Analyzing Large-Scale News Video Databases to Support Knowledge Visualization and Intuitive Retrieval

Hangzai Luo*
Software Engineering Institute
East China Normal University
Shanghai, China

Jianping Fan, Jing Yang, William Ribarsky†
Department of Computer Science
UNC-Charlotte
Charlotte, NC, USA

Shin'ichi Satoh‡
National Institute of Informatics
Tokyo, Japan

ABSTRACT

In this paper, we have developed a novel framework to enable more effective investigation of large-scale news video database via knowledge visualization. To relieve users from the burdensome exploration of well-known and uninteresting knowledge of news reports, a novel interestingness measurement for video news reports is presented to enable users to find news stories of interest at first glance and capture the relevant knowledge in large-scale video news databases efficiently. Our framework takes advantage of both automatic semantic video analysis and human intelligence by integrating with visualization techniques on semantic video retrieval systems. Our techniques on intelligent news video analysis and knowledge discovery have the capacity to enable more effective visualization and exploration of large-scale news video collections. In addition, news video visualization and exploration can provide valuable feedback to improve our techniques for intelligent news video analysis and knowledge discovery.

Keywords: Semantic Video Classification, Knowledge Discovery, Knowledge Visualization.

Index Terms: I.2.6 [Artificial Intelligence]: Learning— Concept learning; I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction Techniques

1 INTRODUCTION

Broadcast video news is a very important information source to most people. To satisfy the extremely diverse demands of the public, it covers large amount of events everyday. As a result, it provides a picture of what is happening now at the local, national, and international levels. 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 rapidly smaller. Therefore, there is an urgent demand for achieving intuitive and effective 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 clips. Before the system can provide an automatic video news exploration service, it must understand the underlying semantics of the input video clips. However, there exists a big **semantic gap** [10, 1] between the low-level visual features and the

---

*e-mail:memcache@gmail.com
†e-mail:{jfan, jyang13, ribarsky}@uncc.edu
‡e-mail:satoh@nii.ac.jp

high-level semantic video concepts. Existing video exploration systems can only support the services based on low-level visual features [3]. Typical users, however, 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, but they can hardly satisfy the requirements of semantic video news database exploration applications. With a limited video semantics, how to provide intuitive applications for video news database investigations is still an open problem.

The second problem is how to *extract the most useful knowledge* from the large-scale video news database and *display such knowledge to the users*. Because the total amount of knowledge for a large-scale video news database is very large (e.g., thousands of hours of video), most of the information is irrelevant to the point of interest. If all information is delivered to the analysts or audiences, they may easily get lost and miss the important information. For example, "Bush is the president of the USA" 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. Abnormal information is more useful and interesting for the users. Thus, there is an **interest gap** [14] between the underlying information collection and the user's interest.

The third problem is how to *launch personalized knowledge retrieval upon receiving input from the users*. Before the users submit any input to express their preferences and information needs, the system can only display a general overview of all knowledge. The general overview of all knowledge discloses the global overview of the database but does not disclose enough details to fit the user preferences or the user's current information needs. As a result, the knowledge structure must be reorganized after the system receives user input, so that more details related to the user input can be disclosed. Existing systems adopt retrieval techniques to extract a few most relevant items from the database. However, the traditional retrieval techniques can only provides a set of possibly relevant items. How to directly disclose the personalized knowledge structure via these relevant items is still an open problem.

Researchers have proposed different approaches to resolve these problems. Semantic video classification is one of the potential solutions to bridge the semantic gap. To achieve video understanding via semantic classification, the semantic classification algorithms first extract features from the video clips and then classify the video clips to semantic concepts according to their feature vectors via machine-learning algorithms. However, they are optimized toward keyword-based search applications. As a result, they may not be optimal for video news exploration and analysis applications. Therefore, how to extract suitable semantics for video news exploration and analysis applications is still an open problem.

Visualization approaches have been proposed to help the users explore in information spaces and find interesting parts intuitively. InSpire [16] transforms the text document collection of interest to a spatial representation for visualization and analysis. For example, statistical information of news reports [9] could be put on a world map to inform the audience of the "hotness" of regions and the relations among the regions. TimeMine [13] is able to detect the most
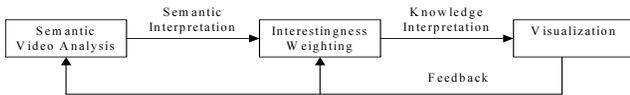
Figure 1: The workflow of the framework.

important reports and organize them through a timeline with statistical models of word usage. Another system, called newsmap [15], organizes news topics from Google news on a 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 [4] and ThemeView [5] can visualize a large collection of documents with keywords or themes over time or as aggregations of related themes. ThemeRiver and ThemeView can represent intuitively the distribution structure of themes and keywords of the database. However, all of these visualization systems cannot directly provide knowledge to the users. Rather, they disclose relevant information that the users must investigate to form their own conclusions. Although the tools provided 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.

In addition, all existing algorithms address the above problems separately. Therefore, they may optimize the solution for different purposes. On the one hand, the semantic video classification algorithms are generally optimized for keyword-based video retrieval applications [1, 7]. As a result, the semantic video concepts implemented may be suitable for search but not suitable for visualization. On the other hand, the visualization approaches focus on providing new techniques for information representation [14] and assume the information is somehow available for immediate use. However, the most useful information can only be extracted by using state-of-the-art semantic analysis algorithms. By addressing the two problems together, these mismatches can be avoided and the performance of the system can be improved significantly.

To resolve the above problems, we offer here a knowledge visualization framework that can integrate achievements on semantic video analysis, information retrieval, knowledge discovery, and knowledge visualization. The framework is introduced in Section 2. Sections 3 through 5 introduce algorithms to implement different components of the framework. Finally we conclude in Section 7.

