Integrating Multi-Modal Content Analysis and Hyperbolic Visualization for Large-Scale News Video Retrieval and Exploration

H. Luo[a], J. Fan[1b], S. Satoh[c], J. Yang[b], W. Ribarsky[b]

[a] *Software Engineering Institute, East China Normal University, Shanghai, China*
[b] *Department of Computer Science, UNC-Charlotte, Charlotte, USA*
[c] *National Institute of Informatics, Tokyo 101-8430, Japan*

Abstract

In this paper, we have developed a novel scheme to achieve more effective analysis, retrieval and exploration of large-scale news video collections by performing multi-modal video content analysis and synchronization. First, automatic keyword extraction is performed on news closed captions and audio channels to detect the most interesting news topics (i.e., keywords for news topic interpretation), and the associations among these news topics (i.e., contextual relationships among the news topics) are further determined according to their co-occurrence probabilities. Second, visual semantic items, such as human faces, text captions, video concepts, are extracted automatically by using our semantic video analysis techniques. The news topics are automatically synchronized with the most relevant visual semantic items. In addition, an interestingness weight is assigned for each news topic to characterize its importance. Finally, a novel hyperbolic visualization scheme is incorporated to visualize large-scale news topics according to their associations and interestingness. With a better global overview of large-scale news video collections, users can specify their queries more precisely and explore large-scale news video collections interactively. Our experiments on large-scale news video collections have provided very positive results.

Key words: Multi-Modal Content Analysis, Interestingness Assignment, Association Determination, Hyperbolic Visualization.

1. Introduction

The broadcast news videos have extensive influence and carry abundant information, different organizations and individuals are utilizing them for different purposes. For example, the government can boost up the morale by publishing positive reports. On the other hand, the terrorists can also use it to display their announcement (e.g., Al Jazeera), formulate their plans, raise funds, and spread propaganda. By watching and analyzing the international news reports, the intelligence agents can translate the raw news videos into useful information and the private investors can also evaluate the political, economic, and financial status for making good decisions. Unfortunately, watching large-scale news video collections could be very tedious. Due to the large number of broadcast channels, it is too expensive and time-consuming to hire people to do manual process, and such manual process of large-scale news video collections may also delay our response for critical news events. Therefore, there is a growing interest of developing more effective techniques for automatic news video analysis and exploration of large-scale news video collections. In addition, supporting automatic news video analysis and exploration is able to acquaint general audiences with a better global overview of large-scale news video collections, so that they can also harvest our research achievements. Unfortunately, most existing tools for automatic news video analysis and exploration still suffer from the following challenging problems.

The first problem is how to automatically *extract the video semantics* from news video clips and bridge the **semantic gap** [7, 17, 40] between the low-level visual features and the high-level human interpretation of video semantics. Most existing systems can only support automatic extraction of low-level visual features and search video clips via similarity matching according to their low-level visual features [1, 2, 5, 30, 33, 41, 42, 44, 51, 52]. On the other hand, most users may not be familiar with the low-level visual features and they can only express their information needs via high-level video concepts (i.e., keywords for video concept interpretation) [16, 17]. Due to huge diversities of visual properties and semantics in news videos, most existing tools for semantic video classification cannot directly be extended to achieve automatic analysis and exploration of large-scale news video collections [1, 2, 5, 9, 33, 41, 42, 44, 51, 52].

The second problem is how to *extract most significant news*

[1] Corresponding author. Tel.: 704-687-8556, FAX: 704-687-3516, E-mail: jfan@uncc.edu (J. Fan)

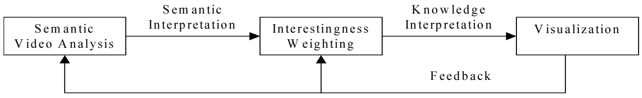


Fig. 1. The workflow of our proposed framework for analysis and exploration of large-scale news video collections.

topics from large-scale news video collections. This problem is becoming more critical because of the following reasons: (a) The amount of news topics could be very large because there are so many broadcast channels; (b) Different news topics may have different importance; (c) Different users may have different interestingness on the same news topic and there is no existing approach for user interestingness modeling. Thus there is an **interest gap** [47] between the available collections of news topics and the user's real information needs.

The third problem is how to *present the significant news topics to the users efficiently and effectively*. Large-scale collections of news videos may carry huge amount of information. However, it is very slow to exchange information between the user and the computer. Generally, a person can accept information at the speed of tens of bits per second. But our database has 10 GB to 1TB or larger scale. As a result, there is an **interaction gap** between the computer and the user.

Based on the above observations, we have developed a new framework with the workflow given in Figure 1. Firstly, visual semantic items for video content interpretation are automatically extracted from raw video clips by using our semantic video analysis technique. The keywords (e.g., text semantic items) for news topic interpretation and their associations (e.g., contextual relationships between the news topics) are extracted from the closed captions and audio channels, and they are further synchronized with the most relevant visual semantic items. Secondly, the interestingness measurements (weights) for these multi-modal (i.e., visual, auditory, and textual) semantic items are assigned simultaneously via statistical analysis. Finally, a novel hyperbolic visualization framework is incorporated to visualize large-scale news relation network (i.e., news topics and their associations) and support interactive news exploration, so that users can have better understanding of large-scale collections of news videos and make better query decisions. In addition, the users can also communicate with our system by providing their feedbacks, and such feedbacks can further be integrated to improve our algorithms for semantic video analysis and interestingness assignment.

Each step of the proposed framework bridge one of the three gaps as described above. Our proposed scheme is not a simple aggregation of the related techniques. It optimizes all steps toward the final goal. Therefore, the techniques used in our scheme may not be the best one in the relevant area, but it is the best suitable one for our proposed scheme. Firstly, because the vision is the fastest way to accept information for human beings, we adopt visualization to bring the interaction between the users and our system. The visualization technique can allow our system to present more information at the same time to the users than traditional techniques adopted by other retrieval or exploration systems, such as classified exploration and keyword-based search. Secondly, to fully utilize the capability of the visualization technique, we adopt a very small semantic unit (compared to the units adopted in other video semantic analysis and retrieval systems) to characterize the semantics and interestingness of the video news reports. On the one hand, it is more easy to detect and evaluate small semantic items, thus our algorithms for semantic video analysis and interestingness weighting can have higher performance. On the other hand, more accurate evaluation of the interestingness weighting enables more effective visualization. As a result, our proposed scheme may have better performance than aggregating the most sophisticated techniques in the relevant areas.

In this paper, we have developed multiple algorithms to address the following problems: (1) How to extract semantic items from large-scale news video collections that may have huge diversities of visual properties and semantics? (2) How to extract news topics and their associations? (3) How to assign the interestingness with the news topics and semantic items automatically? (4) How to achieve synchronization between multimodal news video contents? (5) How to visualize large-scale relation network (i.e., news topics and their associations) in a limited-size scene?

This paper is organized as follows: In section 2, we briefly review the most relevant works. In section 3, we introduce our algorithm for news topic detection and relation network generation. Section 4 describes our scheme for user interestingness modeling. Section 5 introduces our semantic video analysis and multi-modal video content synchronization algorithm. In section 6, we introduce our work on hyperbolic visualization of large-scale relation network. Finally, we conclude in Section 8.

2. Related Works

In the past, researchers have proposed some interesting approaches for bridging these gaps (i.e., semantic gap, interest gap and interaction gap). In this section, we give a brief review for some of these existing algorithms which are most relevant to our proposed work.

2.1. *Bridging Semantic Gap*

Semantic video classification is one promising solution to bridge the semantic gap [1, 2, 4, 5, 12, 14, 33, 41, 42, 44, 51, 52], but its performance largely depends on two inter-related issues: (1) suitable algorithms for video content representation and feature extraction; (2) effective algorithms for video classifier training.

To address the first issue, many approaches have been proposed for video content representation and feature extraction, and most existing approaches can be classified into three categories: (1) shot-based approaches that extract the global visual features from whole video shots [19, 36]; (2) region-based approaches that extract the local visual features from motion, color, or texture consistent homogeneous video regions [10, 18, 28]; and (3) object-based approaches that extract the representative visual features from semantic-sensitive video objects [24]. To enhance the power of the visual features on discriminating various video concepts, the underlying video patterns for feature extraction should be able to capture the intermediate video semantics at the object level effectively and efficiently (i.e., semantics for interpreting the real world physical objects in a video clip). Using the segmentation of real world physical objects (i.e., video objects) for feature extraction can significantly enhance the ability of the visual features on discriminating various video concepts. Unfortunately, automatic detection of large amounts of semantic video objects with diverse perceptual properties is still an open problem in computer vision. On the other hand, both the shot-based and the region-based approaches are easy for implementation, but their visual features may not be representative and cannot be used to discriminate various video concepts effectively. Consequently, there is an urgent need to develop new framework for video content representation and feature extraction, which is able to take the advantages for all these three existing approaches and avoid their shortcomings.

