MMSS: Multi-modal Story-oriented Video Summarization[*]

Jia-Yu Pan, Hyungjeong Yang,[†] Christos Faloutsos
Computer Science Department
Carnegie Mellon University
{jypan, hjyang, christos}@cs.cmu.edu

Abstract

We propose multi-modal story-oriented video summarization (MMSS) which, unlike previous works that use fine-tuned, domain-specific heuristics, provides a domain-independent, graph-based framework. MMSS uncovers correlation between information of different modalities which gives meaningful story-oriented news video summaries. MMSS can also be applied for video retrieval, giving performance that matches the best traditional retrieval techniques (OKAPI and LSI), with no fine-tuned heuristics such as tf/idf.

1. Introduction and related works

As more and more video libraries [9] become available, video summarization is in great demands for enabling users to efficiently access these video collections. Most previous work focuses on summarizing an *entire* video clip into a more compact movie to facilitate browsing and content-based retrieval [8, 4]. For story-oriented summarization, research has been done mainly under the context of multi-document summarization [3] in the textual domain. Little work has been done on story-oriented video summarization using the multi-modal information.

Identifying footages of the same evolving story is difficult. Broadcast news production commonly shows a small icon beside an anchorperson to represent the story on which the anchorperson is reporting at the time [1]. The same icon is usually reused later in the shots about the follow-up development of the story, as an aid for the viewers to link current coverage to past coverage. We call these icons "*news-logos*", and the associated stories *logo stories*. The property of logos makes them a robust feature for linking separated footages of a story.

In this paper, we propose a method, *MMSS*, to generate multi-modal summary of a logo story. *MMSS* integrates multi-modal (visual/textual) information, treating it in a uniform, modality-independent fashion, with no need of parameter tuning. In fact, *MMSS* uncovers cross-modal correlation which, not only gives good story summaries, but also video retrieval performance matches the best finely tuned traditional information retrieval techniques.

The paper is organized as follows. Section 2 introduces the proposed method, *MMSS*. Sections 3 presents our experimental results on two applications, namely, story-oriented video summarization and video retrieval. Section 4 concludes the paper.

2. Proposed method: Video mining with *MMSS*

MMSS introduces a general framework for mining the cross-modal correlations among data of different modalities (frames/terms/logos) in video clips. The found cross-modal correlations are then used for story-oriented summarization and video retrieval.

The data set we used in this work is the TRECVID 2003 [7] data set. The data set is a collection of news programs. Each news program is broken into shots, each of which is associated with a keyframe and transcript words. For the words, we keep only the nouns and filter out the stop words.

In our experiments, logos are identified and extracted from the shot keyframes, using off the shelf algorithms for iconic matching [1, 2]. Figure 1 shows the keyframes and the associated words of three logos in the CNN news from the TRECVID 2003 data set.

Observation 1 *Logos provide robust visual hints and help alleviate the problems of tracking shots of a same story.*

Our goal is to exploit the logos, to facilitate video mining tasks. Particularly, we focus on the following two applications:

[*] Supported by the NSF under Grants No. IIS-0121641, IIS-0083148, IIS-0113089, IIS-0209107, IIS-0205224, INT-0318547, SENSOR-0329549, EF-0331657, IIS-0326322, and by the Pennsylvania Infrastructure Technology Alliance (PITA) Grant No. 22-901-0001.

[†] Supported by the Post-doctoral Fellowship Program of Korea Science and Engineering Foundation (KOSEF).

winter olympics nagano ceremony night round competition result headline news superticker medal watch headline news superticker result

(a) "Winter Olympics"

money trie source lawyer house intern monica information counsel starr immunity learned starr

(b) "Lewinsky"

arab league secretarygeneral strike iraq reaction arab brent sadler report iraqi president saddam hussein kind sentiment attack

(c) "Iraq"

Figure 1. News logos

- (Story summarization) How do we generate high-quality textual and visual summaries of a story?
- (Video retrieval) How can we exploit the logos, to retrieve the video clips that are relevant to a text query?

In addition, we want to perform the above two tasks in a principled way, that is, using the same framework for both tasks, integrating all multi-modal sources easily, with no parameter tuning.

