Exploring Large-Scale Video News via Interactive Visualization

Hangzai Luo[*]
Department of Computer Science
UNC-Charlotte
Charlotte, NC, USA

Jianping Fan[†]
Department of Computer Science
UNC-Charlotte
Charlotte, NC, USA

Jing Yang[‡]
Department of Computer Science
UNC-Charlotte
Charlotte, NC, USA

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 visualization framework to enable more effective visual analysis of large-scale news videos, where keyframes and keywords are automatically extracted from news video clips and visually represented according to their interestingness measurement to help audiences find news stories of interest at first glance. A computational approach is also developed to quantify the interestingness measurement of video clips. Our experimental results have shown that our techniques for intelligent news video analysis have the capacity to enable more effective visualization of large-scale news videos. Our news video visualization system is very useful for security applications and for general audiences to quickly find news topics of interest from among many channels.

Keywords: News Visualization, Semantic Video Classification

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

1 Introduction

The news industry is a large and profitable industry. Hundreds of news companies and producers report a tremendous amount of news stories in the format of text, audio or video everyday. Much of this content, including news from many countries, is now available to viewers via cable TV or Internet videocast. Intelligent analysis of such news stories can provide valuable information on science, politics, business and even military matters. On one hand, general audiences can have fun by reading, listening to or watching interesting news stories. On the other hand, security experts and business analysts can gather necessary information from news stories for decision making. Not only are news videos publicly available and rich in information, but international TV news provides a global overview of all news reports, which can represent important information about public opinion, discussions, and real thoughts from other countries in the world. Obviously, this information could be very useful for intelligence analysis, investment decision making, political or cultural analysis, and many other uses.

Due to the large amounts of video generated every day, discovering and analyzing news stories of interest is becoming an increasing problem. In news analysis, for example, it is becoming untenable to hire people to manually process all available news videos and produce summarization for them. Manual analysis of large-scale news videos is too expensive, and it may take long time for response. In addition, the manual summarization is generally biased and may mislead decision makers who use these summaries. On the other hand, providing summarization and visualization of large-scale news videos is also very important for general audiences, i.e., to save them time in searching and reading news of interest. Based on these observations, there is an urgent need to develop new techniques for: (a) intelligent analysis of large-scale news videos to extract the news stories of interest; (b) more effective visualization of news video collection.

Targeting these requirements, some researchers have developed keyword-based news retrieval systems and these techniques have been widely used in news industry (i.e., news websites). However, the keyword-based news retrieval systems make the big assumption that the users have clear ideas about what are looking for, but this may not be true. The news video database has the following properties: (1) Dynamic, e.g. changing every day; (2) The topics have large diversity; (3) Unpredictable (or else it's no longer news). In addition there are broad correlations among events. For example, sales may be affected by new products, interest rate changes, new political policies or even large foreign investments. Thus neither experts nor general users may have pre-specified preferences when they explore the news database. They may not be able to decide what relevant keyword to search for because there are too many potential keywords. Some news systems have provided another approach for news retrieval by classifying the news stories into a set of categories for browsing, but they also suffer from the same problem as the keyword-based systems, i.e., the users do not know what the news stories of interest are during that day or time period. In addition, the lack of preference beforehand does not mean the users will have interest in every news story provided by the system. For a certain user, he/she may just have interest in a few news stories among thousands of available. Unfortunately, most existing systems for news retrieval, filtering, ranking and summarization have not explored user attention models to enable more effective news stories organization, indexing and visualization.

1.1 Our Contributions

Based on these observations, we have developed a novel framework for news video analysis and visualization. Our framework has the following **advantages**: (a) First, a new algorithm for statistical video analysis is developed to extract the news stories of interest from large-scale TV news collections. (b) Second, a user attention model is developed by automatically assigning importance weights for the available news stories of interest. (c) Third, a novel news video visualization framework is developed to organize and visualize large-scale news stories more effectively and attract the

[*] e-mail: hluo@uncc.edu
[†] e-mail:jfan@uncc.edu
[‡] e-mail:jyang13@uncc.edu
[§] e-mail:rebarsky@uncc.edu
[¶] e-mail:satoh@nii.ac.jp

IEEE Symposium on Visual Analytics Science and Technology 2006
October 31 - November 2, Baltimore, MD, USA
1-4244-0592-0/06/$20.00 © 2006 IEEE

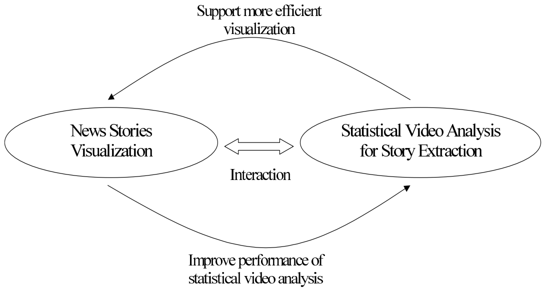

Support more efficient visualization
News Stories Visualization
Statistical Video Analysis for Story Extraction
Interaction
Improve performance of statistical video analysis

Figure 1: The interaction between two relevant research areas

users' attentions. As shown in Figure 1, our new framework is able to bring together two research areas: visualization and statistical video analysis. We have observed that more effective techniques for statistical video analysis enable more efficient visualization of large-scale news stories. On the other hand, efficient visualization is able to provide useful feedback to improve our techniques for statistical video analysis.