To address the second issue, robust video classifier training techniques are needed to model both the inter-concept variations and intra-concept similarity effectively. Two approaches are widely used to train video classifiers [1, 2, 4, 5, 12, 14, 20, 27, 33, 41, 42, 44, 51, 52]: (a) GMM-based approach uses Gaussian mixture models to approximate the underlying distributions of video data; (b) SVM-based approach uses support vector machines to maximize the margin between the positive samples and the negative samples. As a generative model, the major advantage of the GMM-based approach is that prior knowledge can be effectively incorporated into training suitable concept models for accurate video classification and annotation by predefining part of the model structure. Due to the diversity and richness of video contents, GMM models for video semantics interpretation may contain hundreds of parameters in a high-dimensional feature space, and thus large-scale labeled videos are needed for accurate classifier training. In addition, there may be mismatch between Gaussian functions and the real distributions of video data. On the other hand, the SVM-based approach enables more effective classifier training with smaller generalization error rate in high-dimensional feature space. However, searching the optimal model parameters (i.e., SVM parameters) is computationally expensive, and its performance is very sensitive to choices of kernel functions. Furthermore, automatic kernel function selection heavily depends on the implicit geometric property of the video data in the high-dimensional feature space. Because the high-dimensional feature space may be heterogeneous, it is difficult to select one common kernel function that can effectively characterize the implicit geometric properties of the video data. Another shortcoming of the SVM-based approach is that its training complexity depends on the number of training videos, and the prior knowledge (i.e., the correlations between video concepts) is not incorporated into SVM video classifier training.

The Informedia Digital Video Library project at CMU has achieved significant progresses on analyzing, indexing and searching of large-scale news video collections. A detailed survey can be found in [21]. Several applications have been reported, such as keyword-based video retrieval and query results visualization. The DELOS project also has significant progress on multiple areas including information access and personalization[11], object detection[3] and visualization[8]. Our proposed work has significant differences from these existing works: (a) Rather than performing semantic video classification to achieve automatic news video content understanding, multi-modal content analysis results from multiple information sources are synchronized and integrated for indexing and exploring large-scale news video collections. Even though many video classification algorithms have been proposed [1, 2, 4, 5, 12, 14, 33, 41, 42, 44, 51, 52], they cannot be used to achieve reliable classification of large-scale news video collections because of their huge diversities of visual properties and semantics. Most existing techniques for semantic video classification can perfectly work on same specific video domains with strong constraints of video contents. However, none of them can effectively handle the huge diversity of visual properties and semantics in the news videos. (b) An interestingness score is automatically assigned to each news topic via statistical analysis, and such interestingness scores are further used to decide the importance of the relevant news topics for filtering less interesting news topics. (c) The associations among the news topics are extracted for achieving more effective visualization and interactive exploration of large-scale news video collections. (d) A hyperbolic visualization tool is incorporated to acquaint the users with a better global overview of large-scale news video collections, so that they can explore large-scale news video collections interactively.

2.2. *Bridging Interest Gap*

To bridge the interest gap, visualization is widely used to help the users explore large amounts of information and find interesting parts interactively. In-spire [50] has been developed for visualizing and exploring large-scale text document collections, where statistics of news reports [31] is put on a world map to inform the audiences of the “hotness” of regions and the relations among the regions. TimeMine [45] is proposed to detect the most important reports and organize them through timeline with statistical models of word usage. Another system, called newsmap [49], can organize 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 relevant news topic. ThemeRiver [22] and ThemeView [23] can visualize a large collection of documents with keywords or themes. ThemeRiver and ThemeView can intuitively represent the distribution structure of themes and keywords in the collections.

Rather than recommending the most interesting news topics to the users, all of these existing news visualization systems prefer to disclose all the available information to the users, and thus the users have to dig out the interesting information by themselves. When large-scale news collections come into view, such available information could be very large and displaying all of them to the users may mislead them. To address this problem, some of these existing news visualization systems also disclose different distribution structures of large-scale news collections, but such distribution structures may not make any sense to the users because there is an interest gap between the distribution structures and the user's real information needs [47]. Therefore, it is very attractive to incorporate the interestingness of news topics for achieving more effective visualization and exploration of large-scale news video collections. Several researchers use predefined ontology to assistant visual content analysis and retrieval [13, 15, 46]. However, such pre-defined ontology is unacceptable for news content representation because the news content is highly dynamic. How to extract the semantic structure automatically from large-scale news video collections is still an open problem.

2.3. *Bridging Interaction Gap*

As addressed above, the vision is the fastest way for human beings to accept information. Thus supporting news video visualization is one potential solution for exploring large-scale news video collections. Therefore, visualization techniques are potential solution. To fully utilize user's vision capability, the adopted visualization technique must adapt to the human being's vision property. It is reported [26, 48] that a human being will focus on the details of a single point but still check the global context at the same time. Because the hyperbolic visualization technique [26, 48] implements the fish-eye effect with a uniform framework, which is very suitable for the goal of our system. However, the hyperbolic visualization technique is only a single layer exploration technique for a specific format of information. The information carried in large-scale news video collections may need hierarchical interactive exploration. Therefore, how to transform the data upon user input to implement user-adaptive hierarchical exploration is still an open problem.

In [29], a preliminary framework is proposed to resolve the above problems. This paper is the extension of [29]. In this paper, we developed the text term extraction and relation network generation algorithm, introduced a new visualization approach and proposed an algorithm to perform user-adaptive visualization. These extensions greatly improve the overall performance of our previous work [29].

3. Text Term Extraction and Relation Network Generation

For news programs, the closed captions can provide abundant information and such information can be used to detect news topics of interest with high accuracy. Rather than performing semantic video classification to obtain the news topics [1, 2, 5, 20, 27, 33, 41, 42, 44, 51, 52] (i.e., which can extract only a limited number of semantic concepts via multi-modal visual features), we have taken the advantage of cross-media to clarify the video contents and achieve automatic news topic detection from news closed captions. To do this, the closed captions are first segmented into sentences, and the text sentence is further segmented into keywords.

In news videos, 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 the content of the news report. Therefore, they are not appropriate for news semantics interpretation and should be removed. Because there may have clear and fixed patterns for these sentences, we have designed a context-free syntax parser to detect and mark this particular information. By incorporating 10-15 syntax rules, the parser can detect and mark such specific 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 the 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 sentences and recover the most capital information by using part-of-speech (POS) [39] and lemma information. We use TreeTagger [39] to perform the part-of-speech tagging. Capital information can be recovered automatically by using the TreeTagger parsing results.

After such specific sentences are marked and capital information is recovered, an open source text analysis package LingPipe [25] is used to perform the named entity detection and resolve co-reference of the named entities. The named entities referring to the same entity are normalized to a most representative format to enable statistical analysis, where the news model of LingPipe is used and all the parameters are set to default value.

Finally, the normalized results are parsed again by TreeTagger 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. All these detected keywords are treated as a basic vocabulary for interpreting the relevant news topics.

The relations among different news topics are also very important to enable more effective retrieval and exploration of large-scale news video collections. Because the news topics have already been extracted, the relations (associations) among these news topics can further be extracted according to their co-occurrence probabilities. Based on this observation, a weighted news topic relation network is used for knowledge interpretation. Based on this 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 collection of news videos of interest and K_D to represent

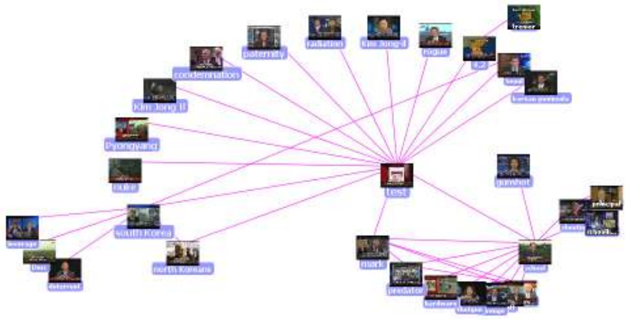


Fig. 2. Links of the semantic item “test” disclose details of the event and response of the international community during the North Korean nuclear weapon test.