2 KNOWLEDGE VISUALIZATION FRAMEWORK

Based on the above observations, one finds that it is very difficult, if not impossible, to resolve the problems addressed in the above section independently in a single research area. A solution can only be achieved by integrating achievements on *semantic video analysis*, *information retrieval, knowledge discovery,* and *knowledge visualization*. In addition, all components must be optimized toward a single target to achieve intuitive and intelligent exploration of large-scale video databases. Based on this understanding, the workflow of our framework is shown in Figure 1. First, the semantic interpretation is extracted from raw video clips via semantic video analysis techniques. Second, the knowledge interpretation is extracted by weighting the semantic interpretation according to an interestingness measurement. Third, visualization techniques are adopted to represent the knowledge intuitively. Finally, the semantic interpretation and the knowledge interpretation can be improved through the user input received via the visualization interface.

To establish the best visualization design, the knowledge interpretation must be carefully selected to satisfy as many user needs as possible. Therefore, the user needs must be carefully analyzed. Our system is targeted at two major types of users: analysts and general audiences. Their needs are discussed below.

First, the purpose of both analysts and general audiences in using the system is to gain knowledge of the database. To disclose as much knowledge of the database as possible, the knowledge interpretation must be in a format that is intuitive and easy to visualize.

Second, the users may not be interested in most knowledge in the database. Large-scale news video databases carry a large amount of knowledge, but much of it is common sense and thus not interesting for most users. As a result, the knowledge interpretation must be able to suppress general, uninteresting knowledge and emphasize abnormal, interesting knowledge.

Third, the user's interest viewpoint may change during the exploration. When a user finds an interesting event, it is preferable to disclose the semantic structure of the event and other relevant events to the user, so that the most useful knowledge can be explored easily. Then, the knowledge interpretation must be in a format that is easy to modify according to the user's changing viewpoints. In addition, the knowledge interpretation must be able to automatically cluster relevant events together.

Based on the above observations, we use a weighted news topic relation network as the knowledge interpretation. The network uses news topics (i.e., keywords and keyframes) as nodes and their relations as edges, and the edges are weighted according to their *interestingness* for the users. If we use D to represent the database of interest and K_D to represent the knowledge interpretation of D (i.e., the weighted semantic network), K_D can be represent as:

$$K_D = \{(k_i = (s_a, s_b), w_U(k_i)) \mid 1 \le i \le N\} \qquad (1)$$

where k_i is a relation between a pair of news topics s_a and s_b, U is the user who is using the system, and $w_U(k_i)$ is the interestingness weight of k_i based on U's preference. An example of the network is given in Figure 2.

s_a and s_b are defined as related when they occur in a closed caption or automatic speed recognition (ASR) script sentence simultaneously. By collecting all these relations together, the semantics of the whole database can be represented. Therefore, we use the pairs of news topics, k_i, as the knowledge items to interpret the knowledge of video news databases. However, not all of these relations are interesting for the users. For example, the relation between "Bush" and "President" is not interesting because it is well known. On the contrary, the relation between "Iraq War" and "Gas Price" may be interesting for many users. To resolve this problem, we also compute an interestingness weight $w_U(k_i)$ for each knowledge item k_i. Our knowledge interpretation extraction algorithm will extract the interestingness weights for the knowledge items. The algorithm will be introduced in the next section.

The weighted news topic relation network is suitable for our system because of the following reasons. First, networks can be intuitively visualized, so that most information of the news topic relation network can be delivered to the users. Second, uninteresting knowledge can be suppressed by the interestingness weights used in the semantic network. Third, the interestingness weights can be easily modified to adapt to the user's changing viewpoint during the exploration. Finally, the analysis of closed caption sentences together with the news topic pairs provide strong relations and rich knowledge content, as shown in the networks.

It is therefore quite worthwhile to use the weighted news topic relation network as the format of the knowledge interpretation and to optimize all components of our knowledge visualization framework toward a single target. The semantic analysis algorithm is optimized to extract news topics and their relations for knowledge interpretation and knowledge visualization. The interestingness weighting algorithm is optimized to weight the news topics and their relations extracted from the semantic analysis algorithm so that common, uninteresting knowledge can be suppressed. The visualization interface is optimized to disclose as much about the
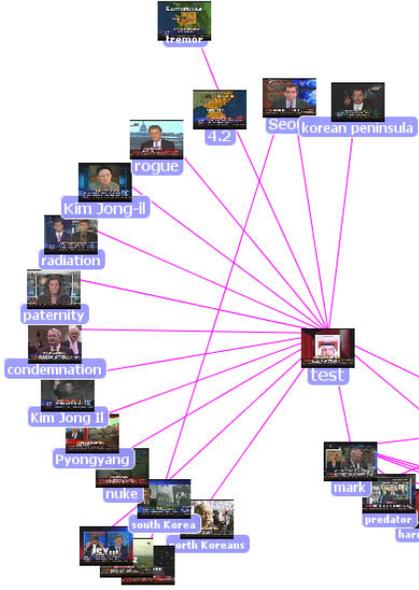
Figure 2: News topic relation network discloses interesting knowledge to the users. Links of the news topic "test" disclose details of the event and response of the international community during the North Korean nuclear weapon test.

Figure 3: Showing the entire news topic relation network is too messy. The database has news reports between Oct. 1, 2006 and Oct. 10, 2006 from CNN, FOX, and MSNBC.

news topic relation network as can be clearly shown. All these optimizations result in maximizing the amount of relevant knowledge delivered to the users during the exploration.

To implement this optimized system, algorithms from different research areas must be integrated. The following sections of this paper introduce these optimizations.