Graph G_{MMSS} We integrate the information of shot-word co-occurrence with the logo information into a graph G_{MMSS}. The graph G_{MMSS} is a three-layer graph with 3 types of nodes and 2 types of edges. The 3 types of nodes are *logo-node*, *frame-node* and *term-node*, corresponding to the logos, keyframes (each representing a shot), and terms, respectively. The 2 types of edges are the *term-occurrence edge* and the "*same-logo*" edge. Figure 2 shows an example graph G_{MMSS} with 2 logo-nodes $\{l_1, l_2\}$, 5 frame-nodes $\{f_1, \ldots, f_5\}$, and 10 term-nodes $\{t_1, \ldots, t_{10}\}$. The term-occurrence edges are the solid lines, and the "same-logo" edges are the dotted lines.

A logo-node l_i is connected to a frame-node f_j by a "same-logo" edge, if the logo $O(l_i)$ appears in the frame $O(f_j)$. A frame-node f_j is connected to a term-node t_k by a term-occurrence edge, if the term $O(t_k)$ occurs in the shot whose keyframe is $O(f_j)$.

For logo story summarization and video retrieval, the essential part they share is to select objects pertaining to a query object. In logo story summarization, frames and terms forming the summary are selected based on their "relevance" to the query object, the logo(-node) of the story. As for video retrieval, we rank and select video shots by their "relevance" to the set of query terms. With the graph G_{MMSS}, we can turn the problem of computing "relevance" with respect to the query objects, into a random walk on the graph G_{MMSS}, as we show next.

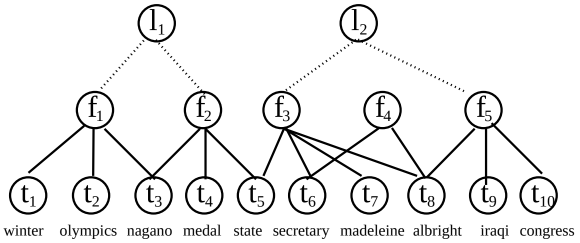

Figure 2. (The MMSS graph G_{MMSS}) Three types of nodes, and two types of edges: logo-nodes l_i's, frame-nodes f_i's and term-nodes t_i's; "same-logo" edges (dotted) and the term-occurrence edges (solid).

Random walk with restarts (RWR) In this work, we propose to use *random walk with restarts* ("*RWR*") for estimating the *relevance* of node "v" with respect to the restart node "s". The "random walk with restarts" operates as follows: to compute the relevance of node "v" for node "s", consider a random walker that starts from node "s". At every time-tick, the walker chooses randomly among the available edges, with one modification: before he makes a choice, he goes back to node "s" with probability c. Let $u_s(v)$ denote the stationary probability that our random walker will find himself at node "v". Then, $u_s(v)$ is what we want, the relevance of "v" with respect to "s", and we call it the *RWR score* of "v" (with respect to "s"). The intuition is that if the random walker who restarts from s (with probability c) has high chance of finding himself at node v, then node v is close and relevant to s. Details about RWR can be found in [5].

To use RWR to summarize a logo story $O(l_i)$, we set the restart node s be the logo-node s=l_i. The frame(-node)s and term(-node)s with the highest RWR scores are then selected as the story summary. Similarly, for video retrieval, the restart nodes are set to the term-nodes corresponding to the query terms. The query result is the shots (frame-nodes) with the highest RWR scores.

3. Experimental Results

The experiments are designed to answer the following questions: (a) For visual story summarization, how good are the shots that *MMSS* chooses? (b) For text story summarization, how good are the terms that *MMSS* chooses? (c) For video retrieval by text query, how well does *MMSS* compare to successful existing text retrieval methods, like OKAPI and LSI?

We should emphasize that OKAPI and LSI can only answer queries of the form "given a query word, find relevant video shots”. Our *MMSS* method, being modality-independent, can answer any type of query, like “given a shot (without a logo), find the best logo for it”; or “given a logo, find the best shots and/or terms for it”.

In our experiments, we follow the guidelines from [6] and set the restart probability c=0.65 for our 3-layer G_{MMSS} graph.