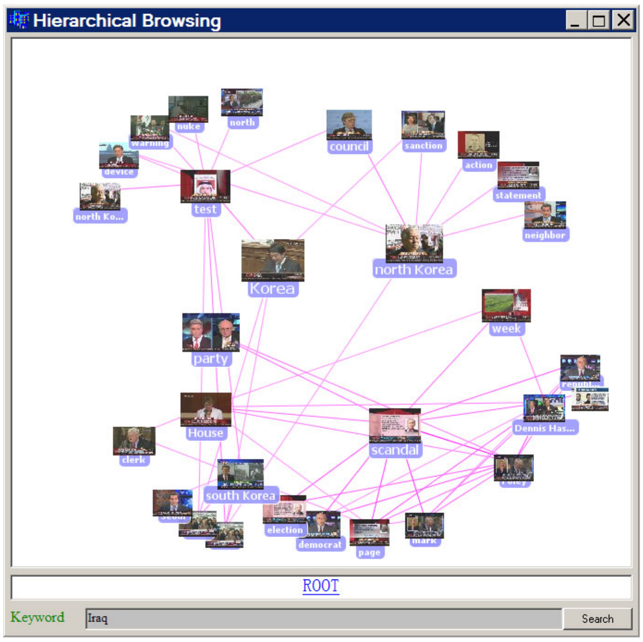


Fig. 3. North Korean nuclear weapon test event in the global news reports context.

the knowledge interpretation of D (i.e., the weighted relation network), K_D can be represent as:

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

where k_i is a relation (i.e., 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. A pair of news topics s_a and s_b is defined as a relation when they occur in a closed caption or automatic speech recognition (ASR) script sentence simultaneously. By collecting all these relations together, the semantics of the large-scale collections of news videos can be represented. Examples of the relation network are given in Figure 2 and Figure 3.

Even though the relation network is a good semantic representation of large-scale collections of news videos, not all of these relations are interesting for the 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 is introduced in the next section.

4. User Interestingness Modeling

The interestingness of the news topics and their relations must be quantified to implement our visualization system, but most existing visualization techniques [22, 49] use the raw frequency or probability to organize the visualization, and the raw frequency or probability may not indicate user's information need. For example, “Bush” is a keyword with very high frequency in 2006, but it may not be interesting because it is already well known. As the same reason, the relation between “President” and “Bush” is uninteresting for most users because it is already known. 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** of the news topics and their relations must be assigned for visualizing large-scale news video collections. Ideally, the interestingness is related to the user's preference and their prior knowledge, but it is very difficult to gather user preference information. Because the users may learn information from different sources that that our system may not know, such as friends. In addition, 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 is proposed in this paper.

Google has implemented a great search engine based on the PageRank [35] 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 for us to gather the provider behavior information. Thus we can quantify the interestingness of news topics by using the statistical information extracted from many TV channels. To quantify the interestingness of news topics 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 past news collection can be used to extract a general knowledge model of the user's knowledge. We use a probability distribution to represent the knowledge model:

$$G = \{g(x) \,|x \in S\} \tag{2}$$

where x is the given news topic or relation, S is the set of all the news topics and their relations, and $g(x)$ is the probability of the news topic or relation x. The general knowledge model can be used as a predictor. If a news topic or relation can be predicted completely, then it may not be interesting at all for the users. Only those news topics or relations that cannot be predicted well may be interesting for the users. According to information theory, the unpredictability of a message (i.e. news topics or their relations in our system) can be quantified by the information it carries. To quantify the amount of the information a news topic or relation carries, the local probability model of news topics or relations of news reports for a specific time interval of interest is defined as:

$$L(t) = \{l_t(x) \,|\, x \in S_t \subseteq S\} \tag{3}$$

where t is the specific time interval of interest, such as one specific day, S_t is the set of all the news topics or their relations in the specific time interval t and $l_t(x)$ is the probability of the given news topic or relation x in t. The difference between the local probability model $L(t)$ 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:

$$D(L(t) \parallel G) = \sum_{x \in S_t} l_t(x) \log \frac{l_t(x)}{g(x)} \tag{4}$$

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 news topic or relation $x \in S_t$. By examining Eq. (4), one can observe that $D(L(t) \parallel G)$ is the weighted sum of a set of components, and each component is only related to one single news topic x. Therefore, the contribution for one certain news topic $x \in S_t$ can be obtained by the relevant component of $D(L(t) \parallel G)$ in Eq. (4). Based on this observation, we can quantify the interestingness of x as:

$$w_t^r(x) = l_t(x) \log \frac{l_t(x)}{g(x)} \tag{5}$$

From Eq. (5), one can observe that the interestingness measurement $w_t^r(x)$ depends on two factors: $l_t(x)$ and $\frac{l_t(x)}{g(x)}$. The first factor $l_t(x)$ characterizes the probability of news topic x in local probability model $L(t)$ and the second factor $\frac{l_t(x)}{g(x)}$ characterizes the probability difference of x in between $L(t)$ and G (i.e., unpredictability). Because the purpose of Kullback-Leibler divergence is to measure the overall difference of $L(t)$ and G, as a result the unpredictability factor $\frac{l_t(x)}{g(x)}$ must be scaled by $l_t(x)$ in Eq. (4) so that $D(L(t) \parallel G)$ is adapt to the local distribution $L(t)$. However, our purpose is to measure the interestingness of individual news topic x. Therefore, the scale $l_t(x)$ is irrelevant to our purpose. Consequently, only the unpredictability factor $\frac{l_t(x)}{g(x)}$ should be used to characterize the interestingness measurement:

$$w_t^d(x) = \frac{l_t(x)}{g(x)} \tag{6}$$

Because multiple media channels (audio, video and closed caption text) are involved in the visualization, $w_t^d(x)$ in Eq. (6) needs to be normalized to simplify the multi-modal data fusion:

$$\overline{w}_t(x) = \frac{w_t^d(x)}{\max\limits_{x \in S_t} \left\{ w_t^d(x) \right\}} \tag{7}$$

Eq. (7) normalizes the interestingness weights to $[0, 1]$. The normalized weights are preferable to fuse with other factors. In our system, we use Eq. (7) to compute $w_U(k_i)$, i.e., $w_U(k_i) = \overline{w}_t(k_i)$. And Eq. (7) is a general weighting algorithm that can be applied to weigh news topics, relations and other visual properties such as video semantic concepts. Therefore, we use Eq. (7) to compute not only $w_U(k_i)$ but also interestingness weights of other items.

With the above algorithm for assigning the weights for the news topics and their relations, our system is able to extract the interesting news topics and suppress less interesting news topics. Because only statistical important semantic items can have high interestingness measurement, the random error of semantic analysis, which outputs a number of different incorrect semantic items, can be filtered out automatically and more robust results can be achieved.

To enable more efficient visualization of large-scale news video collections, special visual features must be considered, or else the results may not be visually important. There are multiple types of visual features that may be important. Some types of visual features can be processed by using Eq. (7), such as the people and the semantic concepts of news video shots. Other types of visual features may not be characterized by using the same statistical analysis algorithms as described above, such as video production rules. Based on this understanding, we have developed a series of semantic video analysis algorithm to extract the needed video semantics. These algorithms are introduced in the next section.

5. Semantic Video Analysis

Even though semantic video analysis and understanding are still very challenging, supporting semantic video analysis plays an important role in enabling more efficient retrieval and exploration of large-scale news video collections. Based on this observation, we have developed multiple solutions to automatically detect multi-modal semantic items (i.e., video, audio, text) and extract the representative visual features for video content representation. In addition, the interestingness weights for these semantic items are assigned automatically via our statistical video analysis algorithm.

5.1. *Statistical Property Analysis of Video Shots*

The video shots are the basic units for video content representation, and thus they can be treated as one of the semantic items for automatic weight assignment. However, unlike the keywords in text documents, the re-appearance of video shots cannot be detected automatically via simple comparison of their visual properties. Thus new techniques are desired for detecting the re-appearance of video shots in news videos [29], so that we can assign the importance weights of video shots automatically.

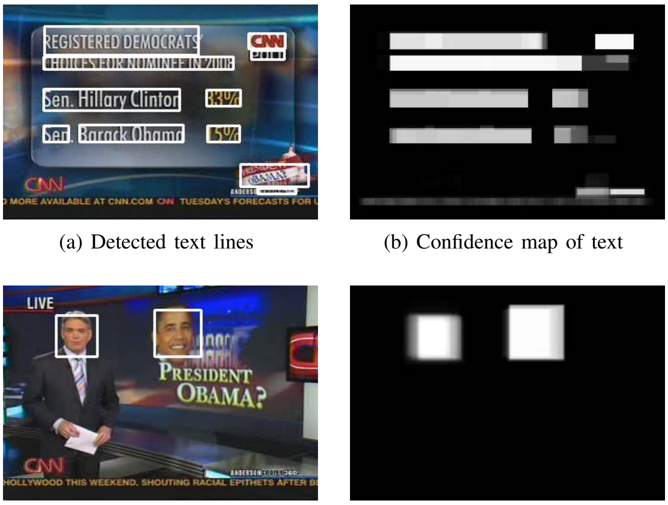
One certain video shot may be repeated multiple times because of the following reasons: (a) video shots for the anchors may repeat multiple times in the same news program; (b) video shots for the participants of an interview may appear multiple times in the same news program; (c) video shots for interpreting the important news may appear in both the news summary at the beginning and the detailed report later in the same program; (d) 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 re-appearance indicate the importance for the corresponding video shots. Nevertheless, the first two situations of video shot re-appearance may not indicate that the corresponding video shots are important. To quantify the effect of above four rules, the intra-program re-appearance number $r_{intra}(i)$ and inter-program re-appearance number $r_{inter}(i)$ for each video shot are computed. The two re-appearance 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 re-appearance patterns (i.e., different re-appearance situations) may provide different semantics and different weights should be assigned. Based on these understandings, the weights for different re-appearance numbers of video shots are approximated by using a bell shaped curve:

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

where ϱ_{intra} and ϱ_{inter} are the parameters. For intra-program repetition of video shots, Rules (a), (b) and (c) apply. However, the shots satisfying Rules (a) and (b) are not very attractive while the shots satisfying Rule (c) may be attractive. To discriminate the re-appearance pattern of Rule (c) with that of Rules (a) and (b), $\varrho_{intra} = 2$ is adopted in our experiments. As a result, anchor and interview shots re-appearing many times are suppressed because they are not "visually" important. On the other hand, news shots repeated in both the summary and the detailed report are emphasized, because their intra-program re-appearance number is exactly 2. For inter-program repetition, we found via experiments that little shot is repeated more than 5 times in different programs. In addition, if a news report is repeated channel by channel and program by program, it quickly becomes well known and therefore no longer "news". Therefore, we set $\varrho_{inter} = 5$ in our experiments.

5.2. *Video Objects Detection*

For news videos, video objects, such as 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 using suitable computer vision techniques [29]. 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 on all the video frames within the same video shots are integrated and the corresponding confidence maps for the detection results are calculated. As shown in Figure 4, such confidence maps can provide valuable information for evaluating our detection results.



(a) Detected text lines

(b) Confidence map of text

(c) Detected faces

(d) Confidence map of faces

Fig. 4. Automatic text and face detection results

The confidence region is generated by transforming the relevant confidences for our detection results into a binary image via thresholding. The threshold for generating the confidence region of text is set to the recall of the object detection algorithm. The reason of using recall as the threshold is that our object detection algorithm is tuned with high precision. Therefore, object regions will have confidence value higher or equal to the recall. Obviously, the size ratio between the confidence region and the size of video frames provides some valuable information for weight assignment. Therefore the size ratios for text and human faces regions are obtained, $\alpha_{text}(i)$ and $\alpha_{face}(i)$. However, the ratios cannot be directly used as the weight of the video shot. News editors may emphasize a person by putting a close-up, resulting in a large $\alpha_{face}(i)$. But a small face generally does not mean the person is unimportant, it only means the shot is an event shot. The text object is analog to the face object. As a result, the ratios must be converted by using a lower-bounded curve:

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

where ν is lower-bound ratio that news editors adopt for emphasizing information. For the face object, it generally equals the average ratio in anchor shots. For the text object, it generally equals the average ratio in shots without extra open captions (e.g., only with open captions overlayed on every shot). When α is less than ν, we cannot tell whether the news editor wants to emphasize the object or not. Therefore, the curve of Eq. (9) assigns the same weight to all these shots. For shots with α larger than ν, λ controls the conversion from α to the weight. To decide appropriate values for α and λ, we compute the $\alpha_{text}(i)$ and $\alpha_{face}(i)$ of all shots on a 10-hour news video database. We found that the average $\alpha_{face}(i)$ for anchor shots is close to 0.01 and average $\alpha_{text}(i)$ for shots without extra open captions is close to 0.05. Therefore, we adopt $\nu_{text} = 0.05$ and

$\nu_{face} = 0.01$. To compute λ, we select the control point of the conversion at 0.8, e.g., if a shot has a $\alpha(i)$ value that is larger than 80% of the shots (only counting shots with $\alpha(i) \geq \nu$), then it is assigned the weight 0.8. According to experimental data, $\lambda_{text} = 0.1593$ and $\lambda_{face} = 0.04096$. For a given video shot, the importance weight for human face area $w_{faceArea}(i)$ and the importance weight for text area $w_{textArea}(i)$ can be determined by:

$$\begin{aligned} w_{testArea}(i) &= \frac{1}{1+e^{-\frac{\max\{\alpha_{text}(i)-\nu_{text},\,0\}}{\lambda_{text}}}} \\ w_{faceArea}(i) &= \frac{1}{1+e^{-\frac{\max\{\alpha_{face}(i)-\nu_{face},\,0\}}{\lambda_{face}}}} \end{aligned} \qquad (10)$$

By performing face clustering [38], face objects can be clustered into several groups. As a result, the human objects can be identified and be treated as one of the semantic items for weight assignment. The face object is similar to text term. Therefore, they can be weighted by using news topics weighting algorithm as Eq. (7), which is introduced in Section 4. In addition, there may be more than one human object in a video shots. We just pick up the one with the highest weight as the representative for the video shot:

$$w_{human}(i) = \max\left\{\max_{x \in HUMAN(i)}\{\overline{w}_t(x)\},\ 0.5\right\} \qquad (11)$$

where $HUMAN(i)$ is the set of human objects of shot i. The first $\max\{\bullet\}$ lower-bounds the human weight at 0.5 for the same reason discussed above. $\bar{w}_t(\bullet)$ is the news topics weighting algorithm of Eq. (7). The second $\max\{\bullet\}$ ensures our algorithm to detect the most important person in the shot and ignore unimportant people.

5.3. *Semantic Video Classification*

The video concepts associated with the video shots can provide valuable information to enable more effective visualization and retrieval of large-scale news video collections, and semantic video classification is one promising solution to detect such video concepts. To detect the video concepts automatically, we have adopted our previous work reported in [17].

Two types of information about video concepts can be used for weight assignment. First, each video concept has an intrinsic importance. For example, a shot with a person reading an announcement is more important than a shot with a journalist introducing background information. The importance of the video concepts, $w_c(C(i))$, is determined by user study. Results are in Table 1. Where $C(i)$ is the video concept in the video shot i.

Second, the video concept can be treated as a text term. As a result, it can also be weighted as news topics by Eq. (7), which is introduced in Section 4. Finally, the weight for the given video concept is determined by:

$$w_{concept}(i) = w_c(C(i)) \times \overline{w}_t(C(i)) \qquad (12)$$

where $\bar{w}_t(\bullet)$ is the news topics weighting algorithm of Eq. (7).

Table 1
Video 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

5.4. *Semantic Visual Weights Fusion*

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 do not be misled by missing some visual patterns, we first use max operation to fuse weights for the same visual property:

$$\begin{aligned} w_{repeat}(i) &= \max\{w_{intra}(i), w_{inter}(i)\} \\ w_{object}(i) &= \max\{w_{faceArea}(i), w_{textArea}(i)\} \\ w_{semantics}(i) &= \max\{w_{human}(i), w_{concept}(i)\} \end{aligned} \qquad (13)$$

They are all derived from visual characteristics of video shots. Where w_{repeat} is the weight reflecting repetition of the shot, w_{object} reflecting prominent objects in the shot, and $w_{semantics}$ reflecting semantic categories. Consequently, the overall visual importance weight for a given video shot is determined by the geometric average of above three weights:

$$w_{video}(i) = \sqrt[3]{w_{repeat}(i) \times w_{object}(i) \times w_{semantics}(i)} \qquad (14)$$

5.5. *Multi-Modal News Content Synchronization and Decision Fusion*

For news programs, the closed captions have good matching with the relevant audios. Therefore, they can be integrated to take advantage of cross-media to clarify the video contents and remove the redundant information. The text documents for the closed captions may not synchronize with the video well and generally have a delay of a few seconds. Nevertheless, the audio generally synchronizes very well with the video but the accuracy of most existing techniques for automatic speech recognition (ASR) is still low. By integrating the results for automatic speech recognition with the results of closed caption analysis, the closed captions can be synchronized with the video contents with higher accuracy.

After the closed captions are synchronized to the relevant news 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 these 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 into keywords, as discussed in Section 3. Finally, all the video shots are associated with the relevant keywords in the same text sentence. After all shots have been associated with keywords, the keyword weight of a shot is computed by:

$$w_{keyword}\left(i\right) = \max_{x} \left\{\overline{w}_t\left(x\right) | x \text{ is a keyword of } i\right\} \qquad (15)$$

where $\bar{w}_t\left(\bullet\right)$ is the news topics weighting algorithm of Eq. (7), which is introduced in Section 4. The reason to use $\max\left\{\bullet\right\}$ to compute the overall keyword weight is that almost every shot is associated with several high-frequent uninteresting keywords, and really interesting keywords are generally rare. If any average algorithm is used, the calculated value is dominated by the low weights of high-frequent uninteresting keywords. The $\max\left\{\bullet\right\}$ operation ensures our algorithm to detect the existence of really interesting keywords robustly.

