

Integrating Semantic Video Understanding and Knowledge Visualization for Large-Scale News Video Exploration

Hangzai Luo^a, Jianping Fan^a, Shin'ichi Satoh^b, Jing Yang^a and William Ribarsky^a

^aDepartment of Computer Science
UNC-Charlotte, Charlotte, NC 28223
Tel: 704-687-8538
{hluo, jfan, jyang13, ribarsky}@uncc.edu

^bNational Institute of Informatics
Tokyo 101-8430, Japan
satoh@nii.ac.jp

Abstract

In this paper, we have developed a novel framework to enable more effective visual analysis and exploration of large-scale news videos via knowledge visualization. A novel interestingness measurement for video news reports is proposed to enable analysts and general audiences to find news stories of interest at first glance and catch the valuable knowledge in large-scale video news databases. Keyframes, keywords and their relations are automatically extracted from news video clips and visually represented according to their interestingness measurement. Our techniques for intelligent news video analysis have the capacity to enable more effective visualization and exploration of large-scale news videos. Our visualization-based news video analysis and exploration system is very useful for analysts and general audiences to quickly find the news stories of interest from large-scale news videos extracted from many channels.

Index Terms

Semantic Video Classification, Video Visualization, Knowledge Visualization.

I. INTRODUCTION

Broadcast video news is a very rich and immediate source of breaking news. It provides a picture of what is happening now at the local, national, and international levels. It offers different perspectives

depending on whether it is a local news broadcast talking about local events, a broadcast in one country talking about what is happening in another country, a news channel broadcast, and so on. Broadcast news provides not only reports on events but insight into the social and political framework from which the broadcast originates. For these reasons, broadcast news is watched and closely analyzed by individuals, government organizations, and companies. However, with the rapidly increasing number of broadcasts, especially in developing countries, the fraction that can be successfully watched in detail or even monitored by any individual or entity is growing ever smaller.

In news analysis, for example, it is becoming untenable to hire people to manually process all available news videos and produce summarizations from them. Manual analysis of large-scale news video reports is too expensive, and it may take too long. There is an urgent demand for achieving automatic exploration of large-scale video news databases. However, automatic video news analysis still suffers from the following challenging problems.

The *first problem* is how to *extract the underlying semantics* from the video news clips. Before the system can provide automatic video news exploration, it must be able to understand the underlying semantics of the input video clips. However, there exists a big **semantic gap** [1] between the low-level visual features and the high-level semantic video concepts. Existing video exploration systems can only support the services based on low-level visual features. The users, on the other hand, can only express their information needs via high-level semantics and concepts [2]. Semantic video classification approaches can extract limited video semantics from video clips. However, they cannot satisfy the requirements of video news database exploration applications for the video semantics. How to provide intuitive video news database exploration via the limited video semantics is still an open problem.

The *second problem* is how to *extract most useful knowledge* from the large-scale video news database. The total amount of information for a large-scale video news database is very large. Most of the information is irrelevant. If all information is delivered to an analyst or audience, they may easily get lost and miss the important information. For example, “Bush is the president of US” is a piece of well-known information. Disclosing this information to an analyst does not make sense. And most general audiences may not be interested in such kind of information. Only the abnormal information (e.g. the **new knowledge**) is useful and interesting for the users. There is an **interest gap** [3] between the underlying information collection and the users’ interest.

Visualization approaches have been proposed to help the users intuitively explore information space and find interesting parts. In-spire [4] transforms the text document collection to a spatial representation for visualization and analysis. Statistical information of news reports [5] is put on a world map to inform

the audiences of the “hotness” of regions and the relations among the regions. TimeMine [6] is able to detect the most important reports and organize them through timeline with statistical models of word use. Another system, called newsmap [7], organizes news topics from Google news on a two dimensional rectangle, where each news story covers a visualization space that is proportional to the number of related news pages reported by Google. News titles are drawn in the corresponding visualization space allocated to them. ThemeRiver [8] and ThemeView [9] can visualize a large collection of documents with keywords or themes. The distribution structure of the themes and keywords on the database can be intuitively represented to the users by ThemeRiver and ThemeView. However, all of these visualization systems cannot directly provide the knowledge to the users. They disclose all information to the users. The users must “mine” the information of interest with the provided tools. Even though these tools disclose different distribution structures of the database, most of the distribution structures are uninteresting for many users. Only the **unexpected events**, such as the announcement of Osama bin Laden, can catch the eyes of these users.

To bridge these two gaps, we have developed a novel framework for news video analysis and visualization. As shown in Figure 1, our proposed framework is able to integrate the achievements of three research areas for supporting more effective video exploration: *semantic video analysis*, *video retrieval* and *knowledge visualization*. The semantic video analysis techniques extract the underlying semantics for video retrieval and knowledge visualization. The video retrieval technique provides a good approach of reasoning in support of knowledge visualization. And the video retrieval improves the semantic video analysis performance by statistical analysis. By integrating the visualization interfaces, the system is able to take advantage of human intelligence to significantly improve the performance of semantic video analysis and video retrieval. In addition, the visualization techniques enable the system to support intuitive exploration of video news reports.

The methods and results of this paper are a significant example of the visual analytics approach, where automated analyses are closely coupled with interactive visualization to attack large, complex data, supporting exploration and discovery. The knowledge retrieved provides essential material for analytical reasoning and decision-making.