3   NEWS TOPIC RELATION NETWORK VISUALIZATION

The purpose of the visualization interface is to represent as much of the knowledge interpretation (i.e., the weighted news topic relation network) as possible to the users clearly and intuitively. To achieve this purpose, the visualization interface could show as many nodes and edges as possible to the users. However, because the network may be very large for large-scale video databases, it may be too messy to show the entire network to the users simultaneously. An example is given in Figure 3. The user may focus on a certain local detail at any time, so uninteresting detail should not get in the way and perhaps be removed. On the other hand, the points of interest for the user may change rapidly. The user may also need to explore from one part of the database to another part frequently. As a result, the global semantic context must be displayed at the same time as the interesting details are shown.

To enable users to examine the local details under the global semantic context, we use the hyperbolic browsing technique [6] to visualize the semantic network. The hyperbolic browsing technique lays out the network on a hyperbolic plane and then projects the hyperbolic plane to the 2D screen space. There is a nice property of the hyperbolic plane for network visualization: the space increases exponentially along the distance. Therefore, the hyperbolic plane is able to hold exponentially increasing nodes as the network expands in depth. When the hyperbolic plane is projected to the 2D screen space, an appropriate projection can be selected so that a fisheye effect can be automatically implemented to enlarge the nodes around the focus and shrink the nodes far from the focus. As a result, the local details of interest can be represented in the global semantic

context. An example is given in Figure 4. When a user wants to examine details at other points, she can simply drag the network to move the focus to a new position. As a result, the details of the network at any position can be checked in the global semantic context. And much more nodes can be represented simultaneously than without the hyperbolic interaction. Other visualization techniques could be used here as long as they retain the ability to see details in a global context. But efficient design requires that the technique be highly interactive and that relevant portions of the network be visible because they retain such rich knowledge content and potential for inference.

By using the hyperbolic browsing technique, hundreds or thousands of nodes and edges can be visualized on the screen simultaneously or navigated to coherently. However, the news topic relation network of a large-scale news video database may have tens of thousands of nodes and edges. Therefore, it's still untenable to display all the nodes and edges simultaneously. As a result, the network must be slimmed according to the user's current point of interest. To resolve this problem, we integrate an information retrieval algorithm in our system.

At any time of the exploration, the user can express her special interest by double clicking the nodes of the network. Consequently, the system knows that the user's new point of interest is the selected node, s_I. Then, the system must take two actions in response to the user input: (1) modify the network to disclose the relevant knowledge associated with s_I, and (2) retrieve the most relevant news reports from the database and present them to the user.

To perform Action (1), we must extract a new news topic relation network that is relevant to the user input but still preserves the global semantic context. To achieve this purpose, we "boost" the relevant knowledge items relevant to the user input on the global network by a relevance factor:

$$
\begin{aligned}
\acute{K}_D(s_I) &= \left\{ (k_i, w_U(k_i) \times \varpi(k_i, s_I)) \mid 1 \le i \le N \right\} \\
&= \left\{ \left(k_i, w_U(k_i) \times c^{\max\{r(s_a, s_I), r(s_b, s_I)\}} \right) \right\}
\end{aligned}
\tag{2}
$$
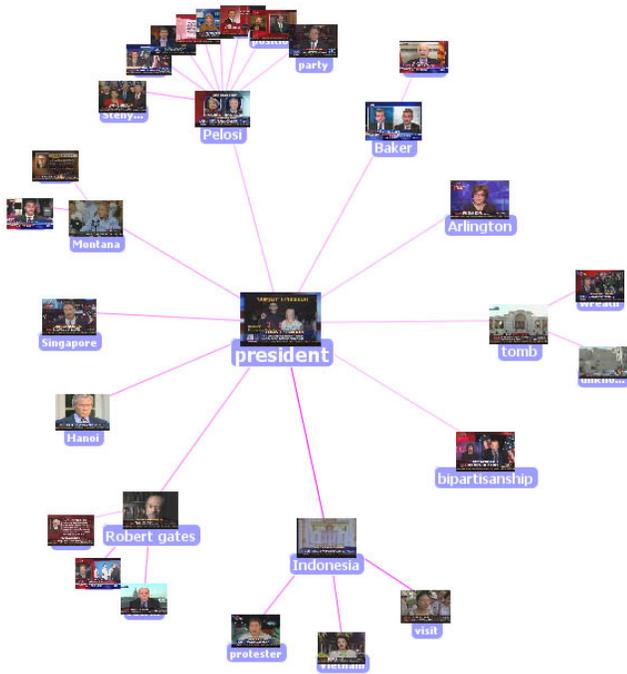
Figure 4: Hyperbolic visualization of the news topic relation network. Local details are embedded in the global semantic context.

(a) Global semantic network

(b) Focused semantic network for "schoolhouse"

Figure 5: Global and focused news topic relation network. On the bottom, "schoolhouse" is now the center and more relevant news topics are displayed and linked.

where s_I is the clicked item, $\varpi(k_i, s_I)$ is a boosting factor to emphasize items most relevant to s_I, $c \geq 1$ is the boosting constant, and $r(s_*, s_I) \in [0,1]$ is the relevance between s_* and s_I. By applying Eq. (2) to the global network, irrelevant knowledge items with $r(s_*, s_I) = 0$ stay unchanged and relevant knowledge items with $r(s_*, s_I) > 0$ have interestingness weights increased according to their relevance to the user input. As a result, more relevant knowledge items are selected for visualization on the new network. Constant c balances the local details and the global context. Larger c enables more local details to be included in the new network. Smaller c preserves more global context in the new network.