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 $w_{keyword}$:

$$w\left(i\right) = \gamma \times w_{video}\left(i\right) + \left(1-\gamma\right) \times w_{keyword}\left(i\right) \qquad (16)$$

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

6. Visualization for Large-scale News Video Exploration

After the news topics and semantic items are available, we introduce our visualization framework to enable more effective retrieval and exploration of large-scale news video collections. Because the available scene size for news video visualization is limited and there are large amount of interesting news topics, we need to answer three questions: (a) Which news topics are more interesting to users and should receive more space for display? (b) How can we display large amount of interesting news topics and their associations more effectively on a limited-size scene? (c) Which visualization method is the most suitable? (d) How can we visualize the changing trend of news topics along the time?

6.1. *Visualization of Relation Network*

As the news reports are unpredictable, both the general audiences and the news analysts may first want to have a global overview of all the available news reports. This means that the **global overview** is the first piece of information that is meaningful for most audiences, and such global overview is not necessarily formed by the whole news stories. It can be composed of the news topics with their interestingness weights and their associations. For providing the global overview of large-scale news video collections, the news topics and their associations are better than the whole news stories because the audiences may be interested in a certain small piece of information of a news report, such as a name or an interesting video shot.

The news topics can provide a good hint to the users, and thus the users can quickly make the decision of which news topic is more interesting than others according to their own preferences. By acquainting the users with a better global overview of large-scale news video collections, they can easily and intuitively specify their queries by clicking the relevant news topics interactively, and our system will return the most relevant news video clips which are strongly related to the corresponding news topics. Compared with the traditional keyword-based news video retrieval systems, our system can visually acquaint the users with: (a) keywords for news topic interpretation; (b) associations between the news topics; (c) interestingness weights to suppress the less interesting news topics. Because the less interesting news topics are suppressed and the unexpected news topics are emphasized, our proposed visualization algorithm can allow users to find the most interesting information easily. On the other hand, the associations among the news topics can also disclose interesting and meaningful information to the users.

For naive users to harvest our research achievements, it is very important to develop more comprehensive framework for visualizing such relation network of news topics, so that they can specify their queries easily and intuitively or explore large-scale news video collections interactively. Visualizing large-scale relation network in two-dimensional system interface is not a trivial task [26, 32, 34, 37, 43, 48]. We have developed a framework integrating multiple innovative ideas to tackle this issue effectively: (a) A tree-based approach is incorporated to visualize the relation network in a nested graph view, where each news topic is displayed along with its relevant ones. (b) The geometric closeness of the news topics on the visualization tree is related to their semantic relevance and associations, so that our graphical presentation can reveal a great deal about how these news topics are organized and how they are intended to be used. (c) Both geometric zooming and semantic zooming are integrated to adjust the level of visible detail automatically according to the discerning constraint on the number of news topics that can be displayed per visualization view.

Our approach for relation network visualization exploits hyperbolic geometry [26]. The hyperbolic geometry is particularly well suited to graph-based layout of large-scale relation network because of two reasons. First, the area of a circle on a hyperbolic plane rises exponentially in radius. Because the number of nodes of the relation network increases exponentially as the depth increases, these nodes can be put on a hyperbolic plane in equal density. Second, the hyperbolic plane can be projected to the Euclidean plane by Poincaré disk model [26] to implement a effect similar to the log-polar transformation in a uniform framework. This effect is essential to visualization because it matches the property of the human vision [6].

The essence of our approach is to project the relation network onto a hyperbolic plane according to the contextual relationships between the news topics, and layout the relation network by mapping the relevant news topics onto a circular display region. Thus our relation network visualization framework takes the following steps: (a) The news topics on the relation network are projected to a hyperbolic plane according to their associations, and such projection can usually preserve the original associations between the news topics. (b) After we obtain such context-preserving projection of the news topics, we can then use Poincaré disk model [26] to map the news topics on the hyperbolic plane to a 2D display coordinate. Poincaré disk model maps the entire hyperbolic space onto an open unit circle, and produces a non-uniform mapping of the news topics to the 2D display coordinate. The Poincaré disk model preserves the angles, but distorts the lines. The Poincaré disk model also compresses the display space slightly less at the edges, which in some cases can have the advantage of allowing a better view of the context around the center of projection.

Our implementation relies on the representation of the hyperbolic plane, rigid transformations of the hyperbolic plane and mappings of the news topics from the hyperbolic plane to the unit disk. Internally, each news topic on the graph is assigned a location $z = (x, y)$ within the unit disk, which represents the Poincaré coordinates of the corresponding news topic. By treating the location of the news topic as a complex number, we can define such a mapping as the linear fractional transformation [26]:

$$Z_t = \frac{\theta z + P}{1 + \overline{P}\theta z} \qquad (17)$$

where P and θ are complex numbers, $\mid P \mid < 1$ and $\mid \theta \mid = 1$, and $\overline{P}$ is the complex conjugate of P. This transformation indicates a rotation by θ around the origin followed by moving the origin to P (and $-P$ to the origin).

6.2. *Interactive Exploration of Large-Scale News Video Collections*

After the hyperbolic visualization of the relation network is available, it can be used to enable interactive exploration and navigation of large-scale news video collections at the topic level via change of focus. The change of focus is implemented by changing the mapping of the news topics from the hyperbolic plane to the display unit disk. The positions of the news topics on the hyperbolic plane need not to be altered during focus manipulation. Users can change their focus of news topics by clicking on any visible news topic to bring it into focus at the center, or by dragging any visible news topic interactively to any other location without losing the contextual relationships between the news topics, where the rest of the layout of the relation network transforms appropriately. Thus our hyperbolic framework for relation network visualization has demonstrated the remarkable capabilities for interactively exploring large-scale news video collections. By supporting change of focus, our hyperbolic visualization framework can theoretically display unlimited number of news topics in a 2D unit disk.

Moving the focus point over the display disk unit is equivalent to translating the relation network on the hyperbolic plane, such change of focus can provide a mechanism for controlling which portion of the relation network receives the most space. Through such change of focus on the display disk unit for relation network visualization and manipulation, the users are able to interactively explore and navigate large-scale news video archives. Therefore, users can always see the details of the regions of interest by changing the focus. Different views of the layout results of our relation network are given in Figures 5-8. By changing the focus points, our hyperbolic framework for relation network visualization can provide an effective solution for interactive exploration of large-scale news video collections at the topic level. 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 can disclose another level of knowledge to the users.

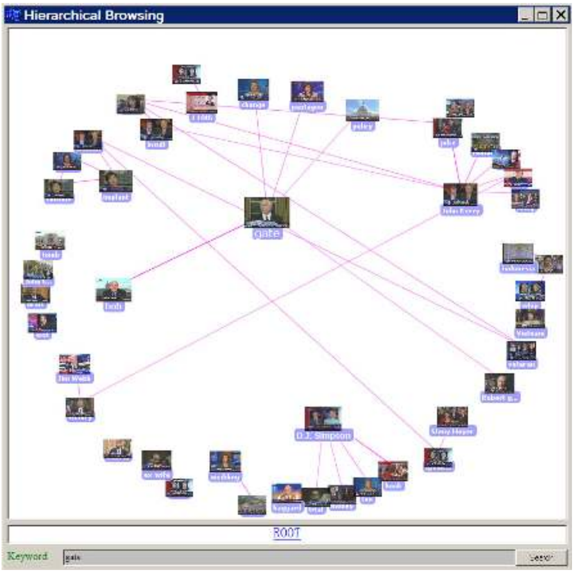


Fig. 5. Display emphasizing the top-left part.

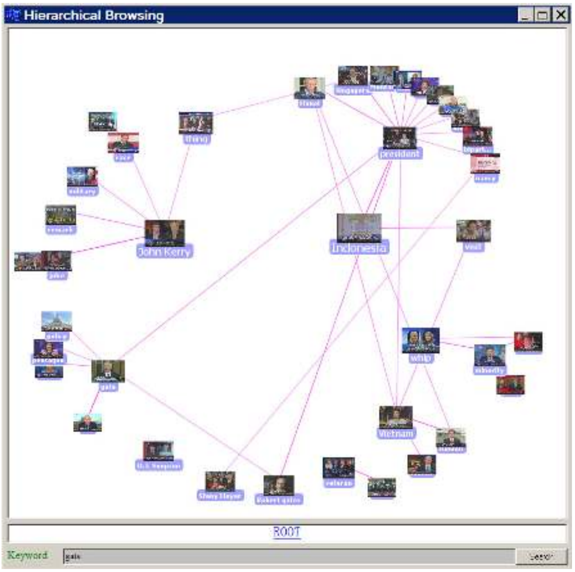


Fig. 6. Display emphasizing the top-right part.