By visualizing the semantic elements (such as keyframes and keywords) that are extracted from large-scale news video collections, our system is able to present abundant information to the users and attract their attention. By incorporating statistical video analysis for knowledge discovery (i.e., extracting news stories of interest), our system is able to relieve analysts from government or business from the burdensome task of watching and reading large volume of news videos and permit them to make decision based on valuable knowledge discovered by our system. In addition, more effective visualization of large-scale news stories of interest can also help general audiences find their personalized news easily. This integration of visual analytics capabilities is quite important in providing users with overviews plus appropriately accessible details for large amounts of streaming news video. Indeed, it is the first step towards an entirely scalable system where the coupling to interactive visualization (eventually with appropriate levels of abstraction) will be essential.

1.2 Related Work

To support more efficient news browsing and organization [23], several systems have been proposed to enable news visualization, and they have received attention from media and researchers. Among them, some systems [16, 12, 7] adopt a world map or regional map to show some specific information. Statistical information of text news [16] or video news [12] is put on a world map to inform the audiences of the "hotness" of regions. Buzztracker [16] also shows the correlations among different regions. Because only location related information is visualized in these systems, users may not have enough information to find the news stories of interest. IDVL [5] can also visualize the result of keyword-based queries. ThemeView [10] can visualize a large collection of documents with a predefined small keywords set. Because they require keywords input from the users, they are still far to be accepted. One commercial system, called 10×10 [9], organizes 100 icon images associated with 100 most important keywords from text news in a 10 by 10 icon grid. Because the icon grid cannot tell the users the real keywords and the importance weight between the keywords or the dynamic trend of the keywords along time, the naive users cannot figure out useful information from the icon grid easily. Another system, called newsmap [24], organizes news topics from Google news on a two dimensional rectangle, where each news story covers a visualization space that is proportional to the number of related news pages reported by Google. news titles are drawn in the corresponding visualization space allocated to them. It has several drawbacks: (1) The relative importance weights of news stories are assigned by the number of related news pages, and they may not be proportional to the interestingness of news stories; (2) news titles are folded in the visualization spaces assigned to the relevant topics and they may be difficult to read; (3) It lacks the capability to show the dynamic trend of news topics along time; (4) It cannot be directly extended for news video visualization. Thus there is an urgent need of developing new techniques for news video visualization.

Supporting efficient visualization of large-scale news video collections is very challenging. Even though there are many sophisticated text statistical analysis algorithms, performing statistical video analysis and understanding is still very hard if not impossible. 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 [21, 17]. In addition, supporting statistical video analysis plays an important role in enabling more efficient visualization of large-scale news videos. Without extracting the semantic topics from large-scale news video collections automatically, it is very hard to visualize them effectively. On the other hand, visualization is also able to provide valuable feedback that can be used to improve the performance of the underlying techniques for statistical video analysis.

Our paper is organized as follows: Section 2 introduces our new framework for news video visualization. Section 3 presents our algorithms for statistical news video analysis and automatic assignment of importance weights for news stories. Section 4 gives the implementation details. We then provide conclusions in Section 5.

2 Video Visualization Framework

There are two conflicting requirements for visualizing large-scale news video collection: (a) The visualization space is limited; (b) We need to show as many news stories of interest as possible in such limited space. Obviously, it is impossible to show all these news stories of interest in such limited space at the same time, we need to select and display only those most important news stories (i.e., most interesting news stories for the users) on the screen.

There are two critical issues that are very important to characterize the news stories and need to be detected and visualized: (1) The relative ratios for the interestingness among the available news stories; (2) The dynamic trend of news stories along time. Based on this understanding, we have developed a good measurement to quantify the interestingness of news stories and this interestingness measurement is incorporated to assign the relative importance weights for the news stories automatically. In addition, we have also developed a new algorithm to detect the news stories of interest and their trends with time automatically from large-scale news video collections. In order to visualize the dynamic trends of the news stories with time, we have created a novel animation framework by automatically assigning an importance weight for each keyframe, so that the sizes of the keyframes can be made proportional to the interestingness of the corresponding news stories. When the interestingness of the relevant news stories changes with time, the sizes for the corresponding keyframes are changed adaptively and the corresponding keyframes are moved automatically to new places according to their importance weights. Thus our animation technique is able to visualize the changes of the news topics with time and provide valuable information for decision making.

By visualizing and animating the map of keyframes for the relevant news stories of interest, the users can have an overview of the daily news stories and their changing trend with time. An example of news video visualization is given in Figure 2. When an interesting keyframe (i.e., certain news story of interest) appears on the screen, the users can click the corresponding keyframe and watch the related news story. This click of the keyframes implies the user's preference of news stories, so we can build a personalized user attention model by using the semantics of the news stories that the user accessed. Our system can then retrieve the news video databases with the given user attention model and the relevant news stories will be returned to the user.