To compute $\acute{K}_D(s_I)$, $r(s_*, s_I)$ must be computed. Because the news topic relation network $\acute{K}_D$ represents the relevance quantities among news topics, $r(s_m, s_n)$ can be computed by exploring $\acute{K}_D$. Between a pair of news topics s_m and s_n, there may be several paths $p_x(s_m, s_n) = (s_m, ..., s_l, ..., s_n)$ on $\acute{K}_D$. The interestingness of $p_x(s_m, s_n)$ is defined as:

$$w_U(p_x(s_m, s_n)) = \min\left\{w_U\left(k_j = (s_a, s_b)\right)\right\} \quad (3)$$
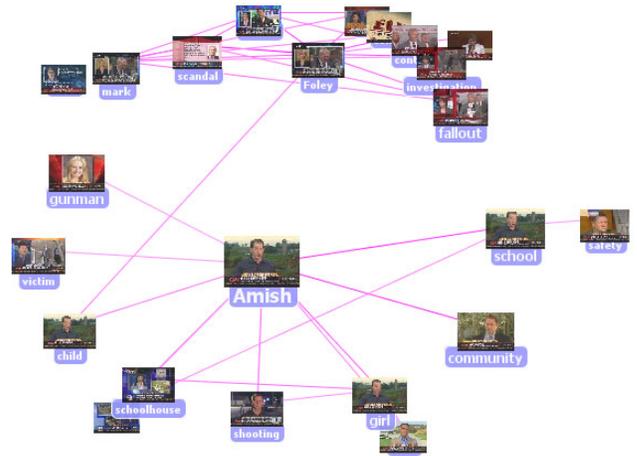
where k_j is a segment of $p_x(s_m, s_n)$. The shortest path is defined as:

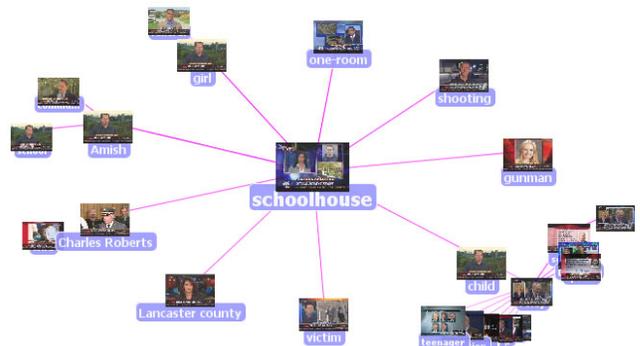$$p_{\min}(s_m, s_n) = \arg\max_x \left\{w_U(p_x(s_m, s_n))\right\} \quad (4)$$

The shortest path $p_{\min}(s_m, s_n)$ represents the most interesting route connecting s_m and s_n. Therefore, it is a good measure of the relevance between s_m and s_n:

$$r(s_m, s_n) = w_U(p_{\min}(s_m, s_n)) \quad (5)$$

By combining Eq. (5) into Eq. (2), a new semantic network $\acute{K}_D(s_I)$ can be generated, which is relevant to the user input s_I. In addition, relevant nodes are automatically laid out close to s_I. Relevant events can then be easily checked. Furthermore, the global semantic context is still preserved, so that the user can quickly switch to new point of interest if she changes her mind. An example is given in Figure 5.

By integrating the focused network, the global network can be slimmed down to keep only the most important news topic relation structure. As a result, the users can learn more high-level semantics and are free from messy fine details. The slimmed global network for the same database as Figure 3 is given in Figure 6. From Figure 6, the users can derive two major events immediately without further exploration: the Foley's scandal and the Amish school shooting. Figure 3 does not have this nice property.

Although the modified news topic relation network is able to disclose more details related to the user input, this is not enough for reasoning. To have best support for reasoning, the relevant original news reports must be retrieved and played for the users. Our system can retrieve from the news video database by using the user input s_I as the query. The 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, at the bottom of Figure 7(a). In addition, the most relevant web news is also retrieved, as shown in 7(b). This feature is very important for audiences who want to know more details and related discussions of the event. This combined process is a good reasoning technique for news video exploration and analysis.

4 KNOWLEDGE INTERPRETATION EXTRACTION

To implement the above visual interface, knowledge interpretations must be extracted from the database. As discussed in the previous section, an appropriate knowledge interpretation is composed of a set of knowledge items (i.e., keywords, keyframes, and their relations) and their interestingness, as in Eq. (1). We now describe how
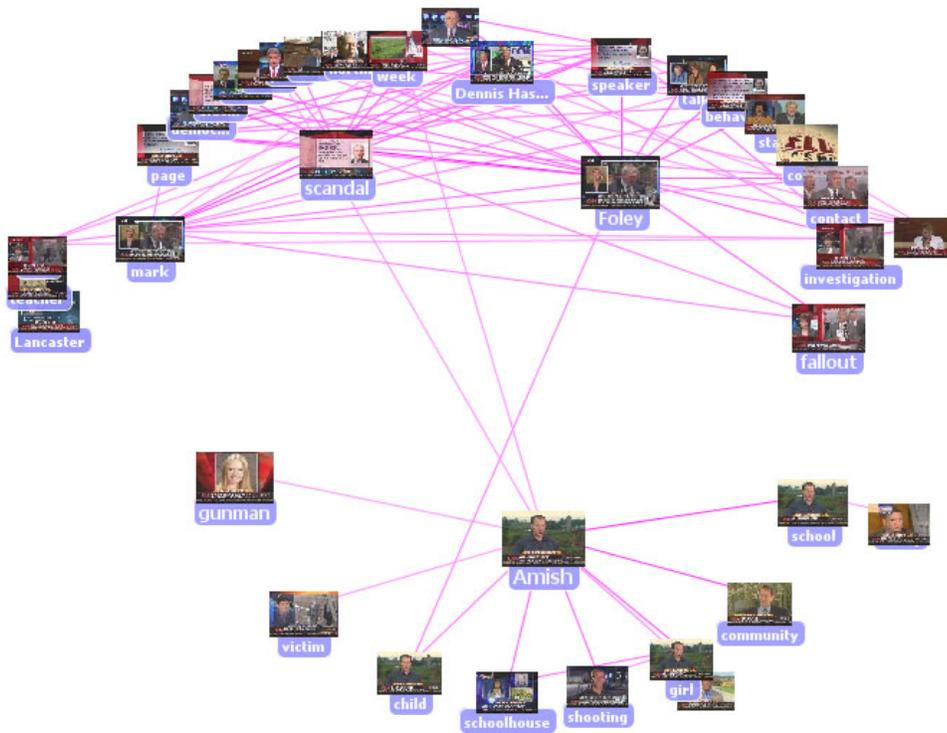
(a) Retrieval results by timeline

(b) Cross-media retrieval results

Figure 6: Slimmed news topic relation network of Figure 3.

