NEWS STORY CLUSTERING WITH FISHER EMBEDDING

Wei-Ta Chu and Han-Nung Hsu
National Chung Cheng University, Taiwan
wtchu@ccu.edu.tw

Outline

- Introduction
- Fish Encoding
- News Story Clustering
  - Describe What Appears
  - Describe How to Evolve
  - Clustering
- Experiments
- Conclusion

Introduction

- Motivation: Large amounts of (redundant) news stories are broadcasted 24 hours a day
- Goal: Cluster news stories of the same topic together
- Challenges:
  - High visual variations
  - Rich semantics
  - Various motion information

Introduction

- Ideas: Describe news stories from both *what* and *how* aspects
  - *Describe what objects or event appear*
  - *Describe how these objects move or how these events evolve*
- What aspect
  - Bag of visual word (BoW)
  - Semantic concepts
- How aspect
  - Motion descriptors

Introduction

- Fisher representation improves the BoW approach
  - Model distribution of features with respect to each visual word, rather than hard quantization

- Contributions
  - Verify that embedding features by Fisher kernels aids news story clustering
  - Investigate impacts of different features

Fisher Encoding

- Fisher representation: describe a feature as the gradient with respect to the probability density function built based on the training data
- Density function modeled by a Gaussian mixture model with μ_i and σ_i
- Given a collection of d-dim features $X = \{x_1, x_2, \ldots, x_N\}$, the gradients with respect to μ_i and σ_i are

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

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

$\gamma(i) = \dfrac{\omega_i u_i(x_j)}{\sum_{k=1}^{K} \omega_k u_k(x_j)}$ is the soft assignment of a feature x_j to the i th Gaussian

Fisher Encoding

- The Fisher vector derived from the feature collection X is the concatenation of $\mathcal{G}_{\mu,i}^{X}$ and $\mathcal{G}_{\sigma,i}^{X}$, and has the dimensionality of $2Kd$ (K Gaussian mixtures, from each of which the obtained gradient is d-dimensional)

News Story Clustering

- Preprocessing
  - 1. Eliminate commercial breaks (based on shot change frequency)
  - 2. Eliminate anchorperson shot
  - 3. Keyframe selection (one out of fifteen frames)
  - 4. Only consider the central part of a keyframe to avoid noise

News Story Clustering

- Describe What Appears
  - Extract 64D SURF feature points, dimension reduction to 32D by PCA
  - 256,000 feature points constitute the GMM consisting of 256 Gaussian mixtures
  - Given a news story, the Fisher vector is $2 \times 256 \times 32 = 16,384$ D.

  - Detect 39 concepts from the VIREO-374 concept detectors, forming a 39D semantic score vector, dimension reduction to 20D by PCA
  - 10,000 score vectors constitute the GMM consisting of 64 Gaussian mixtures
  - Given a news story, the Fisher vector is $2 \times 64 \times 20 = 2,560$ D.

News Story Clustering

- Describe How to Evolve
  - Extract dense trajectories between keyframes, describe them by 192D motion boundary histograms (MBH), dimension reduction to 96D by PCA
  - 256,000 MBHs constitute the GMM consisting of 256 Gaussian mixtures
  - Given a news story, the Fisher vector is $2 \times 256 \times 96 = 49,152$ D.

News Story Clustering

- Clustering
  - Calculate distances between news stories separately based on three types of Fisher vectors, and integrate them to be basis for clustering
  - Apply PCA again to reduce dimensions of Fisher vectors $\boldsymbol{f}_p, \boldsymbol{f}_t, \boldsymbol{f}_m$ to 100.
  - The similarity between two stories S_i and S_j

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

$$D(i,j) = \alpha d_p(i,j) + \beta d_t(i,j) + \gamma d_m(i,j)$$

News Story Clustering

- Clustering
  - The second term is a time factor specially designed to consider temporal distance. The logarithm to base Δ is set according to the approximate period of topic-related news stories would repeat. The value $\log_\Delta |t_j - t_i|$ is larger if two stories are at a larger temporal distance.

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

$$D(i,j) = \alpha d_p(i,j) + \beta d_t(i,j) + \gamma d_m(i,j)$$

  - The affinity propagation (AP) algorithm is used to cluster news stories into groups.

Experiments

- Dataset: 762 news stories covering 329 topics

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

- Performance evaluation: F-measure

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

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

  - 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.

Experiments

- Distance measurement
  - Measure distance between Fisher vectors by L1 norm or L2 norm

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

  - News story clustering in single channels

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**

Experiments

- The improvement of Fisher embedding

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

- News story clustering across channels
  - Do not consider the time factor

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

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

Experiments

- Discussion
  - Big performance gap between single channels and across channels – significant variations of editing, viewpoint, and illumination across channels
  - Select stories of the same topic, broadcasted by different channels or by the same channel multiple times
  - Separately based on SURF and MBH Fisher vectors, we calculate distances between stories broadcasted by the same channel and across channels

Experiments

- Discussion
  - Put more focus on the average variation between two cases. Let the story $S_{1,1}$ as the base, the average variation is

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

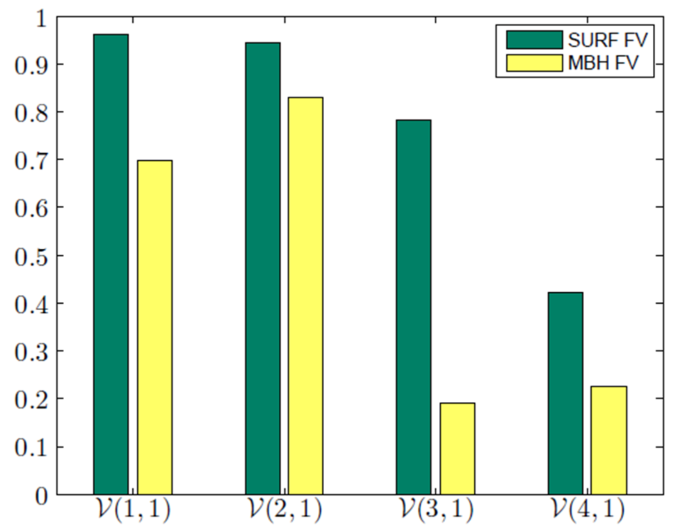
$$\bar{d}^{1,p}_{1,1} = \frac{1}{N_1 - 1} \sum_{S_{1,p} \in \mathcal{S}_1, p \neq 1} d^{1,p}_{1,1},$$

Average distance from the story to others broadcasted by the same channel

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

Average distance from the story to others broadcasted by other channels

  - Variations based on MGH-based Fisher vectors are apparently smaller than that based on SURF-based Fisher vector.
  - MBH is relatively more robust, more resisting visual variations across channels



Conclusion

- WE verify the effectiveness of Fisher representation from both what and how aspects in news story clustering.
  - Comparing with bag-of-words models
  - Combining local features, semantic features, and dense trajectory features
- Study robustness of different features

QUESTIONS?

Wei-Ta Chu

National Chung Cheng University, Taiwan

wtchu@ccu.edu.tw