Zhang CD, Wu X, Shyu ML *et al.* A novel web video event mining framework with the integration of correlation and co-occurrence information. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 28(5): 788–796 Sept. 2013. DOI 10.1007/s11390-013-1377-6

A Novel Web Video Event Mining Framework with the Integration of Correlation and Co-Occurrence Information

Cheng-De Zhang[1] (张承德), Xiao Wu[1,*] (吴 晓), *Member, ACM, IEEE*
Mei-Ling Shyu[2], *Senior Member, IEEE*, and Qiang Peng[1] (彭 强), *Member, ACM*

[1] *School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China*

[2] *Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33146, U.S.A.*

E-mail: chengde66@gmail.com; wuxiaohk@home.swjtu.edu.cn; shyu@miami.edu; qpeng@home.swjtu.edu.cn

Received May 5, 2013; revised August 10, 2013.

Abstract The massive web videos prompt an imperative demand on efficiently grasping the major events. However, the distinct characteristics of web videos, such as the limited number of features, the noisy text information, and the unavoidable error in near-duplicate keyframes (NDKs) detection, make web video event mining a challenging task. In this paper, we propose a novel four-stage framework to improve the performance of web video event mining. Data preprocessing is the first stage. Multiple Correspondence Analysis (MCA) is then applied to explore the correlation between terms and classes, targeting for bridging the gap between NDKs and high-level semantic concepts. Next, co-occurrence information is used to detect the similarity between NDKs and classes using the NDK-within-video information. Finally, both of them are integrated for web video event mining through negative NDK pruning and positive NDK enhancement. Moreover, both NDKs and terms with relatively low frequencies are treated as useful information in our experiments. Experimental results on large-scale web videos from YouTube demonstrate that the proposed framework outperforms several existing mining methods and obtains good results for web video event mining.

Keywords web video event mining, multiple correspondence analysis, co-occurrence, near-duplicate keyframe

1 Introduction

Due to the increasing popularity of Internet, the users are able to easily get a large number of relevant web videos about ongoing incidents or events through search engines or video sharing websites, such as Google, Baidu, YouTube and Youku. In addition, news videos are published on newswires in digital versions like CNN, BBC, and CCTV. In January 2009, 14.8 billion online videos were watched by 147 million U.S. Internet users[1]. However, it is still a challenge for the users to grasp the major events after a glance at the search results, not to mention the development of the topic.

Generally speaking, when searching a topic, most of the users first want to know the major events and then to construct their relationship in minds. However, they have to click on the videos one by one to try to manually summarize the videos after watching all of them at the same time. It is not only highly time consuming but also difficult for users to find out what they want, especially for the unfamiliar topics. Therefore, it has become crucial to be able to automatically group relevant videos together.

To group relevant videos, both visual and textual information are commonly utilized. For visual information, some important shots are frequently inserted into videos or reports as a reminder or support of viewpoints, which carry useful information. If these duplicate shots/keyframes are clustered to form different groups according to visual content, where each group is called an NDK (near-duplicate keyframe) group, it would play the same role as a hot term in the text field. Therefore, NDKs can be used to group videos with similar contents of the same event. However, since web videos are usually short and uploaded by general

Regular Paper

This work was supported by the National Natural Science Foundation of China under Grant Nos. 61373121, 61071184, 60972111, 61036008, the Research Funds for the Doctoral Program of Higher Education of China under Grant No. 20100184120009, the Program for Sichuan Provincial Science Fund for Distinguished Young Scholars under Grant Nos. 2012JQ0029, 13QNJJ0149, the Fundamental Research Funds for the Central Universities of China under Grant Nos. SWJTU09CX032, SWJTU10CX08, and the Program of China Scholarships Council under Grant No. 201207000050.

[*] Corresponding Author

users, lighting conditions, editing operations, and poor quality may lead to many relevant keyframes being split into different segments. To solve this problem, co-occurrence is used to group relevant NDKs together with NDK-within-video information. On the other hand, for the textual information, web videos generally have much less text information (such as titles and tags) than documents. In addition, even with titles and tags, many of the texts are noisy, ambiguous, and sometimes incomplete. Users could also add irrelevant hot terms (words) to attract attention. As a result, the textual information derived from the limited titles and tags from the web videos may get worse performance than traditional documents. It is believed that the integration of both visual and textual information can complement each other to improve the performance of event mining.

In this paper, a novel framework is proposed that integrates visual and textual information for web video event mining. First, Multiple Correspondence Analysis (MCA) is explored to extract NDK-level event similarity with the assistance of textual information. Second, co-occurrence is applied to detect the similarity between NDKs and events using the NDK-within-video information. Finally, the information from these two steps is integrated for web video event mining by noisy NDK pruning and positive NDK weight enhancement.

The main novelty and contributions of this paper are summarized as follows:

1) A novel framework is proposed that integrates textual and visual information to improve the performance of web video event mining. First, common negative NDKs, whose weights calculated by both visual and textual information are negative, are pruned. Second, the weights of other NDKs' are enhanced by combining them together.

