar moving evolutions. A video shot d that consists of many trajectories, therefore, is transformed into a BoT word histogram $\boldsymbol{t} = \{n_{1,d}, n_{2,d}, ..., n_{K,d}\}$, in which $n_{i,d}$ denotes the number of trajectories corresponding to the ith BoT word b_i. The value K is the number of different BoT words, i.e. number of clusters.

Different BoT words have different influences on describing documents. From the study of natural language processing, we can measure the importance of a BoT word by TF-IDF (term frequency – inverse document frequency):

$$w_i = \frac{n_{i,d}}{n_d} \log \frac{D}{n_i}, \quad (2)$$

where n_d denotes the number of BoT words (number of trajectories) in the document (video shot) d, n_i denotes the number of documents that contain b_i, and D denotes the total number of document. If b_i occurs frequently in the document d but rarely occurs in other documents, it's a more important BoT word to describe the video shot d.

After the processes described above, we transform each video shot as a K-dim vector $\boldsymbol{g} = (w_1, w_2, .., w_K)$, in which w_i denotes the weighting corresponding to the ith BoT word. Finally, a news story i consisting of a sequence of shots is transformed into a sequence of BoT word histograms:

$$BoT_i = (\boldsymbol{g}_1, \boldsymbol{g}_2, ..., \boldsymbol{g}_m), \quad (3)$$

where $\boldsymbol{g}_k$, $1 \le k \le m$, denotes the BoT word histogram for the kth shot in the ith news story, and m is the number of shots in this story.

- Semantic features

Though we describe visual and motion information of news stories, topic-related news stories across channels may have significant visual variations. Therefore, we further detect semantic concepts in videos to describe news stories. Our work utilizes the VIREO-374 concept detectors [19] to detect semantic features of every video shot. The VIREO-374 concept detectors contain 374

concepts which are modeled by support vector machines (SVM). Following the setting of VIREO-374, we select one keyframe for each shot, represent each keyframe by visual words, and estimate concept scores based on the pre-trained concept models. Because we have voluminous video shots, and a lot of time is needed to detect all 374 concepts, we only detect 39 of 374 concepts that are the same as LSCOM-lite [20] to reduce computation. Every keyframe are thus associated with 39 concept scores, and we concatenate these concept scores as a feature vector. This feature vector is normalized and describes presence of concepts in a keyframe. Finally, a news story i consisting of a sequence of shots is transformed into a sequence of concept score vectors:

$$SC_i = (\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_m), \quad (4)$$

where $\boldsymbol{c}_k$ is a 39-dim concept score vector.

4. NEWS STORY CLUSTERING

4.1 Earth Mover's Distance

To cluster similar news stories, we have to measure similarity between news stories of unequal lengths. In this work, we calculate the earth mover's distance (EMD) between news stories based on the representation mentioned above, due to its promising performance in several domains [7][8]. Note that other distance measures, such as SQFD [21], capable to deal with unequal-length data sequences can be used under the proposed framework. In the following, EMDs are respectively calculated based on BoW, BoT, and semantics, and are then integrated to jointly consider *what* and *how* aspects.

Taking BoW-based EMD as the example, a news story P is represented as $P = \{(\boldsymbol{h}_1^p, w_1^p), ..., (\boldsymbol{h}_n^p, w_n^p)\}$, where $\boldsymbol{h}_i^p$ is the average visual word histogram for the ith shot in P, n is the number of shots in P. The weight w_i^p serves as the total supply of the suppliers or the total capacity of consumers in the EMD method. It is defined as ratio of the length of the ith shot to the total length of all shots in P. Similarly, another news story Q is represented as $Q = \{(\boldsymbol{h}_1^q, w_1^q), ..., (\boldsymbol{h}_m^q, w_m^q)\}$. The EMD between P and Q is computed by

$$D_{BoW}(P, Q) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} \hat{f}_{ij} d_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} \hat{f}_{ij}}, \quad (5)$$

where d_{ij} is the ground distance between $\boldsymbol{h}_i^p$ and $\boldsymbol{h}_j^q$ and is defined as the Euclidean distance. The optimal flow $\hat{f}_{ij}$ is determined by solving the following linear programming problem:

$$\hat{f}_{ij} = \arg\min_{f_{ij}} \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij} d_{ij}$$

$$s.t. \quad \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij} = \min\left(\sum_{i=1}^{n} w_i^p, \sum_{j=1}^{m} w_j^q\right); \; f_{ij} \geq 0;$$

$$\sum_{i=1}^{n} f_{ij} \leq w_i^p, 1 \leq i \leq n; \; \sum_{j=1}^{m} f_{ij} \leq w_j^q, 1 \leq j \leq m. \quad (6)$$

Similar method is used to calculate EMD based on BoT and semantic representation. Three EMDs are then linearly combined to form the integrated distance between two news stories P and Q:

$$D(P, Q) = \alpha D_{BoW}(P, Q) + \beta D_{BoT}(P, Q) + \gamma D_{sc}(P, Q), \quad (7)$$

where α, β, and γ control importance between different factors, and range from 0 to 1.

4.2 News Story Clustering

After measuring distance between any two news stories, we would like to cluster similar stories into the same group. Because we don't know the exact number of clusters in advance, the affinity propagation (AP) approach [9] is adopted for this task. The AP algorithm takes similarity between stories as input, randomly chooses an initial subset of stories as examplars, and iteratively exchanges messages between examplars and other data points until convergence. Two types of messages are considered: responsibility and availability. The responsibility message $r(m, n)$ indicates how well point n serves as the examplar for point m. The availability $a(m, n)$ indicates how well point m chooses point n as its examplar. Jointly considering two messages indicates how likely points m and n should be clustered together.

Similarity between two stories v_i and v_j is defined as

$$s_{ij} = e^{-D(v_i, v_j)} \times \begin{cases} \log_k |t_j - t_i|, & \text{if } |t_j - t_i| \leq k, \\ 1, & \text{otherwise.} \end{cases} \quad (8)$$

The first term is derived from the integrated EMD between v_i and v_j. The value s_{ij} is larger if $D(v_i, v_j)$ is smaller. The second terms is specially designed to consider temporal distance between two stories in the same TV channel, in which t_i denotes that the story v_i is the t_i-th story from the beginning of the video. The logarithm to base k is monotonically increasing until $t_j - t_i$ reaches k. The number k can be set according to the approximate period of topic-related news stories would repeat. For example, if topic-related stories tend to be reported every thirty stories, this value can be approximately set as 30. This value may depend on broadcasting styles of different TV channels. In broadcasting news, it's rare to repeat the same news in a short time period. Therefore, the value s_{ij} is larger if v_i is at a larger temporal distance from v_j. Given similarity values between stories, the AP algorithm clusters news stores into several groups, and stories in the same group are viewed as in the same news cluster.

5. EXPERIMENTS

5.1 Evaluation Dataset

We capture news videos from four news TV channels. Each video includes all things being broadcasted, e.g., news programs and commercials, and lasts for eight to ten hours. Table 1 shows detailed information. We manually define ground truths of story boundaries and news topics. There are totally 772 news stories that cover 334 topics. Broadcasting styles of different TV channels are varied. The video from TV3 is especially edited and relatively has bad visual quality, and thus many shot boundaries are falsely detected. Figure 2 shows some snapshots of the evaluation data, from which we can realize visual complexity of these videos and high variations between different channels.

Table 1. Information of the evaluation dataset.

	Duration	# news stories	# topics	# video shots
TV1	8 hours	155	78	7529
TV2	8 hours	176	85	9028
TV3	10 hours	240	91	29088
TV4	10 hours	201	80	7898
Total	36 hours	772	334	53543

5.2 Performance Evaluation

- Performance of news story segmentation

Figure 3 shows performance of news segmentation. Precision and recall in each channel are about between 0.6 and 0.7. Two factors influence this performance: 1) Commercial removal: Some news stories have frequent shot changes because of fancy broadcasting styles (c.f. Figure 2). This largely increases difficulty of commercial detection. 2) Anchorperson detection: Some news stories contain dual anchorpersons, and they interlace to report

news. We assume anchorpersons in a news section appear more than other detected faces. If there is more than one anchorperson, one of them would be miss-detected.