To implement such a system for large-scale news video visualization, we have addressed two critical issues:

(a) **Interestingness measurement** of news stories, which is related to the user preference and cannot be quantified accurately without users' inputs. To address this problem, we assume that each user wants to know as much information as possible when he/she does not have knowledge of video contents in the databases. Thus, our news video visualization framework should try to display as many news stories (certainly different) as possible within a limited space, and the interestingness for a certain news story should be characterized by the information (knowledge) it can provide to the users. The news stories that can provide more information to the users should be assigned with bigger importance weights and cover bigger visualization space.

(b) **Informativeness measurement** of news stories, which is used to characterize the information of news stories. Obviously, it is difficult to measure the total information that a specific news story can provide because every news reporters may give similar but not identical description for a specific event. Such a situation makes it difficult to perform statistical analysis of news videos. The users may pay attention to small items of news videos such as person's name, product's name or a specific video clip. The overall measurement of the whole news story can't characterize all aspects of the user attention model. Rather than quantifying the informativeness measurement of the whole news stories, we first partition the news videos into a set of semantic items and measure the information carried by each semantic item. For the closed captions, the video text and the audio of news videos, the semantic items are keywords. For the visual channel of news videos, the semantic items are the semantic concepts or the semantic objects (for example, human faces) carried by the corresponding video clips. These multi-modal semantics can be integrated to extract the news stories of interest and enable more effective knowledge-level visualization of large-scale news videos.

Based on these observations, we need to develop a good measurement to quantify the informativeness for a certain semantic item in a given news video clip, which is used to organize the visualization in Section 4. From information theory, the information that a semantic item carries largely depends on how well we can predict it. If we can predict a given semantic item completely by using our previous knowledge, this given semantic item may carry no information for us. Based on this observation, we use a global probability model which characterizes the distribution of all semantic items as the predictor:

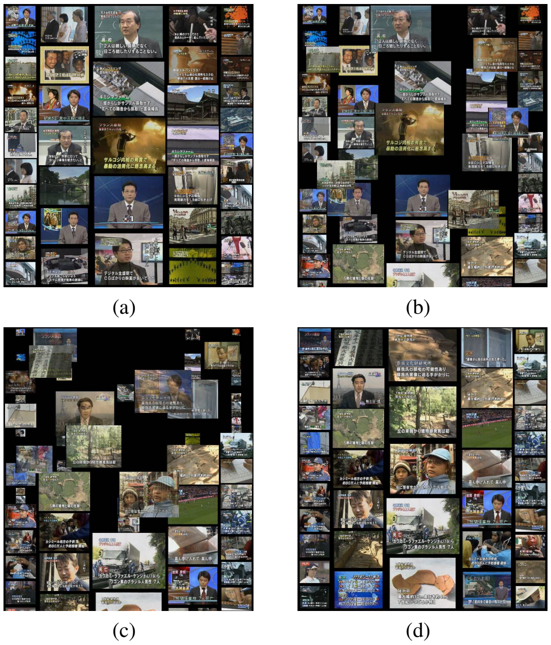

(a) (b) (c) (d)

Figure 2: An example of video news visualization. (a) Keyframes map for Japanese news on Nov. 12, 2005; (b) and (c) Intermediate animation; (d) Keyframes map for Japanese news on Nov. 13, 2005. The keyframes maps show the news topics on given day, the animation represents the trend of topic change over time.

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

where x is a given semantic item, S is the set of all semantic items and $g(x)$ is the probability of the given semantic item x. Obviously, the probability distribution of semantic items may change over time, thus the local probability model for a specific time interval can be defined as:

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

where t is the specific time interval of interest, such as one specific day. We have used different units in our experiments. For the general audiences, the most preferable unit is day. For some long term analysis, month is better. S_t is the set of all semantic items in that specific time interval t and $l_t(x)$ is the probability of the given semantic item x. The difference between the local probability model $L(t)$ for a specific time interval t and the global probability model G is able to tell us how much information we can obtained by knowing $L(t)$. Because both the local probability model $L(t)$ and the global probability model G are probability distributions, the widely used Kullback-Leibler divergence is able to characterize their difference:

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

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

$$\hat{w}_t(x) = l_t(x) \log \frac{l_t(x)}{g(x)} \tag{4}$$

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

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

The $w_t(x)$ in Eq. (5) can be normalized to simplify the multi-modal data fusion:

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

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

When the multi-modal importance weights (i.e., importance weights for video, audio, closed caption, special visual features) for all semantic items are computed by our system, they are combined to determine the overall weight for the given video clip. The unit of video clips (e.g. the partitioning over time) should be carefully selected to have best visualization. Too large unit may cause information loss; too small unit may carry too little information. Because of the natural properties of video, shot is the best suitable unit for visualization. A shot is a continuous set of frames captured by a camera for an uninterrupted period of time. Shot boundaries can be detected with high accuracy. Our system achieves 92% precision and 89% recall on Trec Video 2003 video database.