2) We apply a statistical method MCA to NDK-level web video event mining to explore the correlation between terms and classes, which bridges the gap between NDKs and high-level semantic concepts.

3) NDK-within-video information calculated by co-occurrence is proposed to measure the similarity between NDKs and events. On the one hand, it can group NDK fragments together to enhance NDK detection. On the other hand, it can help enhance the weights of both positive and negative NDKs calculated by noisy textual information.

The rest of this paper is organized as follows. Section 2 gives a brief overview of related work. The details of the framework for web video event mining are provided in Sections 3. The experiments and results are presented in Section 4. Finally, Section 5 concludes this work.

2 Related Work

2.1 Topic Detection and Tracking

Web video event mining belongs to the task of topic detection and tracking (TDT) whose goal is to detect new topics and track the known events in text news streams. Many studies in TDT have been done in text areas[2-6]. For example, bursty relationship was applied to detect bursty events in [7]. The concept of near-duplicate keyframe (NDK) has been abundantly used in real applications[8-9]. An NDK is a set of similar keyframes with certain variations induced by acquisition times, lighting conditions, and editing operations[10]. NDK detection is to calculate the keyframe similarity among videos. The retrieval of NDKs[10-12] plays an important role in measuring video clip similarity and tracking video shots of multi-lingual sources[13]. With the assistance of NDK constraints, news stories are clustered into topics by constraint-based co-clustering[14].

Lately, TDT research has been extended to integrate both textual and visual information. TDT research has also been carried out in multimedia area. For example, topics are tracked with visual duplicates and semantic concepts[15-20]. A trajectory-based approach presented in [21] is used to discover, track, monitor, and visualize web video topics. With textual correlation and keyframe matching, topic clusters are grouped in [22] and news stories from different TV channels are linked in [23]. Topic discovery is deployed by constructing the duality between stories and textual-visual concepts through bipartite graph[13,24]. Events are discovered by text co-occurrence and visual feature trajectory in [25]. Recent research about organizing and utilizing large collections or databases of images and videos is summarized in [26]. A hierarchical framework for event organization is presented in [27], where higher level events are defined upon more detectable low-level events.

2.2 Multiple Correspondence Analysis

Correspondence Analysis (CA) is an exploratory/descriptive data analytic technique designed to analyze simple 2-way and multi-way tables containing some measure of correspondence between the rows and columns[28]. Multiple Correspondence Analysis (MCA) is considered as an extension of CA to more than two variables[29]. MCA has been proved to be able to capture the correlation among nominal variables effectively[29].

Tags inherently have nominal representations, automatically generating the image representation for the landmarks by first using tags and location metadata to detect tags and locations that represent the landmarks, and then applying visual analysis to the images associated with the detected tags and locations[30], which has been successfully applied to content-based video concept detection[28,31-32]. Therefore, the correlation between each tag and an NDK is suitable to be measured by MCA.

Motivated by the functionality of MCA, we explore the utilization of MCA to analyze data instances (i.e., NDKs in this study) described by a set of textual features to capture the correspondence between terms and classes. Moreover, while co-occurrence has been commonly adopted in the textual field, it has not been fully utilized in the visual part of NDKs. Compared with noisy terms which may cross a long range, NDKs in events generally appear in a relatively small range and are less noisy. This motivates the study of using NDK-within-video information to improve the performance of event classification in this paper.

3 Web Video Event Mining

3.1 Proposed Framework

The proposed framework is illustrated in Fig.1. As can be seen from this figure, this framework consists of four stages (enclosed in the dashed rectangular boxes), namely data preprocessing, NDK-within-video information mining, correlation information mining, and integration.

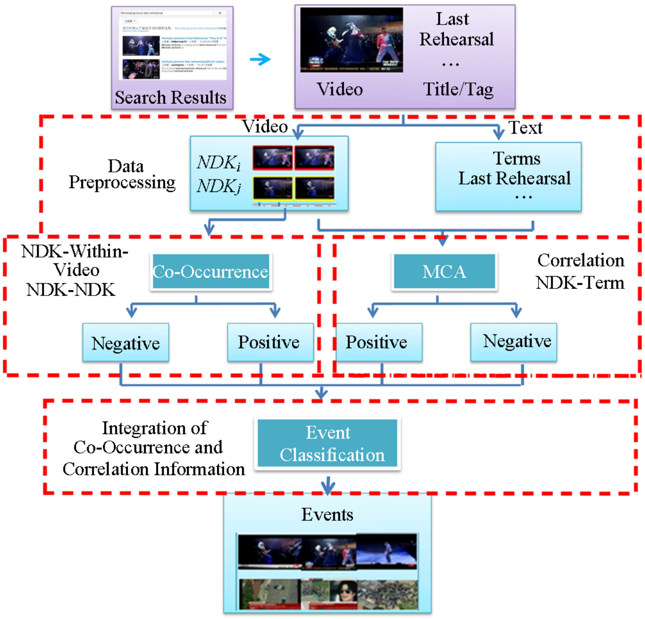


Fig.1. Proposed web video event mining framework.