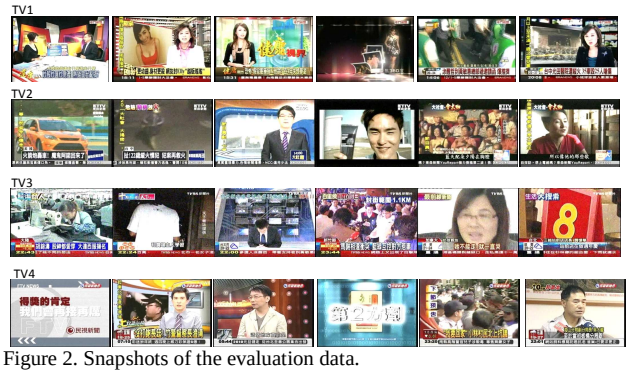

Figure 2. Snapshots of the evaluation data.

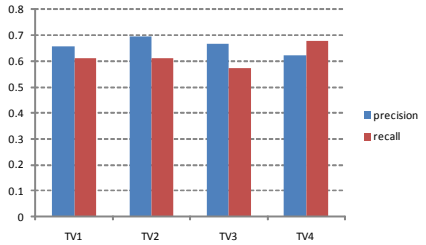

Figure 3. Performance of news story segmentation.

Table 2. Clustering performance based on different features.

	TV1	TV2	TV3	TV4	Average
BoW	0.44	0.58	0.72	0.89	0.66
BoT	0.24	0.38	0.67	0.83	0.53
BoW+BoT	0.44	0.58	0.80	0.91	0.68
BoW+BoT+T	0.46	0.63	0.77	0.92	0.70
BoW+BoT+T+SC	0.68	0.63	0.78	0.95	0.76

- News story clustering in single channels

We use F-measure defined in [11] for performance evaluation. Let $\mathcal{G}$ denote the ground truth and $\mathcal{D}$ denote the detected clusters. The F-measure F is

$$F = \frac{1}{H}\sum_{C_i \in \mathcal{G}} |C_i| \max_{C_j \in \mathcal{D}}\{f(C_i, C_j)\},$$
$$f(C_i, C_j) = \frac{2 \times p(C_i, C_j) \times r(C_i, C_j)}{p(C_i, C_j) + r(C_i, C_j)}, \quad (9)$$

where the precision $p(C_i, C_j) = |C_i \bigcap C_j|/C_j$ and the recall $r(C_i, C_j) = |C_i \bigcap C_j|/C_i$. The term $H = \sum_{C_i \in \mathcal{G}} |C_i|$ is used for normalization. Higher F means better cluster performance.

Table 2 shows clustering performance. The first two rows denote that only BoW ($\beta = \gamma = 0$ in eqn. (6)) and only BoT ($\alpha = \gamma = 0$) is considered, respectively. The first three rows just take the first term in eqn. (8) to measure similarity, while the last two rows further consider temporal information as in eqn. (8). The last row also considers semantic features. From the first two rows, we found that BoW works better than BoT, i.e. describing *what* are in videos should be given higher priority than describing *how* objects move. This result may come from that news videos contain relatively less content with significant motion. The third row shows that with appropriate combination, jointly considering *what* and *how* aspects achieves better performance. We obtain more performance improvement if temporal distance and semantics are incorporated into similarity measure. Comparing the last two rows, we see integrating semantic features effectively improve clustering performance. We have worse performance for TV1 and TV2, because many topic-related stories are extensively edited to have different lengths or interlaced with different content.

