NEWS STORY CLUSTERING WITH FISHER EMBEDDING

Wei-Ta Chu and Han-Nung Hsu

National Chung Cheng University, Chiayi, Taiwan

ABSTRACT

An automatic news story clustering system is presented to facilitate efficient news browsing and summarization. We describe news content by considering both what objects appear and how these objects move in news stories. With Fisher embedding, we respectively encode local features, semantics features, and dense trajectories as Fisher vectors, based on which similarity between news stories can be well evaluated and thus better clustering performance can be obtained. We verify the effectiveness of Fisher encoding, and further show that motion-based features are more effective than appearance-based features through feature analysis.

Index Terms— Fisher vector, news story clustering, what and how aspects

1. INTRODUCTION

In the last two decades, news video analysis has been studied widely from various perspectives. On the one hand, news videos are well structured and well edited to convey information efficiently. On the other hand, events, objects, or scenes that would appear in news videos are unpredictable, and complex visual content give rise to significant technical challenges. Nowadays large amounts of news stories are broadcasted 24 hours by many news TV channels, and efficient access of news stories are largely in demand. In this work, we focus on clustering news stories of the same topic together to facilitate efficient browsing and summarization. News topics, in the representation of various objects and events, and their evolvement, are to be described.

Wang et al. [1] showed that jointly considering object appearance and object motion yields more accurate video event detection. Motivated by this work, we attempt to describe news videos from both *what* and *how* aspects, i.e., describing *what objects or events appear in a news story*, and modeling *how these objects move or how these events evolve*. From the what aspect, one option to describe visual content is using the bag of word (BoW) model derived from local features points. Results of semantic concept detectors are another widely used representation to describe what appears in videos. From the how aspect, the most important information is object motion, which can be described by motion descriptors like motion histograms.

Various studies have been developed to represent videos based on the BoW approach. For example, Wu et al. [2] extracted motion trajectories and described them as a bag of trajectories for video copy detection. Chu et al. [3] integrated results obtained from bag of visual words, bag of semantics, and bag of trajectories to describe news stories. Recently, Fisher representation [4] has been proposed to improve the BoW approach. The Fisher kernel models distribution of features with respect to each visual word, rather than hard quantizing features into one of the visual words. Such representation demonstrates promising performance on image classification [4], and has been extended to capture temporal variations in videos [5]. In this work, we apply the Fisher representation to model news stories, and achieve news story clustering based on similarity calculated from multimodal Fisher representations.

Contributions of this work are twofold. First, we verify that embedding features by Fisher kernels really aids news story clustering. Second, we investigate impacts of different features when they are employed to cluster news stories.

2. RELATED WORKS

Several works have been proposed to cluster similar news video clips to facilitate efficient browsing. Based on available text information, Ide et al. [6] proposed a system to track and search news topics. Although news threads can be elaborately discovered, text information is not always available. Zhai and Shah [7] presented a semantic linking method to find similar news stories across sources. They considered both facial and none-facial keyframes, and discovered language correlation based on automatic speech recognition. Hsu and Chang [8] described news videos by visual features and semantic concepts, and developed a topic tracking system. Although promising performance was obtained, the number of news topics was limited and known in advance. Specific to the news story clustering problem, Wu et al. [9] treated news stories as the basic analysis units, and proposed a constraint-driven co-clustering algorithm to mine news topics. Static visual features, text, and near-duplicate constraints are jointly considered to cluster stories of the same topic together. In contrast to static visual features only, Chu et al. [3] described news stories with the bag of visual words and the bag of motion words. They jointly considered what objects are and how objects move in news stories, and then calculated similarities between news stories to facilitate news story clustering. However, recent studies show that encoding low-level features with the BoW approach is not robust enough [4]. Therefore, in this paper we attempt to exploit the Fisher encoding method to more appropriately characterize news stories.

3. FISHER ENCODING

We briefly introduce Fisher encoding in this section. The Fisher representation describes a feature as the gradient with respect to the probability density function built based on the training features. Generally, the density function is modeled by a Gaussian mixture model (GMM). Let μ_i and σ_i denote the mean and standard deviation of the i^{th} Gaussian mixture. Given a collection of d-dimensional features $X = \{x_1, x_2, ..., x_N\}$, the gradients of these features with respect to μ_i and σ_i are

$$\mathcal{G}^X_{\mu,i} = \frac{1}{N\sqrt{\omega_i}} \sum_{j=1}^{N} \gamma(i) \frac{x_j - \mu_i}{\sigma_i}, \quad (1)$$

Fig. 1. Snapshots of the evaluation dataset.