The first stage is data preprocessing, where NDKs are extracted. As NDKs have the unique property of identifying similar events, all NDKs are considered as the features. More details about NDK detection are introduced in Subsection 3.1. Next, terms extracted from titles and tags are treated as textual features. Due to noisy user-supplied tag information, text words are pruned by word stemming, special character removal, and so on. Finally, all features and the class labels are combined to form an indicator matrix with instances (i.e., NDKs) as rows and the categories of variables (i.e., terms) and class labels as columns.

In correlation information mining stage, MCA is applied to calculate MCA-based transaction weights, targeting for bridging the gap between an NDK and a term. Feature-value pairs[16] are generated for each term, and the similarity between each feature-value pair and a class is calculated. Finally, the weights between each NDK and all classes are calculated by summing the weights of the feature-value pairs along all the features.

In NDK-within-video information mining stage, NDK-within-video information is used to measure the similarities among NDKs to address the problems caused by false NDK detection, video edition, and fragment problems. Finally, this relationship is used to measure the similarity between an NDK and an event.

In the integration stage, after weight pruning and enhancement by combination, every NDK is grouped to the event with the largest similarity value. Actually, NDK-within-video information is used to enhance positive weights and prune negative weights of the NDKs.

3.2 NDK Extraction

An NDK group is a set of keyframes which have similar visual content. Generally, series of web videos are returned from a user query. First, shot boundary detection and keyframe extraction are used to get keyframes, and then each video can be represented with these keyframes. Second, an efficient NDK detection approach is adopted to detect NDKs among videos. In order to guarantee good performance of NDK detection between videos, local points are detected with Harris-Laplace and are described by SIFT[33]. The public available tool proposed in [34] is adopted to detect the NDKs. Third, the detected keyframes with similar visual content are further grouped to form clusters by transitive closure. For example, if $NDK_1 \rightarrow NDK_3$, and $NDK_3 \rightarrow NDK_4$, then we assume that there exists the association $NDK_1 \rightarrow NDK_4$. Ultimately, we get grouped keyframes, where each group represents one visual scene.

After getting NDKs, (1) is used to calculate the probability of an NDK belonging to each event.

$$P(NDK_i, E_j) = \frac{|NDK_i \bigcap E_j|}{|NDK_i|}, \qquad (1)$$

where $|NDK_i \bigcap E_j|$ is the number of common videos between NDK_i and event E_j. $|NDK_i|$ is the number of videos whose keyframes are in NDK_i. If NDK_i contains keyframe K_s, while K_s is the s-th keyframe of video V_t, then V_t is contained by NDK_i too. E_j is the j-th event according to ground truth. At last, each NDK is marked as the class label, which has the biggest probability with this NDK.

3.3 Correlation Information Mining

MCA is an extension of standard correspondence analysis to more than two variables[29]. The correlation among more than two variables in a table can be cached by it. In this study, MCA is adopted to measure the correlations between terms and classes. Thus, a table is represented in a 2-dimensional (2D) NDK-term matrix $\boldsymbol{TG}$ where each element $TG_{i,j}$ in $\boldsymbol{TG}$ is defined in (2).

$$TG_{i,j} = \begin{cases} 1, & \text{if the } j\text{-th term is contained in the} \\ & i\text{-th NDK}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Finally, all features are combined with the class labels C^j to form an indicator matrix with instances (i.e., NDKs) as rows and features (i.e., categories of variables or terms) as columns. Thus, MCA can be be directly applied to calculate the correlation of each feature-value pair of each term with a class. An example of a training dataset after discretization is shown in Table 1. Each of the features is discretized into several feature-value pairs according to the method in Weka[①]. Assuming the s-th feature has j_s feature-value pairs, the number of classes and instances are n and m, respectively, while each instance represents one NDK. And then the indicator matrix is denoted by $\boldsymbol{Z}$ with size $m \times (j_s + n)$. The meaning of the indicator matrix is using the relationship between feature-value pairs and classes to calculate the relationship between a feature and a class. Instead of performing on the indicator matrix, MCA analyzes the inner product of this indicator matrix, i.e., $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$, called the Burt table, which is symmetric with size $(j_s + n) \times (j_s + n)$. Next, singular value decomposition (SVD) is applied to the correspondence matrix, which is transformed from the Burt matrix by centering and standardizing. More details of the MCA technique and the way to generate such correlations can be found in [32]. Now the feature-value pairs and classes can be projected into a 2D space constructed by the first and second principal components.