Figure 7: An example of search results.

to compute the interestingness of knowledge items: $w_U(k_i)$.

The interestingness weight enables our knowledge interpretation to emphasize interesting knowledge and suppress uninteresting knowledge. However, it is a subjective quantity and cannot be quantified accurately without information from the user. Nevertheless, for news retrieval applications, the system generally needs to make recommendations and display something before users have any idea of the database. And for many applications involving large-scale video databases, disclosing the thoughts and and opinions of the public may be more important than satisfying the interestingness of the user. For example, a political consultant may want to know the real feeling of the public for a new policy. In this scenario, the personal preference of the consultant apparently should not be counted. Based on these observations, the system must provide a general interestingness measurement that is reasonable for most users.

To resolve this problem, we propose to use the provider behavior model to quantify the interestingness. Any news provider should have an interestingness measure for each news report. This measure is used to guide the production of news reports, taking into account such factors as story selection, sequence, length of each story, etc. If we can quantify these factors, the interestingness of news stories given by the provider can be quantified. One potential problem of the provider behavior model is that the interestingness measure from providers may be biased. The solution is to integrate multiple channels of large-scale data. The multi-channel database is able to smooth the individual biases and achieve unbiased or near unbiased solutions. Google has used the PageRank [11] technique to quantify the interestingness of web pages according to their links. The PageRank technique is a successful provider behavior model. Because of the great success of Google's search engine, we believe a provider behavior model will also work in large-scale news video database exploration applications.

To quantify the interestingness with the provider behavior model, we need to quantify the factors that news producers use to perform news editing. There are two types of factors can be quantified. The first type is the frequency $\delta(k_i)$ of knowledge item k_i. More impor-

tant news stories will certainly have a higher chance to be selected to broadcast by news producers. Thus, knowledge items repeated again and again on channel after channel may be more important than those that appear only one or two times. However, it may not be true all the time. For example, "George Bush $\Leftrightarrow$President" is one of the most frequent knowledge items in recent news reports. But users are seldom interested in this knowledge item because it's already well known. This implies that the interestingness $w_U(k_i)$ is inversely proportional to the user's prior knowledge, $\mu_U(k_i)$. Based on the above observations, $w_U(k_i)$ can be modeled by integrating the two quantities:

$$\begin{aligned} w_U(k_i) &\propto \delta(k_i) \\ w_U(k_i) &\propto \frac{1}{\mu_U(k_i)} \end{aligned} \Rightarrow w_U(k_i) = \gamma \frac{\delta(k_i)}{\mu_U(k_i)} \quad (6)$$

where γ is the normalization constant. Because the visualization algorithm uses only the relative ratios among $w_U(k_i)$, γ can be simply set to 1 or selected to optimize other targets.

In Eq. (6), $\delta(k_i)$ can be computed by statistical analysis on K_D:

$$\delta(k_i) = P_{K_D}(k_i) \quad (7)$$

where $P_{K_D}(k_i)$ is the probability of k_i in K_D. However, $\mu_U(k_i)$ is subjective and thus more complex to compute. In addition, U may have different prior knowledge of k_i at different times. Therefore, $\mu_U(k_i, t)$ must be used, where t is the time when U uses the system.

There are two factors that may affect $\mu_U(k_i, t)$: the learning factor and the forgetting factor. After k_i is reported by a provider m_l at time t_l, U may not learn it immediately. As time passes, the probability that U knows k_i increases. This means we can adopt a learning curve to compute $\mu_U(k_i, t)$. In addition, U may forget k_i if it is not repeated for a long time. By integrating both factors together, we can model the effect, $\rho_U(t, l)$, of one occurrence of k_i to U at time t (i.e. the time when the user employs the system) as:

$$\begin{aligned} \rho_U(t, l) &= \sigma_U(m_l)\, \varphi_U(t - t_l)\, \phi_U(t - t_l) \\ &= \sigma_U(m_l)\, g_U(t - t_l) \end{aligned} \quad (8)$$

where $\varphi_U(t-t_l)$ is the learning curve, $\phi_U(t-t_l)$ is the forgetting curve, l indicates the broadcast of k_i by provider m_l, σ_U is the efficiency of m_l, t_l is the time of l, and $g_U = \varphi_U \times \phi_U$ is the combined effect of learning and forgetting curves. Then, the prior knowledge of U at time t, $\mu_U(k_i,t)$, is the sum of all occurrence of k_i:

$$\mu_U(k_i,t) = \sum \sigma_U(m_l) g_U(t-t_l) \tag{9}$$

Eq. (9) can be represented as a convolution. If we define $f_U(t) = \sigma_U(m_l)$ when $t = t_l$ and $f_U(t) = 0$ otherwise, $\mu_U(k_i)$ is the convolution of f_U and g_U:

$$\mu_U(k_i,t) = \int f_U(t_l) g_U(t-t_l) dt_l = f_U \circ g_U(t) \tag{10}$$

For a specific exploration task, U may only focus on a relatively short period, such as one day, one week or one month. Therefore, K_D covers video news reports only in the period of interest. However, U may learn a knowledge item at any time. For example, "George Bush $\Leftrightarrow$ President" may have been learned for several years. As a result, $\mu_U(k_i,t)$ must be computed by using a database covering a much longer period than D. We name this database as $\acute{D}$. Consequently, the knowledge interpretation of $\acute{D}$ is $K_{\acute{D}}$. We use $K_{\acute{D}}$ to compute $\mu_U(k_i,t)$ in our system.