3.1 Story summarization

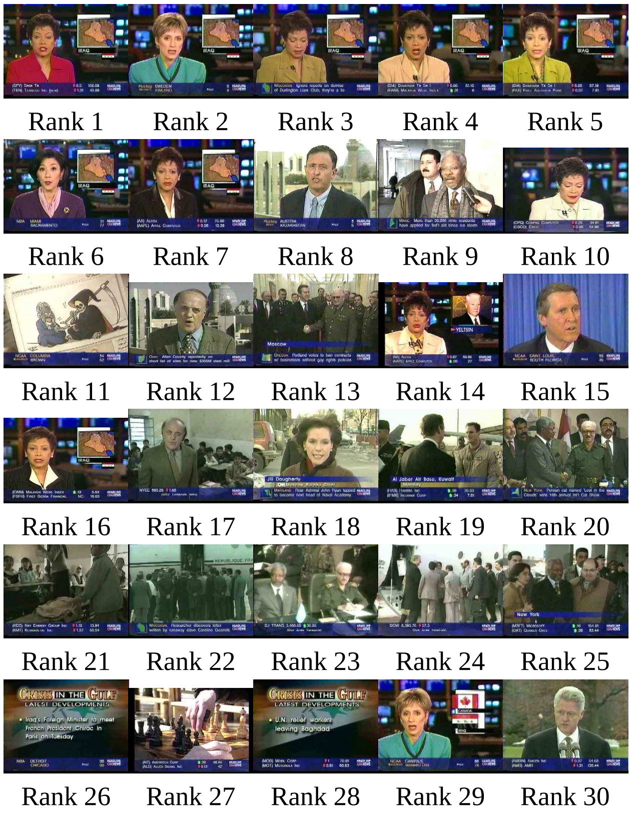

Figure 3. (Visual summary of logo “Iraq”) Frames are sorted (highest score first).

MMSS summarizes news-logo stories using the frames and texts which have high RWR score. Figures 3 shows the top 30 frames selected by *MMSS* for the logo “Iraq”. The top 7 frames are the frames of logo “Iraq” detected by the iconic matching. These frames are ranked high, simply because they are connected directly to the restart logo-node. Interestingly, *MMSS* found extra logo frames (e.g. the logo frame ranked 16) missed by the iconic matching. *MMSS* selects informative frames about the logo story, where faces of the major players are easily seen. For example, Kofi Annan appears in the rank 9 frame. Frames which contain important information in the form of *overlaid text* are also selected, as shown in the frames ranked 26-th and 28-th - the “Crisis in the Gulf”- on which current developments are summarized. We emphasize that the information of overlaid text is important and may not be available to the textual retrieval methods, for they are rarely fully mentioned by the anchorperson and are not in the transcript. Other logos pertaining to the logo “Iraq” are also detected and selected, for example, the “Yeltsin” logo at rank 14 and the “Canada-Iraq” logo at rank 29.

Observation 2 *(Visual summary by MMSS) MMSS summarizes logo stories by selecting relevant frames from the news video collections. Specifically, MMSS selects frames (a) of persons, objects, activities which are significant to the story; (b) of meaningful overlaid text; (c) which contain the “seed” logos but are missed by the “iconic matching” technique; (d) of other relevant logos.*

Table 1 shows the terms selected by *MMSS* for summarizing three logo stories in Figure 1, namely “Winter Olympics”, “Lewinsky” and “Iraq”. Together with the selected frames in Figure 3, we found that *MMSS* successfully select meaningful frames and terms for the logo stories. *MMSS* also picks meaningful frames for the logo stories “Winter Olympics” and “Lewinsky”, but the selected frames are not shown due to the page limit. Detail results can be found in [5].

Story	Summarizing terms
“Winter Olympics”	winter medal gold state skier headline news result superticker olympics competition nagano ceremony watch night round game team sport weather photo woman that today canada bronze year home storm coverage
“Lewinsky”	house lawyer intern ginsburg starr bill whitewater counsel immunity president clinton monica source information money trie learned iraq today state agreement country client weapon force nation inspection courthouse germany support
“Iraq”	iraq minister annan kofi effort baghdad report president arab strike defense sudan iraqi today weapon secretary talk school window problem there desk peter student system damage apart arnett albright secretarygeneral

Table 1. (Textual summary by *MMSS*) Terms are sorted (highest score first).

3.2 Video retrieval

In the task of video retrieval, we are given a query (a set of terms), the goal is to retrieve shots which are most relevant to the query. In other words, we want to rank all shots according to their “closeness” to the set of query words. The queries used in our experiments are: {‘‘lewinsky’’},

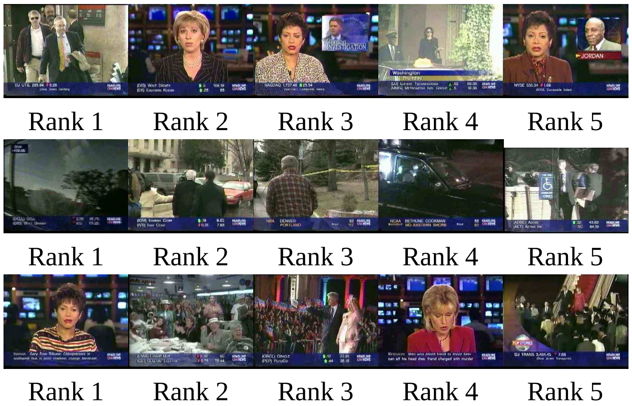

Figure 4. Keyframes of the top 5 shots retrieved by *MMSS* (top row), OKAPI (middle row) and LSI (bottom row) on query {‘‘lewinsky’’, ‘‘clinton’’}. Frames are ranked (highest score first).