Table 1. Training Dataset After Discretization

Feature 1	Feature 2	Feature 3	$\cdots$	Feature m
F_1^1	F_2^1	F_3^1	$\cdots$	F_m^1
F_1^2	F_2^2	F_3^2	$\cdots$	F_m^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

The graphical representation of MCA, called the symmetric map, can be used to visualize the feature-value pairs of a feature and the classes as points in a map with the dimensions depending on the number of principle components. Thus, the correlation between a feature-value pair and a class can be measured by the cosine value of the angle between the two vectors representing the feature-value and the class. Fig.2 shows a 2D symmetric map in which there are a term feature F_i with four feature-value pairs F_i^1, F_i^2, F_i^3 and F_i^4, and two classes C^1 (positive class) and C^2 (negative class). $angle_i^1$ is the angle between feature-value pair F_i^1 and the class C^1. If a feature-value pair is correlated with a class, then the other feature-value pair is negatively correlated with this class to the same amount. A large absolute value of the cosine value indicates a strong correlation between a feature-value pair and a class. Therefore, MCA can be applied to calculate the similarity between each feature-value pair F_i^j and class label C^n. The weight $W_{i,j}^n$ can be calculated using (3).

$$W_{i,j}^n = \cos(angle_i^j), \tag{3}$$

where $angle_i^j$ is the angle between the feature-value pair F_i^j and the class label C^n. If the angle is smaller than 90 degrees, it means that F_i^j has a high relationship with C^n. The cosine value of the angle between the feature-value pair and the class can be taken as a weight, which represents the discriminant capability. A transaction weight between NDK_k and class C^n can be calculated by summing the weights of the feature-value pairs for all features as shown in (4). Positive data NDKs are supposed to have larger transaction weights compared with the negative ones, because a feature-value pair with a larger weight indicates a stronger correlation with the class, compared with a smaller one.

$$TW_{k,n} = \frac{1}{m}\sum_{i=1}^{m} W_{i,j}^n. \tag{4}$$

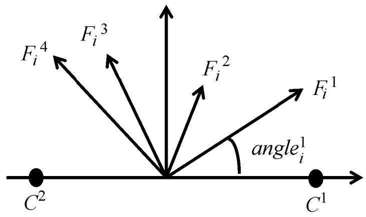


Fig.2. Geometrical representation of MCA.

[①] Weka. http://www.cs.waikato.ac.nz/ml/weka/, Dec. 2012.

Finally, the weights between each NDK and all classes can be obtained, where the weights are sorted in the descending order. The gaps between neighbors in this sorted list are calculated, and the largest gap is used as the threshold to decide the positive or negative results. If $TW_{k,n}$ is greater than the threshold, it means that the similarity between NDK_k and C^n is higher than the largest gap. Then, NDK_k will be taken as a positive instance; otherwise, it will be marked as negative.

3.4 NDK-Within-Video Information Mining

After getting the similarities between each NDK and all classes (events), video content information is added to compensate the defects of textual information. This is due to the facts that there are usually fewer textual features for web videos compared with text documents, and these features are often noisy, ambiguous, and incomplete. Important visual shots are frequently inserted into related videos as a reminder or support of viewpoints, acting as hot terms in the text field. These NDKs usually carry useful video content information and can be used to group videos of similar themes into events. However, due to video editing or NDK detection errors, many NDKs are split into different segments. There are numerous frequently accompanied NDKs which convey useful information. Such NDK-within-video-information can be utilized to identify more related web videos. Therefore, NDK similarity is applied to improve the performance by pruning the noise and enhancing the weight generated by the initially retrieved NDKs. The rationale is that the more common videos they contain, the more similar they are. Let NDK_i and NDK_j be two NDKs, $|NDK_i \bigcap NDK_j|$ be the number of overlapped videos, and $\min(NDK_i, NDK_j)$ be the minimum number of videos contained in NDK_i and NDK_j. To capture the NDK-within-video-information, the Jaccard coefficient is adopted to measure the similarity between NDK_i and NDK_j as shown in (5). A higher score denotes that these two NDKs are more likely correlated.

$$d(NDK_i, NDK_j) = \frac{|NDK_i \bigcap NDK_j|}{\min(NDK_i, NDK_j)}. \quad (5)$$

Similar to the previous stage, the similarities are sorted in the descending order. The gaps between neighbors in this sorted list are calculated. The largest gap is used as the threshold to decide whether NDK_k is positive or negative. The weight of NDK_k is calculated using (6).

$$W_{V_{NDK_k}} = \begin{cases} 1, & \text{if } Sim_V(NDK_k, E_j) > \text{threshold}, \\ 0, & \text{otherwise}, \end{cases} \quad (6)$$

where $Sim_V(NDK_k, E_j)$ is the similarity between NDK_k and E_j and is calculated using (7). If it is larger than the threshold, then NDK_k will be assigned as positive; otherwise, it will be marked as negative. Finally, all NDKs are considered to calculate the similarity between each NDK and E_j, which is defined in (7). Here, m is the number of NDKs that meet the condition of $d(NDK_k, NDK_p) > 0$, and s is the number of NDKs contained in E_n.

$$Sim(NDK_k, E_j) = \frac{1}{s}\sum_{p=1}^{m} d(NDK_k, NDK_p). \quad (7)$$

3.5 Event Mining

After correlation information mining and NDK-within-video information mining, each NDK gets a label of positive or negative from both stages. If an NDK is marked as negative in both stages, it is considered as the noise and needs to be pruned. Otherwise, the weights generated by both stages are integrated to enhance the similarity between NDK_k and E_n using (8).

$$Sim(NDK_k, E_n) = \gamma \times W_{T_{NDK_k}} \times TW_{k,n} + (1-\gamma)\times W_{V_{NDK_k}} \times Sim_V(NDK_k, E_n). \quad (8)$$

Here, γ is a factor to control which can be adjusted according to the different weights of visual and textual information. In this study, we assume that they have equal weight. $W_{T_{NDK_k}}$ and $W_{V_{NDK_k}}$ are the positive or negative status of NDK_k in the textual and visual parts, respectively. $TW_{k,n}$ and $Sim_V(NDK_k, E_n)$ are the similarities between NDK_k and E_n, calculated by MCA and co-occurrence, respectively. NDK_k is assigned to the event which has the largest similarity value.