The overall weight for the given video shot is defined as:

$$w(i) = F\left(W(S_v(i)), W(S_a(i)), W(S_c(i)), W^{'}(V(i))\right) \tag{7}$$

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

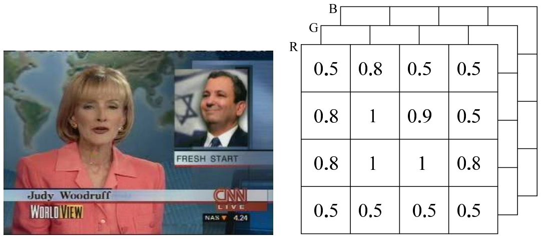
With this foundation, we now turn our attention to a more detailed analysis of the multi-modal video clips. In the next sections, we describe the following techniques: (1) technique for extracting the semantic items from news video clips; (2) technique for extracting some special visual features for automatic weight assignment; (3) multi-modal data fusion technique for determining the final weight for each video shot.

3 Statistical Video Analysis and Automatic Weight Assignment

In order to determine the importance weight for each video shot, we have developed a novel algorithm to extract the multi-modal semantic items (i.e., video, audio, text) and important special visual features automatically and weights are assigned automatically by the proposed statistical video analysis algorithm.

3.1 Semantic items extraction and statistical analysis of video

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



0.5	0.8	0.5	0.5
0.8	1	0.9	0.5
0.8	1	1	0.8
0.5	0.5	0.5	0.5

(a) A sample keyframe (b) Sub-blocks and weights

Figure 3: Repeated shots detection

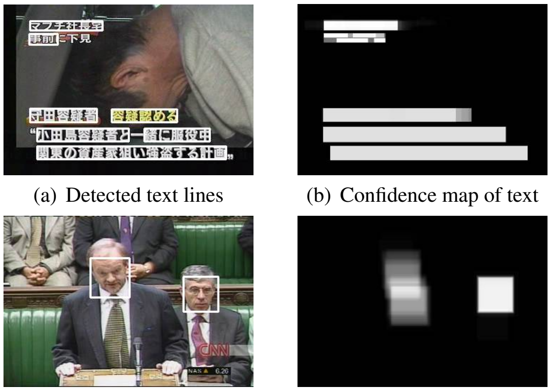
There are three types of semantic units that are critical to determine the importance weights for the corresponding video shots: (a) The first type is the statistical properties of the video shots; (b) The second type is the special video objects that appear in the video shots; (c) The last type is the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

3.1.1 Statistical property analysis of video shots

The video shots are the basic unit for news video interpretation. Thus they can be treated as the semantic items for automatic weight assignment. One certain video shot may be repeated multiple times because of the following reasons: (1) video shots for the anchors may repeat multiple times in the same news program; (2) video shots for the participants of an interview may appear multiple times in the same program; (3) video shots for interpreting the important news may appear in both the news summary at the beginning and the detailed report later in the same program; (4) video shots for the important news may appear in different news programs of the same channel (at different time periods) or different TV channels. The last two situations of video shot repeating indicate the importance for the corresponding video shots. Nevertheless, the first two situations of video shot repeating may not indicate that the corresponding video shots are important. In addition, the repeating of video shots in news videos is very different from the repeating of keywords in text documents, and it cannot be detected automatically by using simple comparison of the video shots. Thus new techniques are desired for detecting the video shot repeating in news videos, such that we can assign the importance weights for the video shots automatically.

One of the authors of this paper has developed an algorithm for detecting the identical video shots from news videos [18]. Even though this algorithm has high accuracy, the video shots repeated in news programs may have different properties due to captions, different capturing devices, different subpictures of anchor shots and different channel marks. To resolve this problem, the keyframes for the video shots are first partitioned into a set of sub-blocks as shown in Figure 3. Three 1-D color histograms are extracted from three color channels for each sub-block. The similarity between two keyframes is computed by the weighted sum of the similarities of the corresponding sub-blocks. The relative weights of sub-blocks are assigned by using numbers as shown in Figure 3. The purpose of the specific sub-blocks weights assignment is to help cluster anchor shots of the same anchor person together. The similarity of each pair of the sub-blocks is computed by the intersection of their color histograms. For each video shot, only 10 frames are selected and used as the keyframes to reduce the computation cost. The similarity of a pair of the video shots, $\varphi(i_1, i_2)$, is computed by the maximal similarity of all their keyframe pairs (i.e., 100 pairs).

Finally, a fixed threshold is used to detect the repeating of the video shots. If $\varphi(i_1, i_2) > \tau_r$, the corresponding video shots i_1 and i_2 are repeated over time. In our system, the threshold τ_r is set to



(a) Detected text lines (b) Confidence map of text

(c) Detected faces (d) Confidence map of faces

Figure 4: Text and face detection

0.9. We found the fixed threshold is good enough for most experiments thus we did not take further effort to design adaptive algorithm.

Since a given video shot may appear several times in the same program or in different programs, connected component analysis is performed by treating video shots as nodes and repeating video shot pairs as edges. The video shots in the same connected components are the multiple occurrence of one video shot. The intra-program repeating number $r_{intra}(i)$ and inter-program repeating number $r_{inter}(i)$ for each video shot can be computed by counting the connected components. The two numbers $r_{inter}(i)$ and $r_{intra}(i)$ for most video shots are equal to 1 because they are not repeated along time. Obviously, some video shots may have these two numbers bigger than 1 and the different repeating modes (i.e., different repeating situations) may provide different semantics so that different weights should be assigned. The weights for different repeating numbers are approximated by using a bell shaped curve:

$$w_{intra}(i) = e^{-\frac{\left(\frac{r_{intra}(i)-2}{2}\right)^2}{2}}, \quad w_{inter}(i) = e^{-\frac{\left(\frac{r_{inter}(i)-5}{6}\right)^2}{2}} \tag{8}$$

3.1.2 Video objects detection and statistical analysis

For news videos, text areas and human faces may provide important information of news stories of interest. By using the technique proposed in [13], text lines in news videos can be detected automatically. Human faces can also be detected by using the programs and models developed by OpenCV [4]. 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 such problem, the detection results for all the video frames within the same video shots are integrated and the relevant confidence maps for the detection results are calculated. As shown in Figure 4, such confidence maps can provide valuable information for evaluating the detection results.