II. LARGE-SCALE VIDEO NEWS EXPLORATION FRAMEWORK

By integrating the semantic video analysis, video retrieval and visualization techniques together, we arrive at a framework with the workflow given in Figure 2. Firstly, the semantic interpretation is extracted from raw video clips via semantic video analysis techniques. Secondly, the knowledge interpretation is

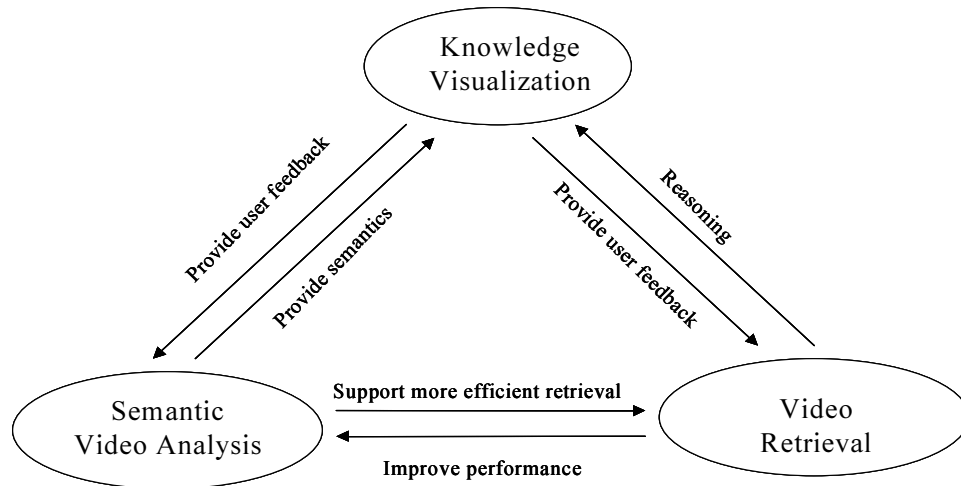


Fig. 1. The interaction among relevant research areas

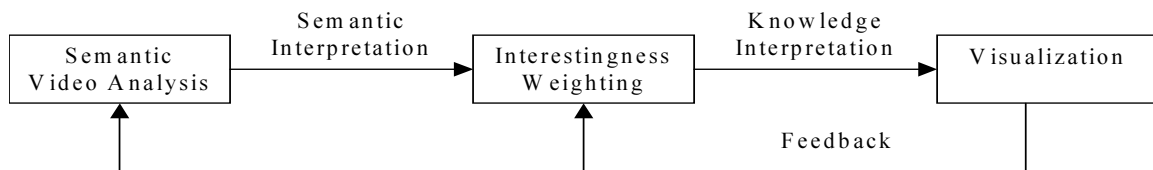


Fig. 2. The workflow of the framework.

extracted by weighting the semantic interpretation according to an interestingness measurement. Finally, visualization techniques are adopted to represent the knowledge interpretation. The users' feedback can be transferred back to the semantic video analysis and interestingness weighting to improve their performance and support reasoning.

To implement such a system, the first problem is how to interpret the underlying semantics of the video news reports. Because a news report may cover multiple entities and events, it must be partitioned to smaller units so that the fine details can be interpreted and processed by the interestingness weighting and the visualization algorithms. An appropriate interpretation unit is very important. Too coarse interpretation may cause information loss, thus the interestingness weighting and visualization performance may be low. Too fine interpretation may carry high noise because the semantic video analysis and natural language processing techniques are unable to extract enough fine details. A natural unit for interpreting the semantics of the video clips is the video shot. Some special visual objects, such as human faces and video text, also carry abundant information and are used to interpret the semantics of the video clips. In addition, the keywords are appropriate for closed caption and automatic speech recognition (ASR) script

interpretation. To disclose as much information as possible to the users, all these units are used in our system to interpret the semantics of the video news reports. These units are called semantic items in this paper. The semantic items carry abundant information of the news video reports.

To have better visualization, the video news reports must be weighted according to their interestingness so that analysts and general audiences can be relieved from the burden of watching a large volume of “normal”, “safe” or “uninteresting” information. Thus the knowledge interpretation of the video news reports must be obtained according to their interestingness measurement.

Based on the above observations, we need to address the following problems: (1) How to extract semantic items from the video news reports. (2) How to extract knowledge interpretations from semantic items (e.g. interestingness weighting). (3) How to visualize the extracted knowledge with the semantic items. Since the semantic interpretation and knowledge interpretation directly service the visualization, the adopted visualization technique will provide requirements for them. Based on this observation, we first introduce the visualization technique in the next section.

III. KNOWLEDGE VISUALIZATION FRAMEWORK

In this section we discuss the visualization techniques that can best serve users for large-scale video news database exploration. To resolve this problem we need to answer two questions: (1) Which information is most useful for users (e.g. knowledge extraction) ? (2) How do we represent the knowledge to the users intuitively from the semantic interpretation and the knowledge interpretation of the video news reports?

As the news reports are unpredictable, both the general audiences and the analysts may want to have a rough idea of all available news reports the system can provide. This means that **global overview information** is the first piece of knowledge that is useful for most audiences. The global overview is not necessarily formed by whole news stories. It can be composed of the semantic items with their interestingness weights. The semantic items are better than the whole news stories for the global overview because not all aspects in a news story are equally interesting. The audiences may be interested in a certain small point of a news report, such as a name or an interesting video shot.