4 Experiment

4.1 Dataset

We use the dataset in [25] to evaluate the performance of the proposed work. The original dataset consists of 19 972 web videos and 803 346 keyframes. These videos were collected from YouTube and most videos were collected in May 2009. The topics were selected based on the top 10 news that happened during 2006 to 2009 as recommended by CNN, TIMES and Xinhua. We selected 15 hot topics shown in Table 2, which basically cover different characteristics of topics for experimental evaluation. For example, topic "Virginia tech massacre" only spanned for one month from April to May of 2007, while the topic "Iran nuclear program" lasted for several years from 2006 to 2009. However, the topic “California wildfires” happens periodically for several years, while the content is relatively homogeneous. To ensure fairness, those events having less than five web videos are regarded as noise and pruned out. Therefore, 10 815 videos, 37 055 NDKs, and 41 461 terms are used as the dataset in the experiments. The detailed information of the dataset is listed in Table 2. Each topic is composed of several events. Taking topic “Iran nuclear program” as an example, it contains 5 events: “Enriched uranium successfully”, “International community response”, “Inspections”, “Reports from IAEA”, “USA’S attitude to Iran”. For the ground truth, we collected the data by searching each topic from Wikipedia and Google, and then manually decided the ground truth.

4.2 Performance Evaluation

The precision (P), recall (R), and F1 measure ($F1$) are used to evaluate the performance of event mining, as defined in (9). Here, B_i^+ is the number of correctly grouped positive videos for cluster A_i, and B_i is the number of positive samples in the ground truth.

$$P=\frac{|B_i^+|}{|A_i|},\quad R=\frac{|B_i^+|}{|B_i|},\quad F1=\frac{2\times P\times R}{P+R}. \tag{9}$$

To evaluate the performance of web video event mining, simplified versions of the proposed methods in [5, 7, 25] (FT$_{\rm T}$, CC$_{\rm V}$, FT+CC) and MCA are used as the baselines. The performance comparison is presented in Table 3. Due to the tradeoff between precision and

Table 2. Dataset Information

ID	Topic	Number of Videos	Number of NDKs	Number of Terms	Number of Events
1	US President Election	737	1 826	3 327	13
2	Mumbai terror attack	423	1 741	1 569	5
3	Russia Georgia war	749	2 823	2 316	7
4	Somali pirates	410	1 405	2 178	5
5	Virginia tech massacre	683	1 865	1 621	2
6	Israel attacks Gaza	802	3 087	3 546	4
7	Beijing Olympic torch relay	652	2 448	1 949	12
8	California wildfires	426	1 631	3 025	6
9	Oil price	759	2 486	3 814	5
10	Myanmar cyclone	613	2 698	1 624	4
11	Kosovo independence	524	9 69	1 593	5
12	Russian president election	1 335	3 930	4 684	6
13	Iran nuclear program	1 056	4 561	3 969	5
14	Israeli Palestine peace	586	3 184	2 275	9
15	Korea nuclear	1 060	2 401	3 971	13
Total		10 815	37 055	41 461	101