- News story clustering across channels

In evaluating performance of news story clustering across channels, temporal relation between news stories is not considered. Although topic-related news stories may be reported many times, some unimportant news stories may be reported only once or twice. Among the 334 news topics, 85 topics were reported once, and 142 topics were reported twice or less. According to the setting in [11], we call these stories outliers. In this experiment, we compare clustering performance based on the dataset with outliers (1st row in Table 3), based on the dataset in which the ones only reported once are filtered out (2nd row in Table 3), and based on the dataset in which the ones only reported less than twice are filtered out (3rd row in Table 3). We see better clustering performance if outliers are filtered out. Furthermore, semantic features just slightly improve story clustering across channels. The reason may be that we only detect 39 concepts (due to time cost consideration). Furthermore, although inaccurate semantic concepts work well for video retrieval at the shot level, how to utilize them to describe a news story needs further study.

Figure 4 shows examples of clustering results. In Figure 4(a), content of topic-related news stories is similar, and thus our method effectively clusters them together. In Figure 4(b), news stories of different topics are erroneously clustered together. This case reveals shortage of current features, which still cannot very accurately bridge the gap between visual characteristics and news topic (the semantic gap problem). We may also fail in the case of Figure 4(c). News stories of the same topic are not clustered together, because content of news stories in a cluster is completely different. This case sometimes happens if different channels have significantly different viewpoints to report this news. The same political event would be reported in opposite viewpoints in different channels. In the cases of Figure 4(b) and 4(c), further information such as speech or closed caption would be beneficial to story clustering, which will be studied in the future.

Table 3. Clustering performance across channels.

	BoT+BoW	BoT+BoW+SC
With outlier	0.36	0.37
Remove outlier 1	0.47	0.48
Remove outlier 2	0.46	0.47

- Relationship to near-duplicate detection

News video clustering sounds to be similar to near duplicate detection or video copy detection. However, we found that topic-related stories have visual variations over the assumptions of these two techniques. We randomly select ten clustering results for several times and evaluate the relationship between clustering accuracy and the ratio of near-duplicate keyframes. Working like a perfect near-duplicate detection module, we manually evaluate the number of near-duplicate keyframes n' in the same cluster and calculate the ratio as

$$r = \frac{n'}{\min(n_1, n_2, \ldots, n_k)}, \quad (10)$$

where n_i denotes the number of keyframe in news story i, and k stories are clustered as the same topic in this example.

Figure 5 shows the relationship. Firstly, near-duplicate ratios are varied at different channels, and at the same channel topic-related

stories may not have consistent properties (especially for TV3). This reflects various editing policies in different channels. Secondly, we implicitly see the trend that higher clustering performance can be obtained when more near-duplicate keyframes exist in the same cluster. Thirdly, even with perfect near-duplicate detection, news story clustering may still be difficult. The near-duplication ratios of TV1 and parts of TV3 are low, and the corresponding F-measure is below 0.5. This shows that news story clustering is a more general problem than near-duplicate detection.

Figure 4. (a) News stories of the same topic that are clustered together. (b) News stories of different topics that are clustered together. (c) News stories of the same topic that are not clustered together.

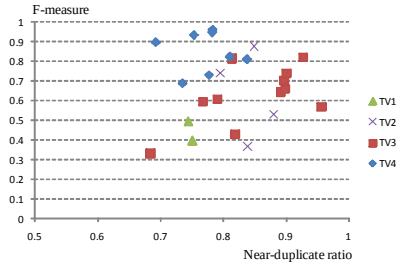

Figure 5. Relationship btw. F-measure and near-duplicate ratio.

6. CONCLUSION

We have presented a news clustering system that automatically segments news stories, represents stories in terms of visual words and trajectories, and then groups stories of the same topic. Based on the captured broadcasted news, we verify that fusing features from two aspects brings the best performance, while the bag of visual word model plays a much more important role than the bag of trajectory. In the future, we would evaluate the proposed framework based on large-scale datasets, such as TRECVID. Moreover, we would improve clustering news stories across different TV channels based on more features, such as more results of concept detection or crowdsourcing knowledge.

Acknowledgement

The work was partially supported by the National Science Council of Taiwan, Republic of China under research contract NSC 100-2221-E-194-061 and NSC 99-2221-E-194-036.

7. REFERENCES

[1] Lienhart, R. 1999. Comparison of automatic shot boundary detection algorithms. Proc. of SPIE Storage and Retrieval for Image and Video Databases VII, vol. 3656, pp. 290-301, 1999.