The semantic items provide a good hint to the users. The users can quickly make the decision of which semantic item is more interesting than others according to their own preference and knowledge and invoke queries by clicking interesting items. Because the semantic items are extracted from the video news collection of the system, the system assures the users that each query via semantic items will return correlated, and most probably interesting news stories. The traditional keyword-based retrieval, however,

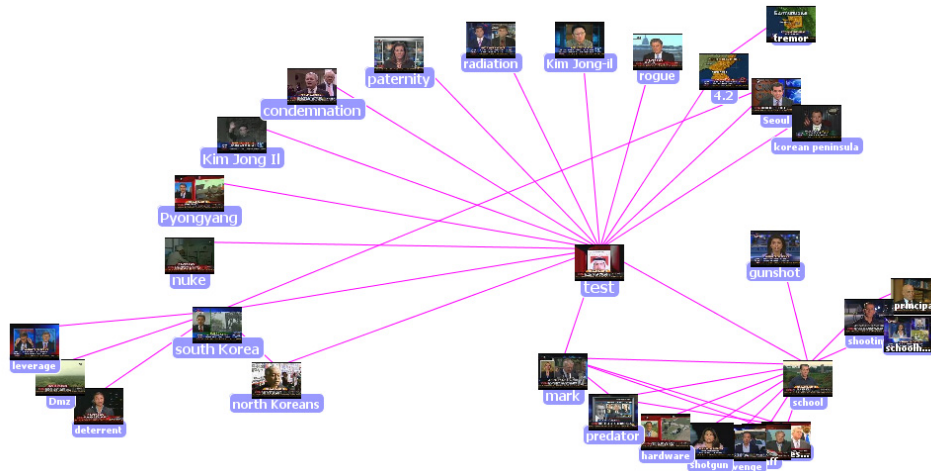


Fig. 3. Links of the semantic item “test” disclose details of the event and response of the international community during the North Korean nuclear weapon test.

does not have such nice property. The users of keyword-based retrieval system may type in keywords that do not have correlated news stories recently.

In addition, the interestingness weights are better than the raw distribution for global overview visualization. By using the interestingness weights, frequent but generally uninteresting semantic items can be suppressed and the really unexpected semantic items emphasized. Thus more knowledge can be represented in the same screen space, and there is much higher chance for the users to find a piece of interesting information.

As the content of news reports are dynamic, the audiences may also want to know the **trend over time** of the semantic items. The dynamic trend is able to tell the development procedure of the whole event. Thus the users can have more complete view of the interesting topics with the dynamic trend.

In addition, the **relations** among semantic items are also interesting. For example, the keywords “North Korea”, “test” and “nuclear weapon” frequently appeared in the October and November 2006 news reports. The relations among these keywords are strong and interesting for the news reports during that time. The semantic items are generally objects; thus they can’t reflect complete events. The relations, on the other hand, can disclose the events. The audiences can learn higher-level semantics from the relations. As shown in Figure 3, the relations disclose interesting and useful event knowledge to the users.

To effectively depict the above information, several visualization techniques are adopted. Firstly, global overview information is represented by using the keyframes map, as shown in Figure 4(a) and 4(d). The size of the keyframe in the map is proportional to its interestingness weight. By organizing the global

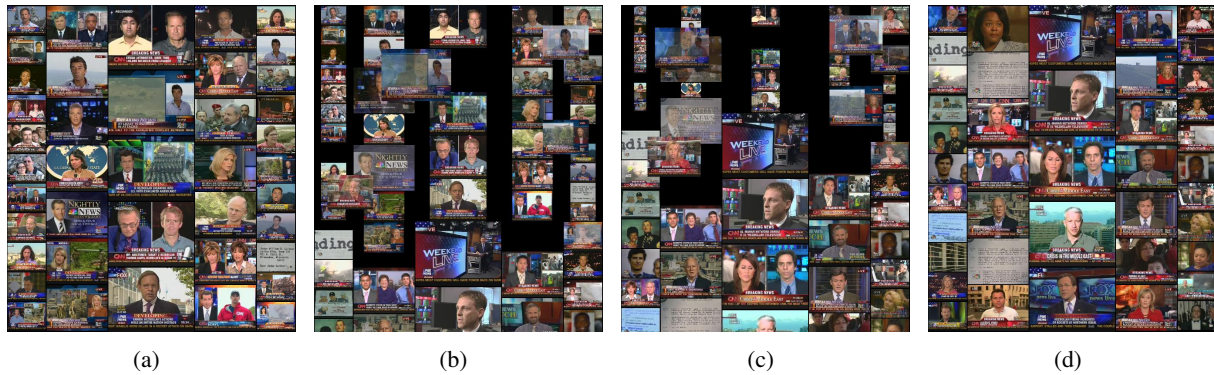


Fig. 4. An example of video news visualization. (a) Keyframes map for U.S. news on July 22, 2006; (b) and (c) Intermediate animation; (d) Keyframes map for U.S. news on July 23, 2006. The keyframes maps show the news topics on given day, the animation represents the trend of topic change over time. An example of animation can be downloaded at <http://webpages.uncc.edu/~hluo/NewsDemo.avi>.

overview information in this way, the users can directly find the interesting news reports on the keyframes map and learn the global structure of all news reports. To represent the dynamic trend of the news reports, an animation of keyframes on the keyframes map is used. Two animation frames are given in 4(b) and 4(c). An example of animation can be downloaded at <http://webpages.uncc.edu/~hluo/NewsDemo.avi>. By watching the animation the users are able to catch the dynamic trend of the news topics.