Table 3. Performance Comparison

Topic	FT$_{\rm T}$[5]			CC$_{\rm V}$[7]			FT + CC[25]			MCA			MCA + CC		
	P	R	$F1$	P	R	$F1$	P	R	$F1$	P	R	$F1$	P	R	$F1$
1	0.26	0.39	0.32	**0.90**	0.15	0.20	0.57	0.35	0.44	0.11	**0.72**	0.18	0.44	0.57	**0.49**
2	0.31	0.14	0.20	**0.88**	0.12	0.15	0.49	0.19	0.28	0.12	0.24	0.14	0.58	**0.30**	**0.40**
3	0.58	0.11	0.19	**0.91**	0.04	0.06	0.72	0.15	0.25	0.37	0.17	0.24	0.64	**0.57**	**0.60**
4	0.49	0.21	0.30	**0.87**	0.05	0.07	0.48	0.25	0.33	0.28	0.25	0.27	0.44	**0.53**	**0.48**
5	0.76	0.05	0.10	**0.99**	0.02	0.03	0.73	0.33	0.46	0.36	0.40	0.38	0.71	**0.56**	**0.63**
6	0.45	0.12	0.20	**0.95**	0.02	0.03	0.54	0.16	0.25	0.21	0.24	0.23	0.64	**0.32**	**0.43**
7	0.52	**0.41**	**0.46**	**0.94**	0.09	0.12	0.52	0.20	0.29	0.10	0.14	0.12	0.48	0.32	0.39
8	0.46	0.12	0.19	**0.95**	0.05	0.07	0.68	0.18	0.29	0.23	0.24	0.24	0.50	**0.30**	**0.38**
9	0.22	0.10	0.14	**0.80**	0.08	0.10	0.58	0.13	0.22	0.04	0.20	0.07	0.62	**0.52**	**0.56**
10	0.39	0.05	0.09	**0.85**	0.05	0.05	0.68	0.34	0.46	0.22	0.27	0.25	0.68	**0.37**	**0.48**
11	0.66	0.07	0.13	**0.99**	0.02	0.03	0.78	0.09	0.17	0.37	0.15	0.22	0.91	**0.30**	**0.45**
12	0.27	0.14	0.16	**0.92**	0.02	0.03	0.61	0.14	0.23	0.04	0.13	0.07	0.71	**0.57**	**0.63**
13	0.60	0.07	0.13	**0.98**	0.02	0.03	0.83	0.10	0.18	0.13	0.17	0.15	0.86	**0.32**	**0.47**
14	0.35	0.17	0.24	**0.92**	0.04	0.07	0.51	0.16	0.25	0.20	**0.83**	0.32	0.62	0.26	**0.36**
15	0.34	0.16	0.22	**0.89**	0.07	0.09	0.46	0.24	0.32	0.24	**0.52**	0.32	0.50	0.41	**0.45**
Average	0.43	0.16	0.20	**0.92**	0.06	0.08	0.62	0.20	0.30	0.21	0.32	0.22	0.44	**0.46**	**0.42**

Note: The best precision, recall and $F1$ for each topic are highlighted in bold.

recall, $F1$ considers both precision and recall, so we focus on the $F1$ measure. It can be easily seen that our proposed framework (MCA+CC) outperforms baseline methods, with 12%∼34% improvement, which is significant.

Overall, it can be observed in Table 3 that our proposed framework achieves promising results compared with all the baseline methods. We can further observe that the recall value and the $F1$-scores for the proposed framework are almost always higher than all the other classifiers. This encouraging observation demonstrates the fact that the proposed framework has the ability to help event mining without missing too many web videos.

For the method of text feature trajectories in [5], word feature trajectories are first extracted, in which each feature is defined as a normalized *df-idf* score. Highly correlated word features are grouped to events by mapping word sets to video sets. Overall, we can see that the performance of this method is poor. It is a challenging mission to detect events from web video search results under the noisy and diverse social web scenario. It indicates that text information alone is not sufficient for event mining.

The method of [7] is taken as baseline 2. It is applied to NDK-level clustering, where both high and low frequency NDKs with close relationships are grouped together. As shown in Table 3, the precision is very high as visual content information has less noise than textual information. Moreover, since NDK detection among videos just has a relationship within the videos with similar content. Therefore, each group of NDK has little connection with the other kind of visual content, so the average of precision is as high as 92%.

For the method of combination of text co-occurrence and visual near-duplicate feature trajectory in [25], the performance of this method is better than [5] and [7]. However, the performance is still not good enough. The reason is that it misses those low-frequency terms and NDKs, which contain a large number of videos. Visual feature trajectory of NDK is not a consistent mining. We have tried to use term co-occurrence to compensate defects for visual information. However, frequent pattern still misses too much information for events.

Targeting for bridging the gap between NDK and high-level semantic concepts, MCA is explored to capture the correspondence between terms and classes, through analyzing the distribution of the terms appeared in each NDK. While low-frequency terms are considered as useful information, noise is an unavoidable problem. Moreover, multi-language, synonyms and the number of videos contained by NDK are still problems. Therefore, the precision of MCA is lower than the other methods. However, statistical properties make MCA express more stable features than the other methods, so the recall values are higher than those of the other baseline methods.

Finally, a novel framework is proposed by the integration of correlation and visual content information. From Table 3, it can be easily seen that the $F1$ values have been significantly improved. More encouraging, the best results can even reach 63%. This can be inferred from the results that a set of representative visual shots are often accompanied with an important event. Therefore, the NDKs are good cues for grouping related videos into events. On the contrary, texts/terms are relatively general, broad, and noisy. Even though low-frequent NDKs and terms may bring in more noisy information, the integration of both co-occurrence and correlation information can complement each other. Ultimately, clustering with textual and low-frequent information could bring more related videos. That is, our proposed framework can group more positive information without misclassifying too many negative web videos.

5 Conclusions

When facing a strapping number of web videos returned from search engines, it is a painstaking task to explore the search list to find out the major event. Due to the unique characteristics of web video scenarios, such as the limited number of features, the unavoidable errors in NDK detection, and the noisy text information, web video event mining has become a challenging task. In this paper, we proposed a novel 4-stage web video event mining framework, which integrates textual and visual information, and aims to improve the performance of web video event mining. Multiple correspondence analysis (MCA) is applied to explore the correlation between different terms and classes to bridge the gap between the extracted NDK and high-level semantic concepts. Moreover, both textual and visual features with relatively low frequencies are considered as useful information in our experiments. However, the videos without overlapped NDK cannot be included into any event. While encouraging, we can see that the result is still unsatisfactory. Apart from the limited and noisy textual and visual information of web videos, we will try to resort to news website for more useful information as the future work.