[2] Wang, F., Jiang, Y.-G., and Ngo, C.-W. 2008. Video event detection using motion relativity and visual relatedness. Proc. of ACM Multimedia, pp. 239-248.

[3] Thomee, B., Bakker, E.M., and Lew, M.S. 2010. TOP-SURF: a visual words toolkit. Proc. of ACM Multimedia.

[4] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. 2008. SURF: speeded up robust features. Computer Vision and Image Understanding, vol. 110, no. 3, pp. 346-359.

[5] Wu, X., Zhang, Y., Wu, Y., Guo, J., and Li, J. 2008. Invariant visual patterns for video copy detection. Proc. of ICPR.

[6] Pei, S.-C., and Chuang, W.H. 2005. Video analysis by means of major face determination. Proc. of IEEE International Midwest Symposium on Circuits and Systems, pp. 1095-1098.

[7] Rubner, Y., Tomasi, C., and Guibas, L.J. 2000. The earth mover's distance as a metric for image retrieval. International Journal on Computer Vision, vol. 40, no. 2, pp. 99-121.

[8] Xu, D., and Chang, S.-F. 2008. Video event recognition using kernel methods with multilevel temporal alignment. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 30, no. 11, pp. 1985-1997.

[9] Frey, B.J., and Dueck, D. 2007. Clustering by passing messages between data points. Science, vol. 315, no. 5814, pp. 972–976.

[10] Katayama, N., Mo, H., and Satoh, S. 2011. News shot cloud: ranking tv news shots by cross tv-channel filtering for efficient browsing of large-scale news video archives. LNCS 6523, pp. 284-295.

[11] Wu, X., Ngo, C.-W., and Hauptmann, A.G. 2008. Multimodal news story clustering with pairwise visual near-duplicate constraint. IEEE Trans. on Multimedia, vol. 10, no. 2, pp. 188-199.

[12] Yeh, J.-H., Chen, J.-C., Kuo, J.-H., and Wu, J.-L. 2005. TV commercial detection in news program videos. Proc. of IEEE International Symposium on Circuits and Systems, vol. 5, pp. 4594-4597.

[13] Zhong, D., Zhang, H.-J., and Chang, S.-F. 1996. Clustering methods for video browsing and annotation. Proc. of IS&T/SPIE Symposium on Electronic Imaging: Science and Technology - Storage and Retrieval for Image and Video Database.

[14] Ide, I., Mo, H., and Katayama, N. 2003. Threading news video topics. Proc. of ACM Workshop on Multimedia Information Retrieval.

[15] Nallapati, R., Feng, A., Peng, F., and Allan, J. 2004. Event threading within news topics. Proc. of ACM International Conference on Information and Knowledge Management.

[16] Hsu, W.H., and Chang, S.-F. 2006. Topic tracking across broadcast news videos with visual duplicates and semantic concepts. Proc. of IEEE International Conference on Image Processing, pp. 141-144.

[17] Wu, X., Ide, I., and Satoh, S. 2009. Large-scale news topic tracking and key-scene ranking with video near-duplicate constraints. Proc. of ACM Workshop on Large-Scale Multimedia Retrieval and Mining.

[18] Zhai, Y., and Shah, M. 2005. Tracking news stories across different sources. Proc. of ACM Multimedia, pp. 2-10.

[19] Jiang, Y.-G., Ngo, C.-W., and Yang. J. 2007. Towards optimal bag-of-features for object categorization and semantic video retrieval. Proc. ACM International Conference on Image and Video Retrieval.

[20] Naphade, M., Smith, J.R., Tesic, J., Chang, S.-F., Hsu, W., Kennedy, L., Hauptmann, A., and Curtis, J. 2006. Large-scale concept ontology for multimedia. IEEE Multimedia, vol. 13, no. 3, pp. 86-91.

[21] Beecks, C., Uysal, M.S., and Seidl, T. 2010. Signature quadratic distance. Proc. of ACM International Conference on Image and Video Retrieval, pp. 438-445.