Eq. (10) gives a good model of the user's prior knowledge. Even though f_U and g_U are still subjective, they may be approximated. By adopting different approximations of f_U and g_U, both general and personalized knowledge extraction can be implemented. For large-scale video database exploration and investigation applications, general knowledge that reflects the thinking of a general user is more interesting than the personalized knowledge extracted for a particular investigator. To achieve this purpose, f_U and g_U must approximate the property of a general not a specific user. In this scenario, $g_U(t) = 1$ is a reasonable approximation. There are two reasons to support this point. First, the time interval from the happening of an event to the perception of the public is very short due to the responsiveness of the modern news industry. Because the time span of the database is much longer than this perception interval, φ_U can be treated as a step function. Second, because we use simple items for knowledge interpretation, users can remember them easily for a long time. In addition, the "refreshing interval" of these items may be much shorter than the time that users can remember them. As a result, $\phi_U = 1$ is a reasonable assumption. Using this reasoning and the property of convolution, $g_U(t) = 1$ is a suitable approximation for general knowledge extraction.

As the audiences have the freedom of selecting programs for watching, the rough average effect of the public is that all news providers may have about the same influence. This means $\sigma_U = 1$ is also a reasonable assumption. Certainly, particular news providers may influence different numbers of users. Consequently, more sophisticated models can be achieved by using market share, revenue or similar factors to compute σ_U.

Based on these observations, $f_U \circ g_U(t)$ is equivalent to the weighted frequency of k_i in $K_{\acute{D}}$. Weights for occurrences of k_i are σ_U. $\sigma_U = 1$ is a special case when all weights are equal. In our experiments we adopt $\sigma_U = 1$. As a result, $\mu_U(k_i,t) = P_{K_{\acute{D}}}(k_i)$. Based on this understanding we can compute the interestingness as:

$$w_U(k_i) = \gamma \frac{P_{K_D}(k_i)}{P_{K_{\acute{D}}}(k_i)} \tag{11}$$

To simplify the post process and fusing with other factors, γ is selected to normalize $w_U(k_i)$ to the range of $[0,1]$:

$$\frac{1}{\gamma} = \max_{k_i \in K_D} \left\{ \frac{P_{K_D}(k_i)}{P_{K_{\acute{D}}}(k_i)} \right\} \tag{12}$$

When personalized knowledge extraction is needed, f_U and g_U can be quantified by using usage history and user preference. For example, if U frequently watches CNN but seldom watches FOX News, we can assign higher weights to knowledge items from CNN. We can also assign a steeper learning curve to users watching news programs more frequently. Although personalized knowledge extraction is very interesting, how to implement it by using usage history and user preference information is complex. Therefore, we do not cover it in this paper and leave it for the future.

Video production rules can also imply the importance assigned by the news producers. To enable more efficient visualization of large-scale news video databases, important visual features should be considered. Video production rules generally do not have the "prior knowledge" problem. However these video production rules are difficult to extract because they are at high-level semantics. In addition, the keyframes, keywords and their relations are all semantic information. To have a complete large-scale news video database exploration system we need to extract these semantic interpretations from the database. The next section introduces the algorithm for semantic interpretation extraction.

5 SEMANTIC INTERPRETATION EXTRACTION

Not all news topics (i.e., keyframes and keywords) are equally interesting for the users. For example, "New York" is a keyword frequently mentioned in many news reports. Thus, many users may not be interested in it. As a result, the news topics also need to be weighted as we do for the knowledge items introduced in above section. In addition, video production rules related to news topics and knowledge items also need to be quantified so that better semantic interpretation and knowledge interpretation can be extracted.

However, 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 [12]. Nevertheless, supporting semantic video analysis plays an important role in enabling more efficient exploration of large-scale news videos. Without extracting the semantics 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 news topics (i.e., from video, audio, text) and video production rules automatically. Weights are assigned automatically with a statistical video analysis algorithm.

5.1 Semantic Video Analysis

The basic unit for news video interpretation is the video shot. Unlike the keywords of text documents, a video shot may carry abundant information (i.e., an image is more than a thousand words). This specific property of the video shot makes it difficult to effectively achieve statistical analysis of its visual properties and assign importance weights 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. The first one is the statistical properties of the shots. The second one is the special video objects in the shots. The last one is the semantic concepts that are associated with the shots. Because these three types of semantic units have different properties, different algorithms are needed to extract such multi-modal news topics.

5.1.1 Statistical Property Analysis of Physical Video Shots

The physical video shot is the basic unit for news video interpretation. Therefore, it can be used as a semantic item. However, unlike the keywords in text documents, the repetition of physical video shots cannot be detected automatically by using simple comparison

(a) Face objects   (b) Face confidence

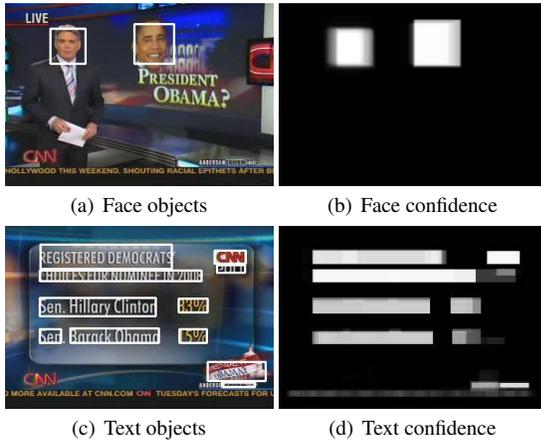(c) Text objects   (d) Text confidence

Figure 8: Video objects detection examples

between the shots. New techniques are needed for detecting the repeat of video shots in news videos [8].