Acknowledgment Cheng-De Zhang gratefully acknowledges the opportunity of the Exchange Visiting Scholar Program from the University of Miami. Special thanks go to Dian-Ting Liu and Qiu-Sha Zhu for their great help in this study.

References

[1] Zhang J, Fan X, Wang J *et al.* Keyword-propagation-based information enriching and noise removal for web news videos. In *Proc. the 18th ACM International Conference on Knowledge Discovery and Data Mining*, Aug. 2012, pp.561-569.

[2] Chen K Y, Luesukprasert L, Chou S *et al.* Hot topic extraction based on timeline analysis and multidimensional sentence modeling. *IEEE Transactions on Knowledge and Data Engineering*, 2007, 19(8): 1016-1025.

[3] Fung G P C , Yu J X, Liu H *et al.* Time-dependent event hierarchy construction. In *Proc. the 13th Int. Conf. Knowledge Discovery and Data Mining*, Aug. 2007, pp.300-309.

[4] Fung G P C, Yu J X, Yu P S *et al.* Parameter free bursty events detection in text streams. In *Proc. the 31st Int. Conf. Very Large Data Bases*, Aug. 2005, pp.181-192.

[5] He Q, Chang K, Lim E P. Analyzing feature trajectories for event detection. In *Proc. the 30th ACM Int. Conf. Research and Develop. in Inform. Retrieval*, Aug. 2007, pp.207-214.

[6] Wang X, Zhai C, Hu X *et al.* Mining correlated bursty topic patterns from coordinated text streams. In *Proc. the 13th ACM International Conference on Knowledge Discovery and Data Mining*, Aug. 2007, pp.784-793.

[7] Yao J, Cui B, Huang Y *et al.* Bursty event detection from collaborative tags. *World Wide Web*, 2012, 15(2): 171-195.

[8] Tan S, Tan H K, Ngo C W. Topical summarization of web videos by visual-text time-dependent alignment. In *Proc. the ACM Int. Conf. Multimedia*, Oct. 2010, pp.1095-1098.

[9] Wu X, Zhao W L, Ngo C W. Near-duplicate keyframe retrieval with visual keywords and semantic context. In *Proc. the 6th ACM International Conference on Image and Video Retrieval*, July 2007, pp.162-169.

[10] Ke Y, Sukthankar R, Huston L. Efficient near-duplicate detection and sub-image retrieval. In *Proc. the ACM Int. Conf. Multimedia*, 2004, Vol.4, pp.869-876.

[11] Ngo C W, Zhao W L, Jiang Y G. Fast tracking of near-duplicate keyframes in broadcast domain with transitivity propagation. In *Proc. the 14th ACM International Conference on Multimedia*, Oct. 2006, pp.845-854.

[12] Zhang D Q, Chang S F. Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In *Proc. the 12th ACM International Conference on Multimedia*, Oct. 2004, pp.877-884.

[13] Wu X, Ngo C W, Hauptmann A G. Multimodal news story clustering with pairwise visual near-duplicate constraint. *IEEE Transactions on Multimedia*, 2008, 10(2): 188-199.

[14] Wu X, Ngo C W, Li Q. Threading and autodocumenting news videos: A promising solution to rapidly browse news topics. *IEEE Signal Processing Magazine*, 2006, 23(2): 59-68.

[15] Martinez-Gil J, Aldana-Montes J. KnoE: A web mining tool to validate previously discovered semantic correspondences. *Journal of Computer Science and Technology*, 2012, 27(6): 1222-1232.

[16] Lu B, Wang G R, Yuan Y. A novel approach towards large scale cross-media retrieval. *Journal of Computer Science and Technology*, 2012, 27(6): 1140-1149.

[17] Feng B L, Cao J, Bao X G *et al.* Graph-based multi-space semantic correlation propagation for video retrieval. *The Visual Computer*, 2011, 27(1): 21-34.

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

[19] Liu D T, Shyu M L, Chen C *et al.* Within and between shot information utilisation in video key frame extraction. *Journal of Information & Knowledge Management*, 2011, 10(3): 247-259.

[20] Meng T, Shyu M L. Leveraging concept association network for multimedia rare concept mining and retrieval. In *Proc. the 2012 IEEE International Conference on Multimedia & Expo*, July 2012, pp.860-865.

[21] Cao J, Ngo C W, Zhang Y D *et al.* Tracking web video topics: Discovery, visualization, and monitoring. *IEEE Trans. Circuits and Systems for Video Technology*, 2011, 21(12): 1835-1846.

[22] Duygulu P, Pan J Y, Forsyth D A. Towards autodocumentary: Tracking the evolution of news stories. In *Proc. the 12th ACM Int. Conf. Multimedia*, Oct. 2004, pp.820-827.