6.3. *Intuitive Query Specification*

Our hyperbolic visualization of the relation network can acquaint the users with a good global view of the overall information of large-scale news video collections at the first glance, so that users can specify their queries visually because the relevant keywords for news topic interpretation and the most representa-

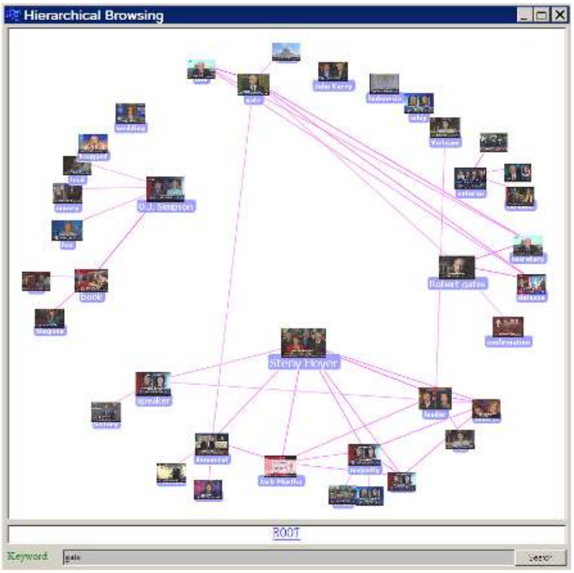


Fig. 7. Display emphasizing the bottom-left part.

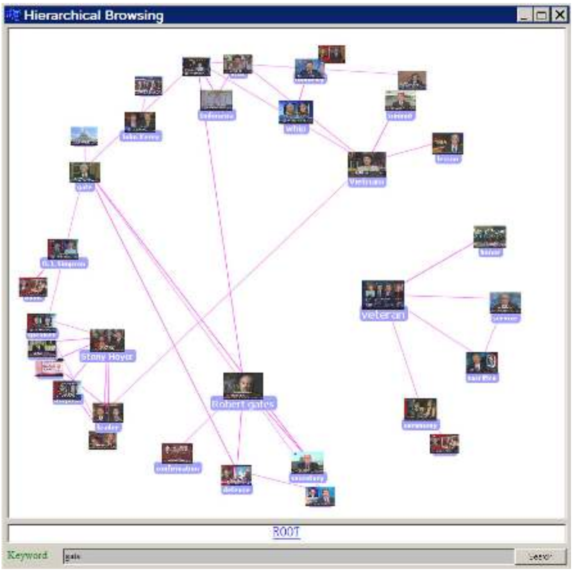
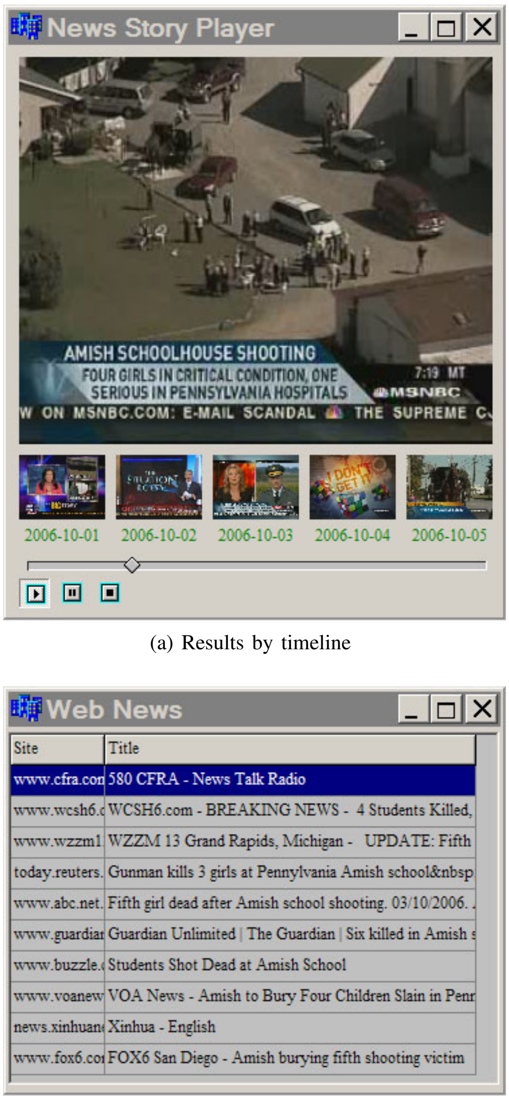


Fig. 8. Display emphasizing the bottom-right part.

tive keyframes are visible. In addition to such specific queries, our framework can allow users to start at any level of the relation network and navigate towards more details by clicking the relevant news topics of interest to change the focus. Therefore, our graphical visualization framework can significantly extend user's ability on video access and allow users to explore large-scale news video collections interactively at different levels of details. After the users click the relevant news topics, our system can then retrieve the news video databases according to the selected news topic 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 history of the whole event, as shown in Figure 9(a). In addition, the most relevant web news can also be retrieved, as shown in 9(b). This feature is very important for audiences who want to know more details and relevant discussions of the event, and


News Story Player
AMISH SCHOOLHOUSE SHOOTING
FOUR GIRLS IN CRITICAL CONDITION, ONE SERIOUS IN PENNSYLVANIA HOSPITALS
7:19 MT
MSNBC
W ON MSNBC.COM: E-MAIL SCANDAL THE SUPREME C
2006-10-01
2006-10-02
2006-10-03
2006-10-04
2006-10-05

(a) Results by timeline

Web News

Site	Title
www.cfra.con	580 CFRA - News Talk Radio
www.wcsh6.c	WCSH6.com - BREAKING NEWS - 4 Students Killed,
www.wzzm1	WZZM 13 Grand Rapids, Michigan - UPDATE: Fifth
today.reuters.	Gunman kills 3 girls at Pennylvania Amish school
www.abc.net.	Fifth girl dead after Amish school shooting. 03/10/2006.
www.guardiar	Guardian Unlimited | The Guardian | Six killed in Amish s
www.buzzle.c	Students Shot Dead at Amish School
www.voanew	VOA News - Amish to Bury Four Children Slain in Penr
news.xinhuan	Xinhua - English
www.fox6.coı	FOX6 San Diego - Amish burying fifth shooting victim

(b) Cross media retrieval results

Fig. 9. An example of search results.

thus our system can provide a good technique of reasoning.

6.4. *User-Adaptive Visualization of Relation Network*

After the users navigate the relation network for large-scale news video collections, they can specify their queries according to their interestingness. Therefore, the users can obtain more relevant news stories by submitting some specific queries. Consequently, 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:

$$\begin{aligned} \acute{K}_D\left(s_I\right) &= \left\{\left(k_i, w_U\left(k_i\right) \times \varpi\left(k_i, s_I\right)\right) \mid 1 \leq i \leq N\right\} \\ &= \left\{\left(k_i, w_U\left(k_i\right) \times c^{\max\{r(s_a,s_I),r(s_b,s_I)\}}\right)\right\} \end{aligned} \tag{18}$$

where s_I is the clicked item, $\varpi\left(k_i, s_I\right)$ 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. (18) 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 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 K_D. The interestingness of $p_x(s_m, s_n)$ is defined as:

$$w_U(p_x(s_m, s_n)) = \min\{w_U(k_j = (s_a, s_b))\} \tag{19}$$

where k_j is a segment of $p_x(s_m, s_n)$. The shortest path is defined as:

$$p_{\text{min}}(s_m, s_n) = \arg\max_x \{w_U(p_x(s_m, s_n))\} \tag{20}$$

The shortest path $p_{\text{min}}(s_m, s_n)$ 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:

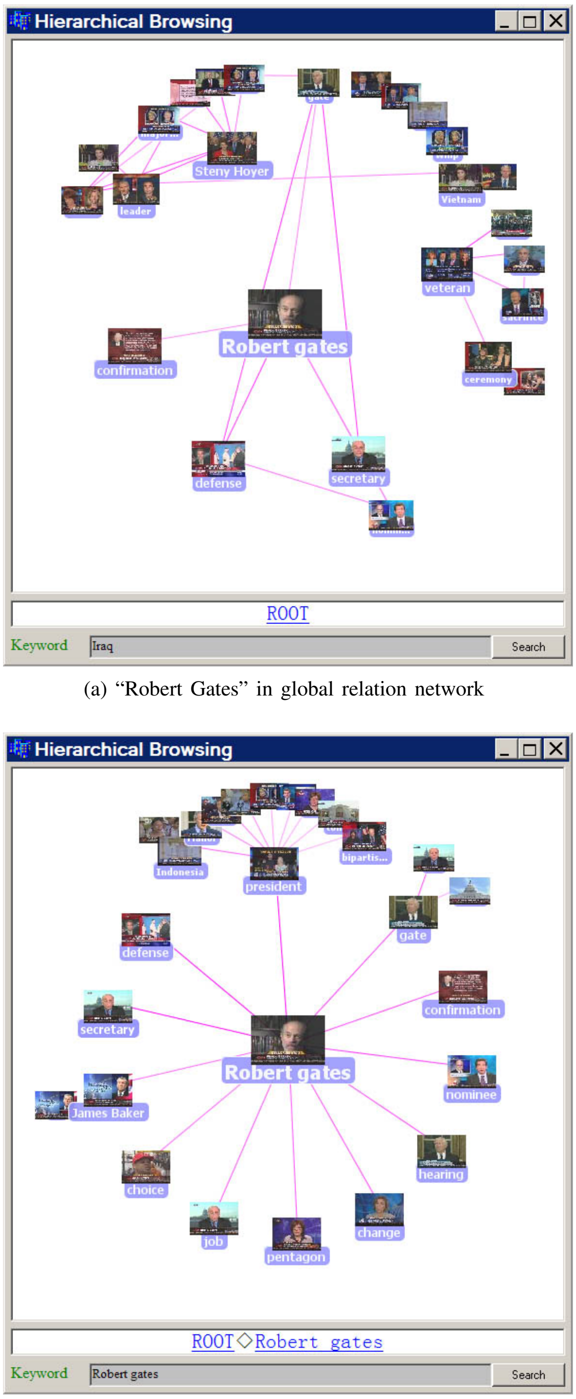
$$r(s_m, s_n) = w_U(p_{\text{min}}(s_m, s_n)) \tag{21}$$

By combining Eq. (21) into Eq. (18), 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. Examples are given in Figures 10-11. Such user-adaptive visualization of relation network can provide the users more details of the relevant news.

6.5. *News Changing Trend Visualization*

As the content of news reports is dynamic (i.e., news are changed along the time), the audiences may also want to see the **news changing trend over time**. Such dynamic news trend is able to tell the history of the whole news event to the users, and thus the users can have more complete view of the interesting topics with the dynamic trend.

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 12(a) and 12(d). The size of the keyframe in the map is proportional to the interestingness weight of the relevant news topic. By organizing the global 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 12(b) and 12(c). By watching the animation the users are able to catch the dynamic trend of the news topics.



Keyword Iraq

(a) “Robert Gates” in global relation network

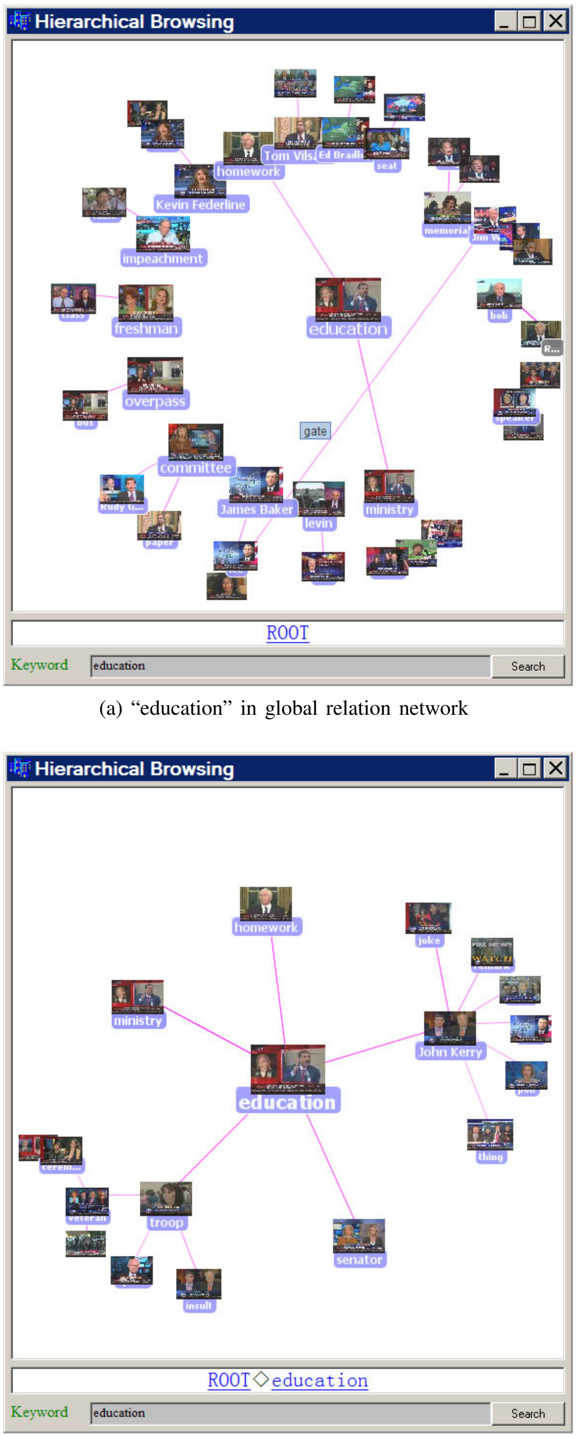
Keyword Robert gates

(b) Detailed view focusing on “Robert Gates”

Fig. 10. Adaptation example for query “Robert Gates”.

7. 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.



(a) "education" in global relation network

(b) Detailed view focusing on "education"

Fig. 11. Adaptation example for query "education".

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).

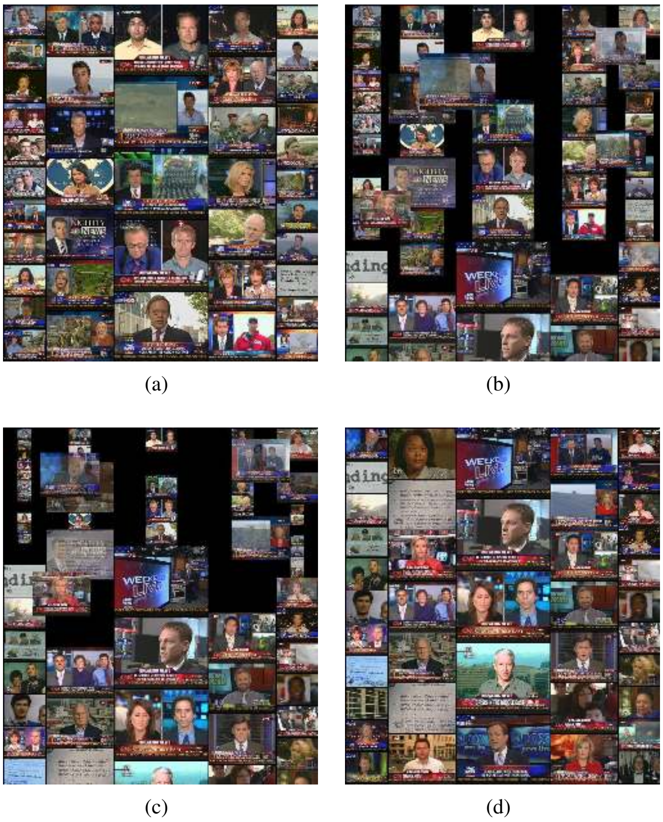


Fig. 12. An example of video news visualization. (a) Keyframes map for U.S. news on July 22, 2006; (b) and (c) Intermediate animation; (d) Keyframes map for U.S. news on July 23, 2006. The keyframes maps show the news topics on given day, the animation represents the trend of topic change over time. An example of animation can be downloaded at http://webpages.uncc.edu/~hluo/NewsDemo.avi.

The first task is to *list several most important news events 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 *summarize 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.

At February 2008, we perform a new experiment for January 2008 database. During this experiment, the average completion time is used as the evaluation criteria. Total 20 undergraduate students participate the experiment. 10 of them are asked to use Google News and another 10 are asked to use our system. A participant only use one of the two systems. None of them has used our system before.

The first task is to *find an interesting news and read it*. The average completion time for Google News is 0.9 minutes, and that for our system is 0.5 minutes. We asked the participants to press the "complete" button only when they read a really interesting news. Therefore, a user may read more than one news to complete this task. As a result, the average completion time is longer than the time to perform the first click.

The second task is to *list 5 most important news events in the past month*. The average completion time for Google News is 2.3 minutes, and that for our system is 5.5 minutes. The result is completely different than our expectation. After checking the answer sheets and inquiring the participants, we found that the users of Google News answered this question primarily by memory, while the users of our system did actually explore the database and check the news reports carefully. As a result, 4 of the Google News users listed less than 5 news reports, while all the users of our system listed 5 news reports.

The third task is to *summarize the first event listed in the second task.* The average completion time for Google News is 3.7 minutes, and that for our system is 3.3 minutes. Our system outperforms Google News slight. The difference is less than what we expected. We again inquired the participant and checked the experiment log, found that watching the video reports consumes more time than reading text reports for summarization purpose. Therefore, even though our system can help the users find relevant reports in shorter time, the users need to use more time to watch the reports. If we exclude the time for reading or watching by using the experiment log, the average completion time becomes 1.5 minutes for Google News users v.s. 0.8 minutes for users of our system.