$$\mathcal{G}^X_{\sigma,i} = \frac{1}{N\sqrt{2\omega_i}} \sum_{j=1}^{N} \gamma(i) \Big[\frac{(x_j - \mu_i)^2}{\sigma_i^2} - 1 \Big]. \quad (2)$$

The value $\gamma(i) = \frac{\omega_i u_i(x_j)}{\sum_{k=1}^{K} \omega_k u_k(x_j)}$ is the soft assignment of a feature x_j to the i^{th} Gaussian, where K is the number of mixtures, and ω_i is the mixture weight of the i^{th} Gaussian to constitute the GMM. We can also calculate the gradient of a feature with respect to ω_i, but the experiments in [4] showed that it brings little additional information. Finally, the Fisher vector derived from the feature collection X is the concatenation of $\mathcal{G}^X_{\mu,i}$ and $\mathcal{G}^X_{\sigma,i}$, and has the dimensionality of $2Kd$ (K Gaussian mixtures, from each of which the obtained gradient is d-dimensional).

4. NEWS STORY CLUSTERING

4.1. Preprocessing

Given news videos that were continuously captured from news TV channels, we adopt the methods described in [3] to eliminate commercial breaks and segment news stories. The anchorperson shot in each news story is further removed to let us purely focus on news content. Errors in news segmentation were manually fixed so that input of the proposed work is new stories with accurate boundaries. For each news story, we extract one frame as the keyframe per fifteen frames. Figure 1 shows some snapshots of the evaluated news videos. We can see that there are scrolling text marquee on screen, which is often irrelevant to the news report. To reduce the influence of irrelevant text, we only consider the central part of a keyframe. Assume that the width and height of a video are W and H, respectively, the region with left-top corner at $(\frac{1}{5}W, \frac{1}{5}H)$ and bottom-right corner at $(\frac{4}{5}W, \frac{4}{5}H)$ is extracted, from which features described in the following sections are extracted. The dash-line box shown in the first image of Figure 1 shows the considered region.

4.2. Describe What Appears

We extract two types of features to describe what appears in videos: local feature points and semantic concepts. From each keyframe of a news story, we extract 64-dimensional SURF feature points [10]. We then reduce dimensionality of feature points to 32 by principal component analysis (PCA), in order to improve GMM clustering by decorrelation [5]. From the training data, 256,000 feature points are randomly selected to constitute the GMM consisting of 256 Gaussian mixtures. Given a collection of SURF features $X = \{x_1, ..., x_{N1}\}$ extracted from a news story S_i, they are encoded as a $2 \times 256 \times 32 = 16,384$-dimensional Fisher vector. Finally, we apply power and L2 normalization to the Fisher vector as in [4], obtaining the final representation $\boldsymbol{f}_p(S_i)$. The number of Gaussian mixtures directly affects the dimensionality of $\boldsymbol{f}_p(S_i)$ (256 in this case). It is determined by varying different settings and selecting the best one that yields the best performance based on our preliminary study. Settings for other types of Fisher vectors in the following are determined similarly.

To further describe what appears in videos in terms of semantics, we utilize the VIREO-374 concept detectors [11] to detect confidence scores of semantic concepts in each keyframe. To reduce computational cost, we only detect 39 of 374 concepts that are the same as LSCOM-lite [12], and thus a 39-dimensional semantic score vector is extracted for each keyframe. Similarly, we reduce its dimensionality to 20 by PCA, and totally 10,000 score vectors are randomly selected from the training data to constitute the GMM. Given a collection of score distribution $Y = \{y_1, ..., y_{N2}\}$ extracted from a news story S_i, they are encoded as a $2 \times 64 \times 20 = 2,560$-dimensional Fisher vector. After applying power and L2 normalization to the Fisher vector, the final representation $\boldsymbol{f}_t(S_i)$ is obtained.

4.3. Describe How to Evolve

From the how aspect, the major representation is derived from motion trajectories. We extract dense trajectories [13] between keyframes and describe them by 192-dimensional motion boundary histograms (MBH). By PCA, their dimensionality is reduced to 96. Totally 256,000 MBHs are randomly selected from the training data to constitute the GMM. Given a collection of MBHs $Z = \{z_1, ..., z_{N3}\}$ extracted from a news story S_i, Fisher embedding is applied to obtain a $2 \times 256 \times 96 = 49,152$-dimensional Fisher vector. After applying power and L2 normalization to the Fisher vector, the final representation $\boldsymbol{f}_m(S_i)$ is obtained.