To represent the relations among the interesting semantic items, a network is used. As in Figure 5, the most interesting relations are represented as the edges and the most interesting semantic items are represented as the nodes. The users can rotate, translate and zoom the network to examine the details at different level. An online demo can be found at <http://webpages.uncc.edu/~hluo/relation/Relation.html>. The relations disclose another level of knowledge to the users.

When the users browse the visualized keyframes and keywords, they can also click the corresponding items to submit a query. Our system can then retrieve the news video databases according to the selected semantic items and most relevant news stories are selected and returned to the users. The retrieved stories can be organized by timeline so that the users can easily learn the development procedure of the whole event, as shown in Figure 6(a). In addition, the most relevant web news can also be retrieved, as shown in 6(b). This feature is very important for audiences who want to know more details and relevant discussions of the event. This video retrieval process is a good reasoning technique for news video exploration and analysis.

To implement the above visualization, the interestingness of the semantic items and their relations must be quantified (e.g. extracting the knowledge interpretation). Thus we discuss the algorithm for extracting

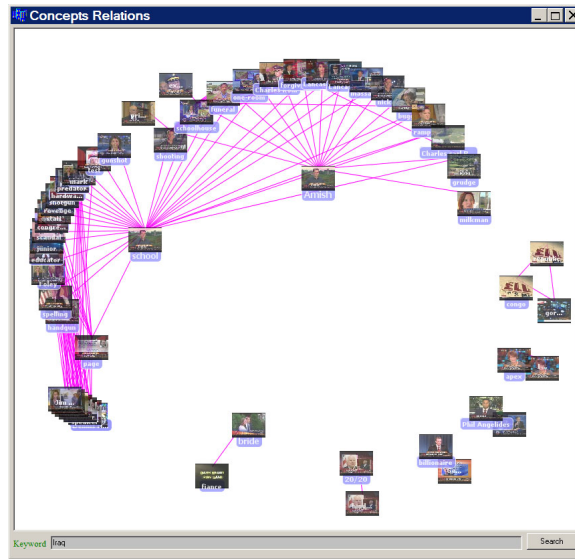


Fig. 5. An example of relations visualization. An online demo can be found at <http://webpages.uncc.edu/~hluo/relation/Relation.html>.

(a) Results by timeline

Site	Title
www.cfra.com	580 CFRA - News Talk Radio
www.wcsh6.com	WCSH6.com - BREAKING NEWS - 4 Students Killed,
www.wzzm.com	WZZM 13 Grand Rapids, Michigan - UPDATE: Fifth
today.reuters.com	Gunman kills 3 girls at Pennsylvania Amish school sp
www.abc.net	Fifth girl dead after Amish school shooting. 03/10/2006.
www.guardian.co.uk	Guardian Unlimited The Guardian Six killed in Amish s
www.buzzfeed.com	Students Shot Dead at Amish School
www.voanews.com	VOA News - Amish to Bury Four Children Slain in Penn
news.xinhuanet.com	Xinhua - English
www.fox6.com	FOX6 San Diego - Amish burying fifth shooting victim

(b) Cross media retrieval results

Fig. 6. An example of search results.

the knowledge interpretation of the video news reports in the next section.

IV. KNOWLEDGE INTERPRETATION EXTRACTION

Based on the above observations, the distribution structure of the semantic items and their relations must be quantified to implement the visualization system. Often visualization techniques use the raw frequency or probability to organize the visualization. However, the raw frequency or probability does not have direct correlation with the users' preference. For example, "Bush" is a keyword with very high frequency in 2006. But many audiences may not be interested in the keyword "Bush". In the scenario of news browsing and retrieval applications, a user may be interested in only the information that he or she did not know before (e.g. the "**really unexpected**" information).

To resolve this problem, the **interestingness** distribution of the semantic items and relations must be extracted. Apparently, the interestingness is related to the users' preference and knowledge. However, it is very difficult to gather user preference information and use it in an integrated way. Because the users may learn information from sources that cannot be controlled by the computers, such as friends, it's even more difficult to have an accurate knowledge model of a specific user in the foreseeable future. Thus a general interestingness measurement that is effective for most users must be proposed.