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.

8. Conclusions

In this paper, we have developed a novel framework to achieve more effective analysis, retrieval and exploration of large-scale news video collections. By integrating cross-media information from multiple sources and synchronizing multimodal content analysis results, our proposed schemes can achieve more effective news topic detection and interestingness assignment and bridge the semantic gap successfully. By incorporating hyperbolic visualization for relation network visualization, our system can also support more effective retrieval and exploration of large-scale news video collections. Our experiments on large-scale news video collections have provided very positive results.

9. Acknowledgment

This work was sponsored by the National Visualization and Analytics Center (NVAC^{TM}) 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] Brett Adams, Chitra Dorai, and Svetha Venkatesh. Towards automatic extraction of expressive elements from motion pictures:tempo. *IEEE Trans. on Multimedia*, 4(4):472–481, 2002.

[2] W. H. Adams, Giridharan Iyengar, Chingyung Lin, Milind Naphade, Chalapathy Neti, Herriet Nock, and John R. Smith. Semantic indexing of multimedia content using visual, audio and text cues. *EURASIP Journal on Applied Signal Processing*, 2003(2):170–185, 2003.

[3] G. Antini, S. Berretti, A. Del Bimbo, and P. Pala. 3d face identification based on arrangement of salient wrinkles. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2006)*, pages 85–88, 2006.

[4] Thanos Athanasiadis, Phivos Mylonas, Yannis Avrithis, and Stefanos Kollias. Semantic image segmentation and object labeling. *IEEE Trans. On CSVT*, 17(3):298–312, 2007.

[5] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M.I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[6] Alexandre Bernardino and José Santos-Victor. Binocular visual tracking: Integration of perception and control. *IEEE Transactions on Robotics and Automation*, 6:15, 1999.

[7] H. Le Borgne, A., Guerin-Dugue, and N.E. O'Connor. Learning midlevel image features for natural scene and texture classification. *IEEE Trans. On CSVT*, 17(3):286–297, 2007.

[8] H. Bruce, B. Cleal, R. Fidel, and A.M.Pejtersen. A multidimensional approach to the study of human-information interaction: a case study of collaborative information retrieval. *Journal of the American Society for Information Science and Technology*, 55(11):939–953, 2004.

[9] D. Bulgarelli, C. Grana, R. Vezzani, and R. Cucchiara. A semi-automatic video annotation tool with mpeg-7 content collections. In *Proceedings of IEEE International Symposium on Multimedia (ISM2006)*, pages 742–745, 2006.

[10] Yixin Chen and James Z. Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5:913–939, 2004.

[11] S. Christodoulakis and C. Tsinaraki. A multimedia user preference model that supports semantics and its application to mpeg 7/21. In *Proceedings of the Multimedia Modeling 2006 Conference (MMM 2006)*, pages 35–42, 2006.

[12] Saman Cooray, Noel O'Connor, Sean Marlow, Noel Murphy, and Thomas Curran. Semi-automatic video object segmentation using recursive shortest spanning tree and binary partition tree. In *Workshop on Image Analysis For Multimedia Interactive Services*, 2001.

[13] S. Dasiopoulou, C. Doulaverakis, V. Mezaris, I. Kompatsiaris, and M.G. Strintzis. *Semantic-Based Visual Information Retrieval*, chapter An Ontology-Based Framework for Semantic Image Analysis and Retrieval. Idea Group Inc., 2007.

[14] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. K. Papastathis, and M. G. Strintzis. Knowledge-assisted semantic video object detection. *IEEE Trans. On CSVT*, 15(10):1210–1224, 2005.

[15] S. Dasiopoulou, C. Saathoff, Ph. Mylonas, Y. Avrithis, Y. Kompatsiaris, S. Staab, and M.G. Strintzis. *Semantic Multimedia and Ontologies: Theory and Applications*, chapter Introducing Context and Reasoning in Visual Content Analysis: An Ontology-based Framework. Springer-Verlang, 2007.

[16] Jianping Fan, Yuli Gao, and Hangzai Luo. Multi-level annotation of natural scenes using dominant image components and semantic image concepts. In *ACM Multimedia*, pages 540–547, 2004.

[17] 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.

[18] Julien Fauqueur and Nozha Boujemaa. Region-based image retrieval: Fast coarse segmentation and fine color description. *Journal of Visual Languages and Computing*, 15(1):69–95, 2004.

[19] Kingshy Goh, Beitao Li, and Edward Y. Chang. Semantics and feature discovery via confidence-based ensemble. *ACM Trans. on Multimedia Computing, Communications, and Applications*, 1(2):168 – 189, 2005.

[20] Alex Hauptmann and Michael Smith. Text, speech, and vision for video segmentation: The informedia project. In *AAAI Fall 1995 Symposium on Computational Models for Integrating Language and V*, 1995.

[21] Alexander G. Hauptmann. Lessons for the future from a decade of informedia video analysis research. In *International Conference on Image and Video Retrieval (CIVR)*, volume LNCS 3568, pages 1–10, 2005.

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

[23] 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.

[24] Derek Hoiem, Rahul Sukthankar, Henry Schneiderman, and Larry Huston. Object-based image retrieval using the statistical structure of images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 490–497, 2004.

[25] Alias i Inc. Lingpipe. http://www.alias-i.com/lingpipe/.

[26] 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.

[27] Wei-Hao Lin and Alexander Hauptmann. News video classification using svm-based multimodal classifiers and combination strategies. In *ACM Multimedia*, 2002.

[28] Tatiana Louchnikova and Stephane Marchand-Maillet. Flexible image decomposition for multimedia indexing and retrieval. In *SPIE Internet Imaging*, pages 203–211, 2002.

[29] 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.

[30] J. Maydt and R. Lienhart. An extended set of haar-like features for rapid object detection. In *Proceedings of the International Conference on Image Processing (ICIP 2002)*, volume 1, pages 900–903, 2002.

[31] 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.

[32] Baback Moghaddam, Qi Tian, Neal Lesh, Chia Shen, and Thomas S. Huang. Visualization & user-modeling for browsing personal photo libraries. *International Journal of Computer Vision*, 56:109 – 130, 2004.

[33] Milind Ramesh Naphade, Igor V. Kozintsev, and Thomas S. Huang. Factor graph framework for semantic video indexing. *IEEE Trans. on CSVT*, 12:40–52, 2002.

[34] G.P. Nguyen and M.Worring. Similarity based visualization of image collections. In *AVIVDiLib'05*, 2005.

[35] 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.

[36] Yossi Rubner and Carlo Tomasi. Texture-based image retrieval without segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1018–1024, 1999.

[37] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. A metric for distributions with applications to image databases. In *IEEE ICCV*, 1998.

[38] Shin'ichi Satoh and Norio Katayama. An efficient implementation and evaluation of robust face sequence matching. In *International Conference on Image Analysis and Processing*, pages 266–271, 1999.

[39] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[40] Arnold W.M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-base im-

age retrieval at the end of the early years. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, December 2000.

[41] Cees G. M. Snoek, Marcel Worring, and Alexander G. Hauptmann. Learning rich semantics from news video archives by style analysis. *ACM Trans. on Multimedia Computing, Communications, and Applications*, 2:91–108, 2006.

[42] Cees G.M. Snoek, Marcel Worring, Jan-Mark Geusebroek, Dennis C. Koelma, Frank J. Seinstra, and Arnold W.M. Smeulders. The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. *IEEE Trans. on PAMI*, 28:1678–1689, 2006.

[43] Daniela Stan and Ishwar K. Sethi. eid: a system for exploration of image databases. *Information Processing and Management*, 39:335 – 361, 2003.

[44] G. Sudhir, John C. M. Lee, and Anil K. Jain. Automatic classification of tennis video for high-level content-based retrieval. In *CAIVD '98*, 1998.

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

[46] David Vallet, Pablo Castells, Miriam Fernandez, Phivos Mylona, and Yannis Avrithis. Personalized content retrieval in context using ontological knowledge. *IEEE Trans. On CSVT*, 17(3):336–346, 2007.

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

[48] Jörg A. Walter and Helge Ritter. On interactive visualization of high-dimensional data using the hyperbolic plane. In *ACM SIGKDD*, 2002.

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

[50] 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.

[51] Yi Wu, Edward Y. Chang, Kevin Chen-Chuan Chang, and John R. Smith. Optimal multimodal fusion for multimedia data analysis. In *ACM Multimedia*, 2004.

[52] Wensheng Zhou, Asha Vellaikal, and C. C. Jay Kuo. Rule-based video classification system for basketball video indexing. In *ACM Multimedia*, 2000.