4.4. Clustering

Based on features of visual appearance and motion, we calculate distances between news stories separately based on three types of Fisher vectors, and then integrate them to be the basis for clustering. More specifically, we calculate the distance $d_p(S_i, S_j)$ between stories S_i and S_j based on $\boldsymbol{f}_p(S_i)$ and $\boldsymbol{f}_p(S_j)$. Similarly, we can calculate distances $d_t(S_i, S_j)$ and $d_m(S_i, S_j)$ based on semantics and motion Fisher vectors ($\boldsymbol{f}_t$ and $\boldsymbol{f}_m$), respectively.

Because we may face the curse of dimensionality due to Fisher vector's high dimensionality, and the high computational cost when dealing with a large number of news stories, we apply PCA again to reduce the dimensionality of Fisher vectors, i.e., $\boldsymbol{f}_p$, $\boldsymbol{f}_t$, and $\boldsymbol{f}_m$, to 100, and calculate distances between reduced vectors. Finally, we integrate distances calculated based on three types of Fisher vectors, as well as consider a time factor [3], and define similarity between two stories S_i and S_j as

$$sim_{i,j} = e^{-D(i,j)} \times \begin{cases} \log_\Delta |t_j - t_i|, & \text{if } |t_j - t_i| < \Delta, \\ 1, & \text{otherwise,} \end{cases} \quad (3)$$

where $D(i,j)$ is the linear combination of three types of distances, i.e., $D(i,j) = \alpha d_p(i,j) + \beta d_t(i,j) + \gamma d_m(i,j)$. According to our preliminary study, the best performance can be obtained when weights α, β, and γ are set as 0.25, 0.375, and 0.5, respectively. This setting will be adopted in the evaluation section. The second term of eqn. (3) is a time factor specially designed to consider temporal distance between two stories in the same TV channel. The value t_i denotes that the story S_i is the t_i-th story from the beginning of the video. The logarithm to base Δ is monotonically increasing until Δ is reached. The number Δ is set according to the approximate period of topic-related news stories would repeat. Generally, a TV channel rarely repeats the same news in a short time period. Therefore, the

Table 1. Information of the evaluation dataset.

ID	Duration	#news stories	#topics	#video shots
TV1	8 hours	155	78	7529
TV2	8 hours	173	84	9028
TV3	10 hours	201	80	7898
TV4	10 hours	233	87	29088

Table 2. F-measure of news story clustering based on different distance measures.

Distance	TV1	TV2	TV3	TV4	Average
L1 distance	0.75	0.76	0.89	0.91	0.83
L2 distance	0.73	0.78	0.94	0.94	0.85

value $\log_\Delta |t_j - t_i|$ is larger if two stories are at a larger temporal distance.

With the similarity between any two stories, we apply the affinity propagation (AP) algorithm [14] to cluster news stories into several groups. News stories of the same topic are to be clustered together. The reason to use the AP algorithm is that it automatically determines the number of clusters, which is unknown before clustering and is dynamic for different TV channels.

5. EXPERIMENTS

We adopt the dataset provided by [3], which consists of totally 762 news videos covering 329 topics broadcasted from four news TV channels. Information of the evaluation dataset is shown in Table 1, and some snapshots from TV1 and TV2 are shown in Figure 1. In the following, we will conduct news story clustering within a single channel and across channels, respectively. Following [3], performance of clustering is measured by F-measure. Let $\mathcal{G}$ denote the ground truth and $\mathcal{D}$ the clustering result. The F-measure F is calculated as:

$$F = \frac{1}{Z} \sum_{C_i \in \mathcal{G}} |C_i| \max_{C_j \in \mathcal{D}} \{f(C_i, C_j)\}, \quad (4)$$

$$f(C_i, C_j) = \frac{2 \times p(C_i, C_j) \times r(C_i, C_j)}{p(C_i, C_j) + r(C_i, C_j)}, \quad (5)$$

where $p(C_i, C_j) = |C_i \cap C_j|/C_j$ is the precision value, and $r(C_i, C_j) = |C_i \cap C_j|/C_i$ is the recall value, respectively. The value $Z = \sum_{C_i \in \mathcal{G}} |C_i|$ is the normalization factor. Higher F means better clustering performance.

5.1. Clustering Results

Distance measurement. We first show clustering performance when we measure distance between Fisher vectors (combining both what and how aspects) by L1 norm and L2 norm, respectively. As shown in Table 2, comparing with L1 distance, L2 distance can improve the F-measure about 0.02, 0.05 and 0.03 in TV2, TV3, and TV4, respectively. On average, the F-measure is improved by 0.02 by using L2 distance, and thus we apply the L2 distance in the following experiments.