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

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

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

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

By performing the face clustering technique developed in [19], face objects can be clustered to several groups and the human objects can be identified and be treated as the semantic items for weight assignment by using Eq. (6). The importance weight for human faces of shot i is computed by Eq. (11):

$$w_{face}(i) = \begin{cases} \max\limits_{x} \{\bar{w}_t(x) \,|\, x \in faces(i)\} & faces(i) \neq \emptyset \\ 0.5 & faces(i) = \emptyset \end{cases} \tag{11}$$

Where $faces(i)$ is the set of all face objects of shot i.

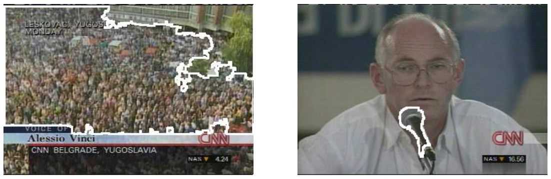
3.1.3 Semantic concept classification

The semantic concepts of the video shot may provide valuable information to enable more efficient visualization and retrieval of large-scale news video collections. Semantic video classification is one of the potential solutions to detect the semantic concepts of video shots. However, semantic video classification is still a challenging problem [3, 17, 21]. Many semantic video classification techniques have been proposed by different researchers. The related techniques can be classified into two categories:

(1) **Rule-based** (i.e., model-based) **approach** by using domain knowledge to define perceptual rules and achieve semantic video classification [25, 22, 2, 8]. One advantage of the rule-based approach is the ease to insert, delete, and modify existing rules when the nature of the video classes changes. However, effective techniques for semantic video classification should be able to discover not only the perceptual rules that can be perceived by human inspection, but also the hidden significant correlations (i.e., hidden rules) among multi-modal inputs. Thus the rule-based methods can only detect the semantic concepts correlated to the video making rules. Unfortunately, most semantic concepts in news videos generally have little correlations with the underlying video making rules.

(2) **Statistical approach** by using statistical machine learning techniques to extract the semantic concepts [1]. The statistical approach can support more effective solutions for semantic video classification by discovering non-obvious correlations (i.e., hidden rules) among different video patterns. However, its performance largely depends on the success of the underlying framework for video content representation and feature extraction. The visual features, which are selected for video content representation, should have the ability to discriminate among various semantic concepts. The difficulty for the existing frameworks for video content representation is the lack of means to relate the low-level visual features to the high-level semantic concepts. Most existing systems for content-based video retrieval (CBVR) use the shot-based or object-based (or, region-based) [3, 21, 17] visual features for video content indexing. Although the shot-based visual features are easy to be extracted, they are too general to be useful for semantic video classification, and thus the classification results are unreliable. Extracting the visual features by using video object regions may be able to capture the middle-level video semantics and thus provide more reliable classification results. However, automatic object extraction in general is a challenging problem because the homogeneous video regions in color, texture or motion do not correspond to the underlying semantic video objects directly.

Based on this understanding, we have proposed a new framework for video content representation, which is able to capture the



(a) Crowd region (b) Microphone

Figure 5: Salient object examples

middle-level semantics of video contents by using principal video shots. The principal video shots are defined as the video units by associating the video shots with the underlying concept-driven multi-modal (visual, auditory and image-textual) salient objects. Thus the principal video shots are semantic-sensitive and have strong correlation with the semantic concepts for the relevant news video clips. In addition, the multi-modal features, which are extracted by using the principal video shots, can be used to discriminate different semantic concepts effectively.

The visual salient objects are not necessarily the semantic video objects but some concept-driven *regions of interest* that are effective to characterize the related semantic concepts. The auditory and image-textual salient objects are not necessarily the recognized speech and image-text but just some auditory and image-textual principal patterns that are related to the semantic concepts of interest. In addition, the salient objects can be extracted effectively by using low-level multi-modal features because they are relatively feature-invariant. Two examples for salient object detection are given in Figure 5.

To clarify this procedure, we generate the visual salient object "crowd regions" as an example to show how we can design our detection functions. As shown in Figure 6, our detection function consists of the following steps: (1) homogeneous image regions on color or texture are first obtained automatically; (2) homogeneous image regions are then classified into two classes that are relevant or irrelevant to the visual salient object "crowd regions"; (3) the visual salient object "crowd regions" is formed by combining all homogeneous regions classified to "crowd regions" in the same frame. The detected salient object is then tracked among frames within the same video shot to eliminate noise. A confidence map can then be generated for the detected salient object. The principal video shot is then defined as the video shot associated with the underlying salient object and its confidence map.