Google has implemented a great search engine based on the PageRank [10] technique. The PageRank technique ranks the web pages according to the provider behavior (e.g. the links of web pages). Even though Google has no user preference data, the PageRank technique can still give reasonable ranks of web pages for most users. Based on this observation, the **provider behavior** is valuable information for quantifying the interestingness. More importantly, it's possible to gather the provider behavior information. Thus we can quantify the interestingness of semantic items by using the statistical information extracted from many TV channels. To quantify the interestingness of semantic items with the provider behavior, we assume that the audiences may have higher probability to know a piece of information if it is repeated more frequently on TV programs. Thus the outdated news collection can be used to extract a general knowledge model of the users' knowledge. We use a probability distribution to represent the knowledge model:

$$G = \{g(x) | x \in S\} \quad (1)$$

where x is the given semantic item, S is the set of all semantic items or their relations and $g(x)$ is the probability of the semantic item or relation x . The general knowledge model can be used as a predictor. If a semantic item or relation can be predicted completely, then it may not be interesting at all for the users. Only those semantic items and relations that cannot be predicted well are interesting for the users. According to information theory, the predictability of a message (e.g. semantic item or relation in our

system) can be quantified by the information it carries. To quantify the amount of the information a semantic item or relation carries, the local probability model of semantic items or relations of news reports for a specific time interval of interest is defined as:

$$L(t) = \{l_t(x) | x \in S_t \subseteq S\} \quad (2)$$

where t is the specific time interval, such as one specific day, S_t is the set of all semantic items or relations in the specific time interval t and $l_t(x)$ is the probability of the given semantic item or relation x . The difference between the local probability model $L(t)$ for a specific time interval t and the general knowledge model G is able to tell us how much information we can obtain by knowing $L(t)$. Because both the local probability model $L(t)$ and the general knowledge model G are distributions, the Kullback-Leibler divergence is used to characterize their difference:

$$D(L(t) \parallel G) = \sum_{x \in S_t} l_t(x) \log \frac{l_t(x)}{g(x)} \quad (3)$$

The distance function $D(L(t) \parallel G)$ is able to characterize the difference between $L(t)$ and G , but we also need to evaluate the information carried by each semantic item or relation $x \in S_t$. By examining Eq. (3), one can observe that $D(L(t) \parallel G)$ is composed of a set of components, and each item x is only related to one single component in $D(L(t) \parallel G)$. Thus the contribution for one certain item $x \in S_t$ can be obtained by the relevant component of $D(L(t) \parallel G)$ in Eq. (3). Based on this observation, we can quantify the interestingness of one certain item x as:

$$w_t^r(x) = l_t(x) \log \frac{l_t(x)}{g(x)} \quad (4)$$

From Eq. (4), one can observe that the interestingness measurement $w_t^r(x)$ for one certain item x depends on two factors: $l_t(x)$ and $\frac{l_t(x)}{g(x)}$. The first factor $l_t(x)$ is used to characterize the local probability model $L(t)$ and the second factor $\frac{l_t(x)}{g(x)}$ is used to characterize its difference with the general knowledge model G . Eq. (4) may emphasize the local probability too much in some situations. For example, in real news video programs, the same anchorperson may appear many times repeatedly in the same news program and may also appear in the news programs offered at different time period from the same TV channel. Thus the semantic item for him/her may have high frequency in the local probability model $L(t)$. If Eq. (4) is directly used to organize the map of the keyframes (i.e., map of news stories of interest), we may select many anchor shots for visualizing the news stories of interest, which is unacceptable. Based on this observation, only the difference between the local probability model $L(t)$ and the general

knowledge model G should be used to characterize the interestingness measurement:

$$w_t^d(x) = \frac{l_t(x)}{g(x)} \quad (5)$$

Because multiple media channels (audio, video and closed caption text) are involved in the visualization, the $w_t^d(x)$ in Eq. (5) needs to be normalized to simplify the multi-modal data fusion:

$$\bar{w}_t(x) = \frac{w_t^d(x)}{\max_{x \in S_t} \{w_t^d(x)\}} \quad (6)$$

To enable more efficient visualization of large-scale news video collections, special visual features should be considered. There are multiple types of visual features that may be important. Some types of visual features can be processed by using Eq. (6), such as the human faces and the semantic concepts of news video clips. Other types of visual features may not be characterized by using the same statistical analysis algorithms as described above (for example, video production rules and text areas in news videos). To extract such kind of visual features, we have also developed some specific statistical video analysis techniques as described later in this paper.

When the multi-modal importance weights (i.e., importance weights for video, audio, closed caption, special visual features) for all these semantic items are determined, they are combined to determine the overall weight for the given video clip. The unit of video clips (e.g., the partitioning over time) should be carefully selected to enable the best visualization. Too large a unit may cause information loss; too small a unit may carry too many unrelated details. Because of the natural properties of video, the video shot is a suitable unit for visualization. The overall weight for a given video shot is defined as:

$$w(i) = F\left(W(S_v(i)), W(S_a(i)), W(S_c(i)), W'(V(i))\right) \quad (7)$$

where i represents the i -th video shot, $S_v(i)$, $S_a(i)$ and $S_c(i)$ are the multi-modal semantic items (i.e., video, audio and closed caption) extracted from the i -th video shot, $W(*)$ represents the set of weights determined by Eq. (6), $V(i)$ is the special visual feature set for the i -th video shot, $W'(V(i))$ is the weight set assigned according to the video production rules and $F(*)$ represents the underlying fusion function. Based on $w(i)$, we can visualize the semantic items and relations more effectively.

With the above semantic items and relations weighting algorithm, our system is able to extract the interesting **knowledge** and suppress uninteresting information. In addition, the knowledge is extracted via statistical approaches. Thus the random error of semantic video analysis can be filtered out and more robust results can be achieved.

V. SEMANTIC INTERPRETATION EXTRACTION

In above sections we have assumed that the semantic items and their relations are available. In this section we propose algorithms to extract the multi-modal semantic items and their relations for the large-scale video news exploration framework.