News story clustering in single channels. Table 3 shows news story clustering performance in single channels. Comparing with clustering results obtained based solely on SURF or MBH, integrating all types of Fisher vectors and the time factor yields the best performance (the average F-measure is 0.85). Combining information about what appear and how they evolve with the time factor advances clustering performance. This result verifies that "what aspect" and "how aspect" are complementary to each other.

Table 3. F-measure of news story clustering in single channels.

Method	TV1	TV2	TV3	TV4	Average
[3]	0.68	0.61	**0.95**	0.78	0.76
SURF-based FV	0.73	0.78	0.87	0.93	0.83
MBH-based FV	0.73	0.78	0.90	0.92	0.83
Semantics-based FV	0.57	0.74	0.83	0.77	0.73
All FVs + time	**0.73**	**0.78**	0.94	**0.94**	**0.85**

Table 4. F-measure of news story clustering based on the bag-of-word approach and the Fisher embedding, from the "what" aspect only.

	TV1	TV2	TV3	TV4	Average
Bag of word	0.58	0.44	0.89	0.72	0.66
Fisher embedding	0.73	0.78	0.87	0.93	0.83

Both [3] and our work integrate what and how aspects in describing news stories, while we utilize Fisher kernels to encode features. Results in Table 3 verifies that our approach outperforms [3], showing that with Fisher embedding we can more appropriately characterize complex news stories.

To further verify the improvement of Fisher embedding, in Table 4 we show performance comparison between the bag-of-word approach and Fisher embedding from the "what" aspect only, i.e., only SURF features are used to construct bag of words and Fisher vectors. From this table we more clearly see the significant superiority of Fisher embedding. The average F-measure is improved from 0.66 to 0.83.

News story clustering across channels. Here we evaluate performance of news story clustering across four channels. Unlike clustering in single channels, we do not consider the time factor as it makes no sense across channels. As shown in Table 5, comparing with [3], clustering performance is significantly improved by Fisher embedding. This again verifies that Fisher embedding is more robust to describe content of news stories and thus obtains performance gain. In this table, we also show the performance difference between considering the whole frame and only considering the central part mentioned in Sec. 4.1. Because there is much noisy scrolling text on screen, considering only the central part largely eliminates noisy features and yields much better performance.

5.2. Discussion

Comparing Table 5 with Table 3, we observe that there is still a big gap between clustering performance in single channels and across channels. The main reasons are due to significant variations of editing, viewpoint, and illumination variations across channels. We have investigated the effectiveness of Fisher encoding in the previous section, and now we give a pilot study to compare effectiveness of features derived from what and how aspects.

We especially select stories of the same topic, which were broadcasted by different channels or were also broadcasted multiple times in a single channel. Let $\mathcal{S}_1 = \{S_{1,1}, S_{1,2}, ..., S_{1,N_1}\}$ and $\mathcal{S}_2 = \{S_{2,1}, S_{2,2}, ..., S_{2,N_2}\}$ denote the set of news stories of the same topic broadcasted by TV1 and TV2, respectively. The story $S_{1,1}$ is the first story of a specific topic broadcasted by TV1. Separately

Table 5. F-measure of news story clustering across channels.

	[3]	Ours (whole)	Ours (central only)
F-measure	0.38	0.66	0.74

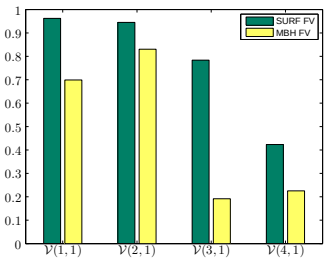

Fig. 2. Sample variations of distances between stories in the same channel and stories in different channels. This figure shows variations calculated based on the stories $S_{1,1}$, $S_{2,1}$, $S_{3,1}$, and $S_{4,1}$, respectively.

based on SURF and MBH Fisher vectors, we calculate distances between stories broadcasted by the same channel and across channels, respectively, and investigate how distances vary in two different cases. It can be expected that distances between stories broadcasted by the same channel would be smaller than that broadcasted by different channels. Let $d_{1,1}^{1,2}$ denote the L2 distance between $S_{1,1}$ and $S_{1,2}$ calculated based on SURF- or MBH-based Fisher vectors. Similarly, $d_{1,1}^{2,1}$ denotes the L2 distance between $S_{1,1}$ and $S_{2,1}$ that were broadcasted by TV1 and TV2, respectively. We put more focus on the average variation between distances obtained from stories in the same channel and that obtained across channels. Let the story $S_{1,1}$ as the base, the average variation specially considered is

$$\mathcal{V}(1,1) = \bar{d}_{1,1}^{q,r} - \bar{d}_{1,1}^{1,p}, \tag{6}$$

$$\bar{d}_{1,1}^{1,p} = \frac{1}{N_1 - 1} \sum_{S_{1,p} \in \mathcal{S}_1, p \neq 1} d_{1,1}^{1,p}, \tag{7}$$

$$\bar{d}_{1,1}^{q,r} = \frac{1}{Z'} \sum_{S_{q,r} \notin \mathcal{S}_1} d_{1,1}^{q,r}. \tag{8}$$