News producers may repeat a certain shot in several ways. By detecting the repeat pattern of shots, we can infer the interestingness weights assigned by news producers. Consequently, we need to discriminate these patterns and assign appropriate weights to them. Through experiments, we found that most repeated shots can be weighted by an intra-program repetition weights and an inter-program repetition weights:

$$
\begin{aligned}
w_{intra}(i) &= e^{-\frac{\left(\frac{r_{intra}(i)-2}{2}\right)^2}{2}} \\
w_{inter}(i) &= e^{-\frac{\left(\frac{r_{inter}(i)-5}{8}\right)^2}{2}}
\end{aligned}
\tag{13}
$$

where $r_{intra}(i)$ is the intra-program repeating number of shot i, and consequently $r_{inter}(i)$ is the inter-program repeating number. More details can be found in [8].

5.1.2 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 [8]. Then the detected objects can be used to quantify the importance of the shot:

$$
\begin{aligned}
w_{textArea}(i) &= \frac{1}{1+e^{-\frac{\max\{\alpha_{text}(i)-v_{text},0\}}{\lambda_{text}}}} \\
w_{faceArea}(i) &= \frac{1}{1+e^{-\frac{\max\{\alpha_{face}(i)-v_{face},0\}}{\lambda_{face}}}}
\end{aligned}
\tag{14}
$$

where α is the ratio that the object is in the frame, v and λ are parameters determined by experiments. Video objects detection examples are given in Figure 8. More details can be found in [8].

By performing face clustering, face objects can be clustered to several groups and the human objects can be identified. The human object is similar to the knowledge items: too frequent items may not be interesting, such as the anchor person. Consequently, the same weighting algorithm introduced in Section 4 is adopted to compute the weight. The importance weight for human face of shot i is computed by:

$$
w_{face}(i) = \begin{cases} \max\limits_{x \in FACE(i)} \{w_U(x)\} & FACE(i) \neq \emptyset \\ 0.5 & FACE(i) = \emptyset \end{cases}
\tag{15}
$$

where $FACE(i)$ is the set of face objects of shot i, and $w_U(x)$ is the weight of x computed by using Eq. (11).

Table 1: Semantic Concept Importance

Concept $(C(i))$	$w_c(C(i))$	Concept	$w_c(C(i))$
Announcement	0.9	Report	0.3
Sports	0.5	Weather	0.5
Gathered People	1	Unknown	0.8

5.1.3 Semantic Video Classification

The semantic concepts of video shots can provide valuable information to enable more efficient and effective visualization and retrieval of large-scale news video collections. Semantic video classification is one method that helps detect the semantic concepts for the video shots. 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. First, the users may have different preferences for different semantic concepts. Therefore, a prior weight can be assigned to each semantic concept according to the user preference. We adopt a scheme that approximates the preference of the public, as assigned in Table 1. Where $C(i)$ is the semantic concept of i, and $w_c(C(i))$ is the weight assignment. Second, semantic concepts are similar to the knowledge items thus can be weighted by the algorithm of Section 4. Finally, the weight of semantic concept is determined by:

$$
w_{concept}(i) = w_c(C(i)) \times w_U(C(i))
\tag{16}
$$

where $w_c(C(i))$ is looked up from Table 1, and $w_U(C(i))$ is the weight of $C(i)$ computed by using Eq. (11).

5.1.4 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. Our purpose of weighting is to detect the existence of some visual properties and emphasize those shots with interesting visual properties. The existence of one visual property may be indicated by different visual patterns. For example, the repeat property may be represented by w_{intra} or w_{inter}. To ensure we detect the existence of interesting visual properties and capture the patterns we are looking for, we first use max operation to fuse weights for the same visual property:

$$
\begin{aligned}
w_{repeat}(i) &= \max\{w_{intra}(i), w_{inter}(i)\} \\
w_{object}(i) &= \max\{w_{faceArea}(i), w_{textArea}(i)\} \\
w_{semantics}(i) &= \max\{w_{face}(i), w_{concept}(i)\}
\end{aligned}
\tag{17}
$$

Where w_{repeat} measures the visual property of physical video shot repetition, w_{object} measures the visual property of salient objects, and $w_{semantics}$ measures the visual property of visual concepts.

Then the overall visual importance weight for a given video shot is determined by the geometric average of the three weights:

$$
w_{video}(i) = \sqrt[3]{w_{repeat}(i) \times w_{object}(i) \times w_{semantics}(i)}
\tag{18}
$$

5.2 Audio and Text Keywords Extraction

The keywords can be extracted from closed caption and ASR scripts. Advanced natural language processing techniques, such as named entity detection, coreference resolving, and part-of-speech (POS) parsing are used in our system to extract appropriate keywords for shots. More details can be found in [8]. Finally each shot is associated a set of keywords. The keyword weight of a shot is computed by:

$$
w_{keyword}(i) = \max_x \{w_U(x) \mid x \text{ is a keyword of } i\}
\tag{19}
$$

In Eq. (19) we use the proposed provider behavior model to weight each keyword.