Even though there are many sophisticated algorithms for statistical text analysis, semantic video analysis and understanding are still very challenging for current computer vision technologies. The problem is caused by the **semantic gap** between the semantics of video clips from the human point of views and the low-level features that can be extracted by computers [1]. However, supporting semantic video analysis plays an important role in enabling more efficient exploration of large-scale news videos. Without extracting the semantic topics from large-scale news video collections, it is very difficult to visualize them effectively. Based on this observation, we have developed novel algorithms to extract the multi-modal semantic items (i.e., video, audio, text) and some special visual features automatically. Weights are assigned automatically with this statistical video analysis algorithm.

A. Semantic Video Analysis

The basic unit for news video interpretation is the video shot. Unlike the keywords of text documents, a video shot may contain abundant information (i.e., an image is more than one thousand words). This specific property of the video shot makes it difficult to effectively achieve statistical analysis on its visual properties and assign importance weights to the corresponding video shots for news video visualization. To overcome this, we have developed a novel framework for statistical video analysis.

There are three types of semantic units that are critical to determine the importance weights for the corresponding video shots: (a) the first one is the statistical properties of the video shots; (b) the second one is the special video objects that appear in the video shots; (c) the last one is the semantic concepts that are associated with the video shots. Because these three types of semantic units have different properties, different algorithms are needed to extract such multi-modal semantic items.

1) Statistical Property Analysis of Video Shots: The video shots are the basic unit for news video interpretation. Thus they can be treated as the semantic items for automatic weight assignment. However, unlike the keywords in text documents, the repeating of video shots cannot be detected automatically by using simple comparison of the video shots. Thus new techniques are desired for detecting the repeat of video shots in news videos [11], such that we can assign the importance weights for the video shots automatically.

One certain video shot may be repeated multiple times because of the following reasons: (1) video shots for the anchors may repeat multiple times in the same news program; (2) video shots for the participants of an interview may appear multiple times in the same news program; (3) video shots for interpreting the important news may appear in both the news summary at the beginning and the detailed report later in the same program; (4) video shots for the important news may appear in different news programs of the same channel (at different time periods) or different TV channels (at different time or same time). The last two situations of video shot repeating indicate the importance for the corresponding video shots. Nevertheless, the first two situations of video shot repeating may not indicate that the corresponding video shots are important. To quantify the effect of above 4 rules, the intra-program repeating number $r_{intra}(i)$ and inter-program repeating number $r_{inter}(i)$ for each video shot are computed. The two numbers $r_{inter}(i)$ and $r_{intra}(i)$ for most video shots are equal to 1 because they are not repeated. Obviously, some video shots may have these two numbers bigger than 1 and different repeating patterns (i.e., different repeating situations) may provide different semantics so that different weights should be assigned. The weights for different repeating numbers are approximated by using a bell shaped curve:

$$\begin{aligned} w_{intra}(i) &= e^{-\frac{(r_{intra}(i)-2)^2}{2}} \\ w_{inter}(i) &= e^{-\frac{(r_{inter}(i)-5)^2}{2}} \end{aligned} \quad (8)$$

2) *Video Objects Detection*: For news videos, text areas and human faces may provide important clues about news stories of interest. Text lines and human faces in news videos can be detected automatically by computer vision techniques [11]. Obviously, these automatic detection functions may fail in some cases. Thus the results that are detected by using a single video frame may not be reliable. To address this problem, the detection results for all the video frames within the same video shots are integrated and the relevant confidence maps for the detection results are calculated. As shown in Figure 7, such confidence maps can provide valuable information for evaluating the detection results.

The confidence region is generated by transforming the relevant confidences for our detection results into a binary image via threshold. The threshold for generating the confidence region of text is set to 0.5. The threshold for generating the confidence region of human faces is set to 0.35. Obviously, the size ratio between the confidence region and the size of video frames provides some valuable information for weight assignment, and thus the size ratios for text and human faces regions are obtained, $\alpha_{text}(i)$ and $\alpha_{face}(i)$. A sigmoid curve is used to determine the importance weights for the text regions and human

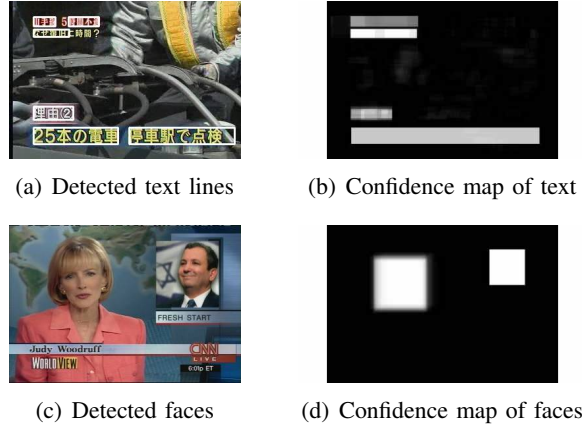


Fig. 7. Automatic text and face detection results

faces:

$$w_{area}(i) = \frac{1}{1 + e^{-\frac{\max\{\alpha(i) - \nu, 0\}}{\lambda}}} \quad (9)$$

where the parameters ν and λ are used to control the shape of the curve. In our current implementation, $\nu_{text} = 0.05$, $\lambda_{text} = 0.1593$, $\nu_{face} = 0.01$ and $\lambda_{face} = 0.04096$. For a given video shot, the importance weight for human faces $w_{faceArea}(i)$ and the importance weight for text regions $w_{textArea}(i)$ can be determined by:

$$w_{textArea}(i) = \frac{1}{1 + e^{-\frac{\max\{\alpha_{text}(i) - \nu_{text}, 0\}}{\lambda_{text}}}} \quad (10)$$

$$w_{faceArea}(i) = \frac{1}{1 + e^{-\frac{\max\{\alpha_{face}(i) - \nu_{face}, 0\}}{\lambda_{face}}}}$$