The value $\bar{d}_{1,1}^{1,p}$ is the average distance from the story $S_{1,1}$ to others also broadcasted by TV1 ($S_{1,p} \in \mathcal{S}_1$), while the value $\bar{d}_{1,1}^{q,r}$ is the average distance from the story $S_{1,1}$ to those broadcasted by other channels ($S_{q,r} \notin \mathcal{S}_1$). The value Z' is the number of stories over which the distance $d_{1,1}^{q,r}$ is computed. Note that these stories $S_{1,p}$ and $S_{q,r}$ are all of the same news topic.

Figure 2 shows sampled average variations when we use different stories as the bases. From this figure we can see that variations calculated based on MBH-based Fisher vectors are apparently smaller than that based on SURF-based Fisher vectors. This indicates that MBH is a relatively more robust feature that resists different post editings and broadcasting styles for the same news topic. This characteristic can be utilized in future study to develop more promising features or to constitute better feature combinations.

6. CONCLUSION

We have verified that describing visual content by Fisher vectors from both what and how aspects achieves promising performance on news story clustering. Comparing with bag-of-word models, encoding local features, semantic features, and dense trajectories with Fisher kernels provides significant improvement, due to more comprehensive representation of complex visual characteristics. Comparing features derived from what aspect with that from how aspect, we discuss robustness of different features and conjecture that features extracted from dense trajectories are more promising in news

Acknowledgement. The work was partially supported by the Ministry of Science and Technology in Taiwan under the grant MOST103-2221-E-194-027-MY3 and MOST104-2221-E-194-014.

7. REFERENCES

[1] Feng Wang, Yu-Gang Jiang, and Chong-Wah Ngo, "Video event detection using motion relativity and visual relatedness," in *Proc. of ACM Multimedia*, 2008, pp. 239–248.

[2] Xiao Wu, Yongdong Zhang, Yufeng Wu, and Jintao Li, "Invariant visual patterns for video copy detection," in *Proc. of International Conference on Pattern Recognition*, 2008.

[3] Wei-Ta Chu, Chao-Chin Huang, and Wen-Fang Cheng, "News story clustering from both what and how aspects: Using bag of word model and affinity propagation," in *Proc. of International ACM Workshop on Automated Media Analysis and Production for Novel TV Services*, 2011, pp. 7–12.

[4] Florent Perronnin, Jorge Sanchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. of European Conference on Computer Vision*, 2010, pp. 143–156.

[5] Ionut Mironica, Jasper Uijlings, Negar Rostamzadeh, Bogdan Ionescu, and Nicu Sebe, "Time matters! capturing variation in time in video using fisher kernels," in *Proc. of ACM Multimedia*, 2013, pp. 701–704.

[6] Ichiro Ide, Hiroshi Mo, and Norio Katayama, "Threading news video topics," in *Proc. of ACM Workshop on Multimedia Information Retrieval*, 2003, pp. 240–246.

[7] Yun Zhai and Mubarak Shah, "Tracking news stories across different sources," in *Proc. of ACM International Conference on Multimedia*, 2005, pp. 2–10.

[8] Winston Hsu and Shih-Fu Chang, "Topic tracking across broadcast news videos with visual duplicates and semantic concepts," in *Proc. of IEEE International Conference on Image Processing*, 2006, pp. 141–144.

[9] Xiao Wu, Chong-Wah Ngo, and Alexander G. Hauptmann, "Multimodal news story clustering with pairwise visual near-duplicate constraint," *IEEE Trans. on Multimedia*, vol. 10, no. 2, pp. 188–199, 2008.

[10] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[11] Yu-Gang Jiang, Jun Yang, Chong-Wah Ngo, and Alexander G. Hauptmann, "Representations of keypoint-based semantic concept detection: A comprehensive study," *IEEE Trans. on Multimedia*, vol. 12, no. 1, pp. 42–53, 2010.

[12] Milind Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis, "Large-scale concept ontology for multimedia," *IEEE Multimedia*, vol. 13, no. 3, pp. 86–91, 2006.

[13] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[14] Brendan J. Frey and Delbert Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.