Even though the principal video shots may contain abundant semantics, they are not equivalent to the semantic concepts. For example, the principal video shot of "microphone" can be classified either "report" or "announcement". To interpret the contextual relationship between a specific semantic concept C_j and the relevant principal video shots, the class distribution of the relevant principal video shots is approximated by using a finite mixture model with κ_j mixture components:

$$P(X, C_j, \Theta_{c_j}) = \sum_{i=1}^{\kappa_j} \omega_i P(X, C_j | \theta_i) \tag{12}$$

In the above expression, $P\left(X, C_j | \theta_i\right)$ is the ith mixture component to interpret one relevant context class. $\Theta_{c_j} = \left\{\kappa_j, \theta_{c_j}, \omega_{c_j}\right\}$ is the parameter tuple that includes the model structure, model parameters and weights, where κ_j is the model structure (i.e., optimal number of mixture components), $\theta_{c_j} = \left\{\theta_i = (\mu_i, \sigma_i) \mid i = 1, \cdots, \kappa_j\right\}$ is the model parameters (mean μ_j and covariance σ_j) for κ_j mixture components, $\omega_{c_j} = \left\{\omega_i \mid i = 1, \cdots, \kappa_j\right\}$ is the relative weights among these κ_j mixture components. Finally, X is the n-dimensional multi-modal features that are used for representing the relevant principal video shots.

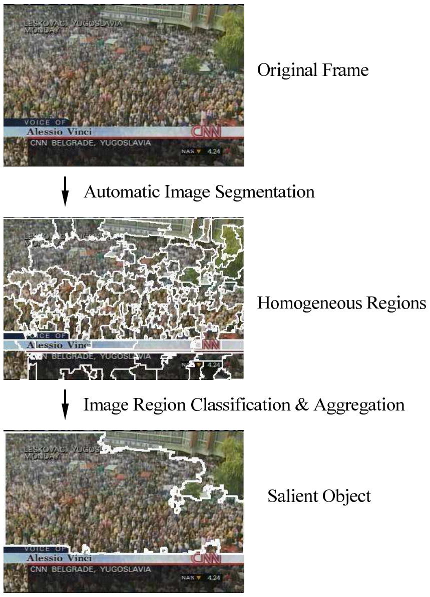

Original Frame
Automatic Image Segmentation
Homogeneous Regions
Image Region Classification & Aggregation
Salient Object

Figure 6: Salient object detection

Maximum likelihood criterion can be used to determine the underlying model parameters in Eq. (12). The optimal parameter set of model structure, weights and model parameters $\hat{\Theta}_{c_j} = \left(\hat{\kappa}_j, \hat{\omega}_{c_j}, \hat{\theta}_{c_j}\right)$ for the given concept C_j is then determined by:

$$\hat{\Theta}_{c_j} = \arg\max_{\Theta_{c_j}} \left\{ L\left(C_j, \Theta_{c_j}\right) \right\} \tag{13}$$

where $L\left(C_j, \Theta_{c_j}\right) = \sum_{X_l \in \Omega_{c_j}} \log P\left(X_l, C_j, \Theta_{c_j}\right)$ is the objective function. To save cost for manually labeling the training samples, it is very important to integrate the unlabeled samples in the model training procedure. When the unlabeled samples are incorporated for classifier training, each unlabeled sample X_u will be assigned a confidence score $\overline{C_j}(X_u) \in [0, 1]$. Thus, the unlabeled samples $\overline{\Omega_{c_j}} = \left\{\left(X_u, \overline{C_j}(X_u)\right) | u = 1, \cdots, N^u\right\}$ are similar to the labeled samples and we can integrate them for classifier training by modifying the objective function:

$$\begin{aligned} \overline{L}\left(C_j, \Theta_{c_j}\right) &= \sum_{X_l \in \Omega_{c_j}} \log P\left(X_l, C_j, \Theta_{c_j}\right) \\ &+ \lambda \sum_{\left(X_u, \overline{C_j}(X_u)\right) \in \overline{\Omega_{c_j}}} \overline{C_j}(X_u) \log P\left(X_u, C_j, \Theta_{c_j}\right) \end{aligned} \tag{14}$$

where the discount factor $\lambda \in [0, 1]$ is used to control the relative contribution of the unlabeled samples for semi-supervised classifier training.

The EM algorithm [15] is the traditional technique to optimize the model parameters toward Eq. (14). However, the traditional EM algorithm has two major problems: (1) it can't utilize the samples from other concepts to optimize the classifier; (2) it can't decide the optimal number of mixture component κ_j. To resolve the two problems and integrate the unlabeled samples in the model training, the adaptive EM algorithm [6] is used to train the optimal semantic concept models.