By performing the face clustering technique, face objects can be clustered to several groups and the human objects can be identified and be treated as the semantic items for weight assignment by using Eq. (6). The importance weight for human faces of shot i is computed by:

$$w_{face}(i) = \begin{cases} \max_{x \in FACE(i)} \{\bar{w}_t(x)\} & FACE(i) \neq \emptyset \\ 0.5 & FACE(i) = \emptyset \end{cases} \quad (11)$$

where $FACE(i)$ is the set of face objects of shot i .

3) *Semantic Video Classification*: The semantic concepts of the video shot can provide valuable information to enable more efficient and effective visualization and retrieval of large-scale news video collections. Semantic video classification is one of the potential solutions to detect the semantic concepts

TABLE I
SEMANTIC CONCEPT IMPORTANCE

<i>Concept</i>	w_c	<i>Concept</i>	w_c
Announcement	0.9	Report	0.3
Sports	0.5	Weather	0.5
Gathered People	1	Unknown	0.8

for the video shots. To incorporate this capability, we adopt a principal video shot-based semantic video classification algorithm [2] in our system.

Two types of information about semantic concepts can be used for weight assignment: (1) the importance of the given semantic concepts; (2) the distribution of the semantic concepts in a given time interval of interest. The importance of the semantic concepts, $w_c(C(i))$, is assigned as in Table I. Where $C(i)$ is the semantic concept in the video shot i . The importance for the concept distribution can be determined by Eq. (6), $w_d(i) = \bar{w}_t(C(i))$. Finally, the weight for the given semantic concept is determined by:

$$w_{concept}(i) = w_c(C(s)) \times w_d(i) \quad (12)$$

B. Audio and Text Items Extraction

For news videos, the text documents from the closed captions match well with the news audio, and thus they can be integrated to take advantage of both media to clarify content and remove redundant information. The text documents for the closed captions may not synchronize with the video and generally have a delay of a few seconds. On the other hand, the audio generally synchronizes very well with the video but the accuracy of most existing techniques for speech recognition is still low. By integrating the results for speech recognition with the results of closed caption analysis, the closed captions can be synchronized with the video with high accuracy.

After the closed captions are synchronized to the relevant videos, we can determine the correlation between the closed captions and the video shots. To do this, the closed captions are first segmented to sentences, and the start time and the stop time for each text sentence can also be obtained automatically. All video shots that locate between the start time and the stop time for the same text sentence are associated with the corresponding text sentence. In addition, the text sentence is further segmented to keywords. All the video shots associated with the same text sentence are associated with all the keywords

in the same text sentence.

In news videos, the news titles shown in video may provide very important keywords and thus they should be detected. Some special text sentences, such as “*somebody*, CNN, *somewhere*” and “ABC’s *somebody* reports from *somewhere*”, need to be processed separately. The names for news reporters in those text sentences are generally not useful to the users. Because there are clear and fixed patterns for these sentences, we have designed a context-free syntax parser to detect and mark this information. By incorporating 10-15 syntax rules, the parser can detect and mark the special sentences in high accuracy.

Most named entity detectors may fail in processing all capital strings because initial capitalization is very important to achieve accurate named entity recognition. One way to resolve this problem is to train a detector with ground truth from closed caption text. However, it’s very expensive to obtain the manually marked text material. Because English has relatively strict grammar, it’s possible to parse the sentence and recover most capital information by using part-of-speech (POS) [12] and lemma information. We use TreeTagger [12] to perform the part-of-speech tagging. Capital information can be recovered automatically by using the TreeTagger parsing results.

After special sentences are marked and capital information is recovered, a opensource text analysis package LingPipe is used to perform the named entity detection and resolve coreference of the named entities. The named entities referring to the same entity are normalized to a most representative format to enable statistical analysis. The model used is the news model of LingPipe. All parameters are set to default value.

Finally, the normalized results are parser by TreeTagger the second time to extract the POS information and resolve the words to their original formats. For example, TreeTagger can resolve “better” to “well” or “good” according to its POS tag. We do not adopt the stemming technique because it may output unreadable words and resolve different words to the same stem. By using the POS tag to resolve the words to their original formats, this problem can be resolved. In addition, the POS tag can be used to remove words without real meanings, such as adverbs and prepositions. Most stop words can be removed by POS tag.