[23] Zhai Y, Shah M. Tracking news stories across different sources. In *Proc. the 13th ACM International Conference on Multimedia*, Nov. 2005, pp.2-10.

[24] Liu L, Sun L, Rui Y *et al.* Web video topic discovery and tracking via bipartite graph reinforcement model. In *Proc. of the 17th ACM International Conference on World Wide Web*, Apr. 2008, pp.1009-1018.

[25] Wu X, Lu Y J, Peng Q *et al.* Mining event structures from web videos. *IEEE Multimedia*, 2011, 18(1): 38-51.

[26] Hu S M, Chen T, Xu K *et al.* Internet visual media processing: A survey with graphics and vision applications. *The Visual Computer*, 2013, 29(5): 393-405.

[27] Parry M L, Legg P A, Chung D H *et al.* Hierarchical event selection for video storyboards with a case study on snooker video visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2011, 17(12): 1747-1756.

[28] Lin L, Ravitz G, Shyu M L *et al.* Correlation-based video semantic concept detection using multiple correspondence analysis. In *Proc. the 10th IEEE International Symposium on Multimedia*, Dec. 2008, pp.316-321.

[29] Salkind N J. Encyclopedia of Measurement and Statistics. SAGA Publications, Inc., 2006.

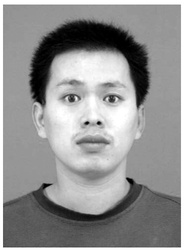
[30] Kennedy L S, Naaman M. Generating diverse and representative image search results for landmarks. In *Proc. the 17th ACM International Conference on World Wide Web*, Apr. 2008, pp.297-306.

[31] Zhu Q S, Lin L, Shyu M L *et al.* Utilizing context information to enhance content-based image classification. *International Journal of Multimedia Data Engineering and Management*, 2011, 2(3): 34-51.

[32] Lin L, Chen C, Shyu M L *et al.* Weighted subspace filtering and ranking algorithms for video concept retrieval. *IEEE Multimedia*, 2011, 18(3): 32-43.

[33] Lowe D G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, 60(2): 91-110.

[34] Zhao W L, Wu X, Ngo C W. On the annotation of web videos by efficient near-duplicate search. *IEEE Transactions on Multimedia*, 2010, 12(5): 448-461.



Cheng-De Zhang received the B.Eng degree in computer science and technology from Xiaogan University, Xiaogan, in 2006, the M.S. degree in computer application and technology from Xihua University, Chengdu, in 2009. He is currently a Ph.D. candidate of Southwest Jiaotong University, Chengdu. From 2012 to 2013, he is with the Department of Electrical and Computer Engineering, University of Miami (UM), USA, as a visiting scholar. His research interests include multimedia information retrieval, data mining, image processing and pattern recognition.

Xiao Wu received the B.Eng. and M.S. degrees in computer science from Yunnan University, Kunming, and the Ph.D. degree in computer science from the Department of Computer Science of City University of Hong Kong, in 2008. Currently, he is an associate professor and the department head of the Department of Computer Science and Technology, School of Information Science and Technology, Southwest Jiaotong University, Chengdu. He was a research assistant and a senior research associate at the City University of Hong Kong from 2003 to 2004, and 2007 to 2009, respectively. From 2006 to 2007, he was with the School of Computer Science, Carnegie Mellon University, Pittsburgh, USA, as a visiting scholar. He was with the Institute of Software, Chinese Academy of Sciences, Beijing, from 2001 to 2002. His research interests include multimedia information retrieval, video computing, and data mining.

Mei-Ling Shyu is a full professor at the Department of Electrical and Computer Engineering (ECE), University of Miami (UM) since June 2013. Prior to that, she was an associate/assistant professor in ECE at UM from January 2000. She received her Ph.D. degree in electrical and computer engineering from the School of Electrical and Computer Engineering and three Master degrees, all from Purdue University, West Lafayette, USA. Her research interests include multimedia data mining, management & retrieval, and security. She has authored and co-authored more than 210 technical papers. Dr. Shyu was awarded the 2012 Computer Society Technical Achievement Award and the ACM 2012 Distinguished Scientists Award. She received the Best Paper Award from the IEEE International Conference on Information Reuse and Integration in 2012, the Best Published Journal Article in IJMDEM for 2010 Award, the Best Student Paper Award with her student from the 3rd IEEE International Conference on Semantic Computing in 2009. She serves/served as an associate editor for several journals including IEEE Transactions on Human-Machine Systems, IEEE Transactions on SMC — Part C: Applications and Reviews, and on the editorial board of many other journals.

Qiang Peng received the B.E. degree in automation control from Xi'an Jiaotong University, the M.Eng degree in computer application and technology, and the Ph.D. degree in traffic information and control engineering from Southwest Jiaotong University, Chengdu, in 1984, 1987, and 2004, respectively. He is currently a professor at the School of Information Science and Technology, Southwest Jiaotong University. He has been in charge of more than 10 national scientific projects, published over 70 papers and holds 10 Chinese patents. His research interests include digital video compression and transmission, image/graphics processing, traffic information detection and simulation, virtual reality technology, multimedia system and application.