After the models for the semantic concepts of interest are formed, we can classify these principal video shots into the most relevant semantic concepts. Several most important concepts for visualization are implemented in our system. Shots of weather forecasting and sport are generally uninteresting. Thus they are detected and assigned smaller importance weights. Shots showing a person announcing something may be important. However, they will be weigh close to shots showing a reporter introducing the details, if we only use human face and text area objects. Both concepts are detected so that they can have different weights. It generally implies

Table 1: Semantic Concept Classification Performance

Concept	*Accuracy(%)*	*Concept*	*Accuracy(%)*
Announcement	75.35	Report	73.44
Sports	77.62	Weather	85.21
Gathered People	81.19		

Table 2: Semantic Concept Importance

Concept	w_c	*Concept*	w_c
Announcement	0.9	Report	0.3
Sports	0.5	Weather	0.5
Gathered People	1	Unknown	0.8

important events that a lot of people gather together. The concept “gathered people” is detected so that shots related to this concept can be assigned a larger weights. The semantic classification accuracy of our system is in Table 1.

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

$$w_{concept}(i) = w_c(C(s)) \times w_d(i) \tag{15}$$

3.2 Semantic item extraction from audio and text

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

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

In news videos, the news titles shown in video may provide very important keywords and thus they should be detected. Some special text sentences, such as “*somebody*, CNN, *somewhere*” and “ABC’s *somebody* reports from *somewhere*”, need to be processed separately. The names for news reporters in those text sentences are generally unattractive to the users. A context-free syntax parser is used to detect and mark this information.

All capital strings will fail most named entity detectors because initial capitalization is very important to achieve accurate named entity recognition. One way to resolve this problem is to train a detector with ground truth from closed caption text. However, it’s very expensive to obtain the manually marked text material. Because the English has relatively strict grammar, it’s possible to parse the sentence and recover most capital information by using part-of-speech and lemma information. We use TreeTagger [20] to perform the part-of-speech tagging. Capital information will be recovered by TreeTagger automatically.

After special sentences are marked and capital information is recovered, LingPipe [11] is used to perform the named entity detection and resolve cross reference. The model used is the news model of LingPipe. All parameters are set to default value.

Finally, LingPipe marked XML files are parsed to extract keywords and associated frequency information. Detected named entities are kept in their original format. Other text strings are segmented by non-alphabet characters and each segment is treated as a word. Because we will compute the weight by using frequency information with Eq. (6), we do not need to consider the stop words. The weight computation will automatically suppress the stop words.

The keyword weight of a shot is computed by Eq. (16):

$$w_{keyword}(i) = \max_{x} \{\bar{w}_t(x) \,|x \text{ is a keyword of } i\} \tag{16}$$

3.3 Multi-modal data fusion

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

$$w_{repeat}(i) = \max\{w_{intra}(i), w_{inter}(i)\} \tag{17}$$

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

$$w_{object}(i) = \max\left\{w_{faceArea}(i), w_{textArea}(i)\right\} \tag{18}$$

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

$$w_{semantics}(i) = \max\left\{w_{face}(i), w_{concept}(i)\right\} \tag{19}$$

The reason to use max operation in above equations is that we want to detect the existence of interesting visual properties (e.g. the repeat pattern, the visual objects and the predefined semantics) of shots. The max operation assures the computed weights do not be suppressed a lot by the missing of a specific property.

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

$$w_{video}(i) = \sqrt[3]{w_{repeat}(i) \times w_{object}(i) \times w_{semantics}(i)} \tag{20}$$

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

$$w(i) = \gamma \times w_{video}(i) + (1-\gamma) \times w_{keyword}(i) \tag{21}$$

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

4 Visualization Implementation

After the importance weight for each video shot is computed, more efficient visualization of large-scale news video collections can be achieved by assigning different sizes for different news stories of interest. In addition, the video frame in the middle of each video shot is selected as the keyframe for the given video shot.

There are several constraints for the layout and animation of the keyframes (i.e., news stories of interest). Firstly, the animation of the keyframes should be able to visualize the dynamic trend of news stories of interest (i.e., changing of news topics with time). Secondly, the sizes of the keyframes should be big enough so that the users can read the video content. In addition, the important keyframes (i.e., news stories of interest) are best visualized in the middle of the screen with larger size. Obviously, it is also very important to visualize as many keyframes (i.e., news stories of interest) as possible within a limited screen at the same time.

First, we introduce our algorithm for visualizing the map of the keyframes in a given time period. Because the aspect ratio of the keyframes is fixed, the treemap algorithm [14] and its variants cannot effectively allocate the proper screen space for each keyframe. To address this problem, we have proposed a new visualization algorithm by using the column structure of screen layout. As shown in Figure 2 (a), 5 keyframes with largest weights are put in the middle of the screen and form a column. Then the other 14 keyframes are separated into two groups and form two columns at the left and right of the middle column. Finally, 22 other keyframes form the outermost two columns. This is a good layout that balances the constraints of overall screen size (standard desktop display in this case), the need to quickly apprehend individual keyframes, and the need to understand the flow of important stories within the context of other stories. If one had a significantly larger or smaller display, the layout could be adjusted accordingly.

To animate the keyframes (i.e., visualize changes of news topics) with time, the association of the keyframes with time should be extracted. Because news reports are rather condensed and tightly formatted, the video materials may not be repeated exactly, but similar video materials with some differences may used repeatedly over time. As we already perform sophisticated analysis on the video materials, we can just use the information obtained in the weight assignment procedure to extract the associations. Two video shots i_1 and i_2 are linked when one of the following three criteria is reached: (1) Both of them contain the faces of the same person; (2) They are a repeating video shot pair, e.g. $\varphi(i_1, i_2) > \tau_r$; (3) Both of them have at least 3 identical keywords.