The TreeTagger results can be associated to video shots according to their time stamps. After all shots have been associated with keywords, the keyword weight of a shot is computed by:

$$w_{keyword}(i) = \max_x \{\bar{w}_t(x) | x \text{ is a keyword of } i\} \quad (13)$$

C. Multi-Modal Data Fusion

To enable more efficient visualization of large-scale news video collections, an overall weight is assigned with each video shot based on the weights described above. First, the w_{intra} and the w_{inter} are fused to compute the weight for the repeating video shot:

$$w_{repeat}(i) = \max\{w_{intra}(i), w_{inter}(i)\} \quad (14)$$

The $w_{faceArea}$ and $w_{textArea}$ are fused to compute the object weight:

$$w_{object}(i) = \max\{w_{faceArea}(i), w_{textArea}(i)\} \quad (15)$$

The w_{face} and $w_{concept}$ are both related to semantics of the corresponding video shot, thus they are integrated to determine the semantics weight:

$$w_{semantics}(i) = \max\{w_{face}(i), w_{concept}(i)\} \quad (16)$$

The video importance weight for a given video shot is determined by the geometric average of above three weights:

$$w_{video}(i) = \sqrt[3]{w_{repeat}(i) \times w_{object}(i) \times w_{semantics}(i)} \quad (17)$$

Finally, the overall weight for the given video shot is determined by averaging w_{video} and $w_{keyword}$:

$$w(i) = \gamma \times w_{video}(i) + (1 - \gamma) \times w_{keyword}(i) \quad (18)$$

In our current experiments, we set $\gamma = 0.6$.

D. Relations Extraction

The relations among semantic items are also important information for video news exploration. As semantic items have already been extracted, the relations among semantic items can be extracted by measuring the concurrence of semantic items. When a pair of semantic items occurs simultaneously within the time period of a closed caption or ASR script sentence, they are considered to be a concurrence. By counting the concurrences of semantic items pairs in the database, the probability distribution of relations can be computed.

However, as with highly frequent semantic items, highly frequent relations may not be always interesting for users. For example, the semantic items pair ‘‘President’’ and ‘‘Bush’’ has very high frequency in recent years’ news reports. But it’s not very useful because almost everyone already knows this

relation. To resolve this problem, the weighting algorithm introduced in Section IV is also applied to the relations. After weighting, the semantic items pairs with highest weights can be used to represent the most interesting relations among semantic items.

VI. CONCLUSIONS

By incorporating knowledge extraction and semantic video analysis for news video exploration, our system enables visualization of large-scale news video collections, disclosing valuable knowledge to the users. The users may often find the news reports of interest at the first glance. This effectiveness and efficiency would be impossible for video content of this scale without the integrated, multimodal analyses described here.

In this paper, we have developed and used a novel interestingness measurement for semantic items and relations. By organizing the visualization with this interestingness measurement, our system is able to bridge the **interest gap**. Analysts and general users can find unexpected and interesting events from the large-scale video news database more efficiently with the resulting knowledge visualization. Figures 3 ~ 6 give examples of the resulting knowledge visualizations that support exploration and discovery.

The methods and results of this paper are a significant example of the visual analytics approach, where automated analyses are closely coupled with interactive visualization to attack large, complex data and support analytical reasoning.

Future work will involve using the knowledge content derived from the multimodal fusion described here to scale up the interactive visualization, permitting exploration over long periods of time and hundreds of geographically dispersed broadcast channels. This work will also be applied to other types of multimedia.

VII. ACKNOWLEDGMENT

This work was sponsored by the National Visualization and Analytics Center (NVACTM) under the auspices of the Southeast Regional Visualization and Analytics Center. NVAC is a U.S. Department of Homeland Security Program led by Pacific Northwest National Laboratory.

REFERENCES

- [1] A. W. Smeulders, M. Worring, S. Santini, AmarnathGupta, and R. Jain, "Content-base image retrieval at the end of the early years," *IEEE Trans. on PAMI*, vol. 22, no. 12, pp. 1349–1380, December 2000.
- [2] J. Fan, H. Luo, and A. K. Elmagarmid, "Concept-oriented indexing of video database toward more effective retrieval and browsing," *IEEE Trans. on Image Processing*, vol. 13, no. 7, pp. 974–992, 2004.

- [3] J. van Wijk, "Bridging the gaps," *Computer Graphics and Applications, IEEE*, vol. 26, no. 6, pp. 6–9, Nov.-Dec. 2006.
- [4] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, "Visualizing the non-visual: spatial analysis and interaction with information from text documents," in *IEEE InfoVis 1995*, 1995.
- [5] A. Mehler, Y. Bao, X. Li, Y. Wang, and S. Skiena, "Spatial analysis of news sources," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 765–772, 2006.
- [6] R. Swan and D. Jensen, "Timemines: Constructing timelines with statistical models of word," in *SIGKDD*, 2000.
- [7] M. Weskamp, "<http://www.marumushi.com/apps/newsmap/index.cfm>."
- [8] S. Havre, B. Hetzler, and L. Nowell, "Themeriver: Visualizing theme changes over time," in *IEEE InfoVis*, 2000.
- [9] B. Hetzler, P. Whitney, L. Martucci, and J. Thomas, "Multi-faceted insight through interoperable visual information analysis paradigms," in *InfoVis'98*, Research Triangle Park, NC, 1998.
- [10] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," <http://dbpubs.stanford.edu:8090/pub/1999-66>, Tech. Rep.
- [11] H. Luo, J. Fan, J. Yang, W. Ribarsky, and S. Satoh, "Exploring large-scale video news via interactive visualization," in *IEEE Symposium on Visual Analytics Science and Technology*, Baltimore, USA, November 2006.
- [12] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *International Conference on New Methods in Language Processing*, Manchester, UK, 9 1994.