With the keyword weight and the visual weight computed above, the overall weight for a given video shot is determined by averaging w_{video} and $w_{keyword}$:

$$w(i) = \gamma \times w_{video}(i) + (1 - \gamma) \times w_{keyword}(i) \qquad (20)$$

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

6 EXPERIMENTS

To evaluate the efficiency of our system, we compare our system with the state of the art news search engine, Google News. We ask users to evaluate the difficulty of answering several news related questions by using our system and Google News. Total 12 users participated the experiments. 10 of them are undergraduate students without any related background, and 2 of them are security experts. Half of the users evaluate our system first. Another half evaluate Google News first. Before a user evaluate our system, the user watches a two-minute introduction video. For each task, the users give out only the difficulty level to complete the task. The difficulty level is defined as a number between 1 and 10, where 1 is the lowest level and 10 is the highest level. The database used in the evaluation contains three channels (CNN, FOX, and MSNBC) of video news reports in the past month (which is October, 2006).

The first task is to *list several most important news reports in the past month*. The average difficulty level for Google News is 9.2, and that for our system is 4.5. Most users said Google News provides little help on completing this task. The two security experts said our system is very helpful to complete this task, and this task is typical for their everyday work.

The second task is to *summary the whole event of North Korean nuclear weapon test*. The average difficulty level for Google News is 6.6, and that for our system is 4.3. The users said that our system places relevant news topics immediately surrounding the point of focus, which is very helpful to figure out the rough aspects of the whole event.

The third task is to *answer when, where, why and how of the Amish school shooting*. The average difficulty level for Google News is 4.1, and that for our system is 6.7. Google News outperforms our system in this task. The most two important reasons given by the users are: (1) The keyword-based search technique of Google News is significantly better than ours; (2) It is much easier to extract fine details from the web news reports than from the video news reports.

Based on the above experiments, one can find that: (1) Our system provides valuable service when the users do not have detailed preference. (2) Sophisticated keyword-based search techniques perform better when the users have detailed preference and need to learn the fine details. Therefore, our system is able to guide the users to build their own preference effectively and efficiently. Then keyword-based search techniques can be adopted to disclose fine details after the system catches the user's fine preference.

7 CONCLUSIONS

In this paper, a large-scale video database exploration and analysis system is proposed by integrating novel algorithms of visualization, knowledge extraction, and statistical video analysis. By optimizing all components toward a single target, the proposed system achieves more effective and intuitive video database mining and exploration.

To implement our system, a knowledge interpretation extraction algorithm is created to extract interesting knowledge and suppress uninteresting knowledge. As a result, the users may find news reports of interest and moreover get overviews of the whole news space for any given time span without the burdensome of mining large volume of uninteresting and useless reports. The knowledge interpretation we presented is able to bridge the **interest gap**.

Experiments disclose that the proposed system is able to help the users build their own preference effectively and efficiently, which is very difficult with systems based on keyword-based systems, such as Google News. As a result, the most system needs to integrate our proposed techniques as the front end to capture the user preference, and keyword-based search techniques as the back end to disclose fine details.

REFERENCES

[1] Nevenka Dimitrova, Hongjiang Zhang, Behzad Shahraray, Lbrahim Sezan, Thomas Huang, and Avideh Zakhor. Applications of video-content analysis and retrieval. *IEEE Trans. on Multimedia*, 9(3):42 – 55, 2002.

[2] Jianping Fan, Hangzai Luo, and Ahmed K. Elmagarmid. Concept-oriented indexing of video database toward more effective retrieval and browsing. *IEEE Trans. on Image Processing*, 13(7):974–992, 2004. (IF: 2.715. Google Cite: 12. SCI Cite: 6).

[3] Amarnath Gupta and Ramesh Jain. Visual information retrieval. *Communications of the ACM*, 40(5):70–79, 1997.

[4] Susan Havre, Beth Hetzler, and Lucy Nowell. Themeriver: Visualizing theme changes over time. In *IEEE Symposium on Information Visualization (InfoVis)*, pages 115–123, 2000.

[5] Elizabeth G. Hetzler, Paul Whitney, Lou Martucci, and Jim Thomas. Multi-faceted insight through interoperable visual information analysis paradigms. In *IEEE Symposium on Information Visualization*, page 137, 1998.

[6] John Lamping and Ramana Rao. The hyperbolic browser: A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. *Journal of Visual Languages and Computing*, 7(1):33–55, 1996.

[7] Beitao Li and Kingshy Goh. Confidence-based dynamic ensemble for image annotation and semantic discovery. In *ACM Multimedia*, pages 195 – 206, 2003.

[8] Hangzai Luo, Jianping Fan, Jin Yang, William Ribarsky, and Shin'ichi Satoh. Exploring large-scale video news via interactive visualization. In *IEEE Symposium on Visual Analytics Science and Technology*, pages 75–82, 2006.

[9] Andrew Mehler, Yunfan Bao, Xin Li, Yue Wang, and Steven Skiena. Spatial analysis of news sources. *IEEE Trans. on Visualization and Computer Graphics*, 12(5):765–772, 2006.

[10] Milind R. Naphade and Thomas S. Huang. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Trans. on Multimedia*, 3(1):141–151, 2001.

[11] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. http://dbpubs.stanford.edu:8090/pub/1999-66.

[12] Arnold W.M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-base image retrieval at the end of the early years. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, December 2000.

[13] Russell Swan and David Jensen. Timemines: Constructing timelines with statistical models of word. In *ACM SIGKDD*, pages 73–80, 2000.

[14] Jarke J. van Wijk. Bridging the gaps. *Computer Graphics and Applications*, 26(6):6–9, 2006.

[15] Marcos Weskamp. Newsmap. http://www.marumushi.com/apps/newsmap/index.cfm.

[16] James A. Wise, James J. Thomas, Kelly Pennock, David Lantrip, Marc Pottier, Anne Schur, and Vern Crow. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *IEEE Symposium on Information Visualization (InfoVis)*, pages 51–58, 1995.