Two keyframes maps are involved in the animation: the old map for the first time period and the new map for the following time period. The analysis of video shots separates the keyframes in the maps into two groups: keyframes without any link and others with links. For keyframes without links, the ones in the old map will gradually zoom to zero size and scroll up until they disappear from the screen; the ones in the new map will scroll up from the bottom until they reach their final positions. For keyframes with links, they will stay in the map. The resulting animation shows a flow of keyframes and some stationary "islands" in the flow. Thus dynamic topic trends are presented to the users, who can intuitively see which topics are growing or diminishing and which are retaining their importance.

5 Conclusion and Future Works

By integrating several methods of statistical video analysis for knowledge discovery, our system is able to provide valuable information to users and enable significantly more effective and efficient visualization of news video from many broadcast channels. It can also help users find news stories of interest in a short time. These results demonstrate the power of integrating visualization with automated analysis techniques working together for a clear purpose. As a result, the complexities of the process are hidden from the user who receives meaningful results in an intuitive visual form.

In the future, visualization for single or several news stories, such as concept-oriented skimming, will be integrated into our system to help users examine specific news selected or retrieved via the visual interface. The system will also be scaled up to a much larger collection of broadcast channels and topics. To achieve this, new interactive visualization techniques that contain levels of abstraction will need to be derived.

Acknowledgment

This work was performed with partial support from the National Visualization and Analytics Center (NVAC$^{\text{TM}}$), a U.S. Department of Homeland Security Program, under the auspices of the Southeastern Regional Visualization and Analytics Center. NVAC is operated by the Pacific Northwest National Laboratory (PNNL), a U.S. Department of Energy Office of Science laboratory.

References

[1] B. Adams, G. Iyengar, Ching Yung 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*, 2:170–185, 2003.

[2] Brett Adams, Chitra Dorai, and Svetha Venkatesh. Towards automatic extraction of expressive elements from motion pictures: Tempo. *IEEE Trans. on Multimedia*, April 2002.

[3] Shih-Fu Chang, William Chen, and Hari Sundaram. Semantic visual templates: linking visual features to semantics. *IEEE Workshop on Content Based Video Search and Retrieval, in conjunction with IEEE ICIP '98*, Oct. 1998.

[4] Intel Cop. Open source computer vision library. *http://www.intel.com/technology/computing/opencv/*.

[5] Mark Derthick, Michael G. Christel, Alexander G. Hauptmann, and Howard D. Wactlar. Constant density displays using diversity sampling. In *InfoVis'03*, Seattle, WA, October 2003.

[6] Jianping Fan, Hangzai Luo, and Yuli Gao. Learning the semantics of images by using unlabeled samples. In *IEEE CVPR*, San Diego, CA, June 20-26 2005.

[7] Michael Gastner, Cosma Shalizi, and Mark Newman. Maps and cartograms of the 2004 us presidential election results. *http://www.cscs.umich.edu/~crshalizi/election/*, 2004.

[8] A. Hanjalic, R.L. Lagendijk, and J. Biemond. Automated high-level movie segmentation for advanced video retrieval systems. *IEEE Trans. on CSVT*, 9(4), 1999.

[9] Jonathan Harris. http://tenbyten.org/10x10.html.

[10] B. Hetzler, P. Whitney, L. Martucci, and J. Thomas. Multi-faceted insight through interoperable visual information analysis paradigms. In *InfoVis'98*, Research Triangle Park, NC, 1998.

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

[12] Informedia-II. http://www.informedia.cs.cmu.edu/dli2/.

[13] A.K. Jain and B. Yu. Automatic text location in images and video frames. *Pattern Recognition*, 13(12):2055–2076, 1998.

[14] Brian Johnson and Ben Shneiderman. Tree maps: A space-filling approach to the visualization of hierarchical information structures. In *the 2nd International IEEE Visualization Conference*, pages 284–291. IEEE ComputerSociety, 1991.

[15] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. John Wiley & Sons, New York, 2000.

[16] Craig Mod. http://www.buzztracker.org/.

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

[18] Shin'ichi Satoh. News video analysis based on identical shot detection. In *ICME*, pages 69–72, 2002.

[19] Shin'ichi Satoh and Norio Katayama. An efficient implementation and evaluation of robust face sequence matching. In *ICIAP'99*, pages 266–271, 1999.

[20] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK, 9 1994.

[21] John R. Smith and Chung-Sheng Li. Image classification and querying using composite region template. *Proc. of Computer Vision and Image Understanding*, 75(1-2):165 – 174, 1999.

[22] Hari Sundaram and Shih-Fu Chang. Determining computable scenes in films and their structures using audio-visual memory models. *ACM Multimedia*, pages 95–104, 2000.

[23] Jeremy Wagstaff. On news visualization. *http://www.loosewireblog.com/2005/05/on_news_visuali.html*, 2005.

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

[25] Hong Jiang Zhang, Shuang Yeo Tan, Stephen W. Smoliar, and Yi Hong Gong. Automatic parsing and indexing of news video. *Multimedia Systems*, 2(6):256–266, January 1995.