{‘‘clinton’’}, {‘‘lewinsky’’, ‘‘clinton’’}, {‘‘annan’’}, {‘‘iraq’’}, {‘‘annan’’, ‘‘iraq’’}, {‘‘olympics’’}, {‘‘white’’, ‘‘house’’, ‘‘scandal’’}.

Since the data set we use does not have ground truth for any query, we do not report the standard precision and recall measures. Instead, we inspect the result by human judgment. We leave the precision/recall experiments to the future works.

We notice that a shot which contains keywords to a query is not necessarily a shot with meaningful content about the query. For example, a “teaser” in the beginning of a news broadcast introduces all headline news and is full of keywords. However, a teaser is usually accompanied with the anchor shots and does not have meaningful scene shots. Traditional textual retrieval methods are likely to retrieve teaser-style shots. On the other hand, *MMSS* is unbiased to the teasers, as we show next.

MMSS successfully ranks relevant shots to the top of the list, as shown in Figure 4. The frontal view of the major players related to the query is at the top of the list, for example, Starr at rank 1 and Monica at rank 4. In addition, *MMSS* avoids the news “teasers” while OKAPI and LSI rank the teaser shots with high scores. For example, in Figure 4, the rank 1 shot chosen by OKAPI (middle row) and the rank 5 shot chosen by LSI (bottom row) are both teaser shots.

Observation 3 *(OKAPI and LSI are biased to teaser shots) Textual retrieval methods such as OKAPI and LSI prefer teaser shots, for example, the “headlines preview” at the beginning of news programs, due to the many keywords the news anchors mentioned in those shots. Unfortunately, these teasers are not major shots of story content.*

4. Conclusions

We propose *MMSS* for story-oriented multi-modal video summarization and cross-modality correlation discovery. *MMSS* encodes both the textual and scene information, as well as logos which link shots of a story, as a graph. The random work with restarts (RWR) stationary probability is used to obtain a story-specific relevance ranking among the terms and shot keyframes. We report experiments on the TRECVID 2003 data set, for two applications, namely, story-oriented summarization and video retrieval. Our experiments show that *MMSS* is effective and gives meaningful summaries. Moreover, *MMSS* matches the best textual information retrieval methods on video retrieval; in fact, it sometimes does better, because it avoids the news “teasers” (Observation 3). Unlike the textual retrieval methods, *MMSS* achieves these with no sophisticated parameter tuning, and no domain knowledge.

References

[1] P. Duygulu, J.-Y. Pan, and D. A. Forsyth. Towards auto-documentary: Tracking the evolution of news stories. In *Proceedings of the ACM Multimedia Conference*, October 2004.

[2] J. Edwards, R. White, and D. Forsyth. Words and pictures in the news. In *HLT-NAACL03 Workshop on Learning Word Meaning from Non-Linguistic Data*, May 2003.

[3] J. Goldstein, V. O. Mittal, J. Carbonell, and J. Callan. Creating and evaluating multi-document sentence extract summaries. In *Proceedings of the Ninth International Conference on Information Knowledge Management (CIKM-00)*, November 2000.

[4] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang. Automatic video summarization by graph modeling. In *Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003)*, 2003.

[5] J.-Y. Pan, H.-J. Yang, and C. Faloutsos. MMSS: Graph-based multi-modal story-oriented video summarization and retrieval. Technical report, CMU-CS-04-114, Carnegie Mellon University, 2004.

[6] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. In *Proceedings of the 10th ACM SIGKDD Conference*, August 2004.

[7] A. F. Smeaton, W. Kraaij, and P. Over. TRECVID 2003 - an introduction. In *Proceedings of TREC 2003*, 2003.

[8] M. A. Smith and T. Kanade. Video skimming and characterization through the combination of image and language understanding techniques. In *Proceedings of CVPR 1997*, pages 775–781, June 17-19 1997.

[9] H. Wactlar, M. Christel, Y. Gong, and A. Hauptmann. Lessons learned from the creation and deployment of a terabyte digital video library. *IEEE Computer*, 32(2):66–73, February 1999.