Tracking News Stories Across Different Sources

Yun Zhai
School of Computer Science
University of Central Florida
Orlando, Florida 32816
yzhai@cs.ucf.edu

Mubarak Shah
School of Computer Science
University of Central Florida
Orlando, Florida 32816
shah@cs.ucf.edu

ABSTRACT

Information linkage is becoming more and more important in this digital age. In this paper, we propose a concept tracking method, which links the news stories with the same topic across multiple sources. The semantic linkage between the news stories is reflected in combination of both of their visual appearance and their spoken language content. Visually, each news story is represented by a set of key-frames with or without detected faces. The facial key-frames are linked based on the analysis of the extended facial regions, and the non-facial key-frames are correlated using the global Affine matching. The language similarity is expressed in terms of the normalized text similarity between the stories' keywords. The output results of the story linking approach are further used in a story ranking task, which indicate the interesting level of the stories. The proposed semantic linking framework and the story ranking method have been tested on a set of 60 hours open-benchmark TRECVID video data, and very satisfactory results for both tasks have been obtained.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Abstracting Methods; H.3.3 [**Information Search and Retrieval**]: Retrieval Models.

General Terms

Algorithm, Performance.

Keywords

Semantic Linking, News Processing, News Ranking.

1. INTRODUCTION

There are many news agencies nowadays that are broadcasting what is happening around us and around the world. Their reporting is instant as real time and comprehensive, covering issues on politics, economics, health, sports, etc. Large networks provide more national and global news, while local stations concentrate more on the regional issues. Due to the numerous variety of audiences, one may only be interested in a few areas or topics, e.g., sports or politics. Thus, finding a particular story based on the user's preference is important. Furthermore, even though everyone in the news industry claims their reporting is objective, the actual attitude in the broadcasting may be biased and be different from network to network due to the background and culture differences. Therefore, watching the same news from multiple sources provides the audience with more comprehensive views. To achieve this goal, the semantic linkage between stories must be established.

Several works have been proposed for the shot-based linking. Tavanapong *et al.* [8] has proposed shot clustering method for the purpose of video scene segmentation. The shot image is constructed from the corresponding key-frames. The links for grouping the shots are established by comparing the sub-blocks in the shot images. Odobez *et al.* [5] used the spectral technique to cluster the video shots. Multiple key-frames were employed for representing a single shot. The color histograms were used for the visual similarity measure. The correlation is further scaled by the temporal distance. Zhang *et al.* [10] proposed a simpler version of the spectral clustering technique. The stories from two sources are modelled as the vertices in a bipartite graph, and the computation of finding the eigenvalues for the similarity matrix is dramatically reduced. The clustering for the stories are based on the analysis of textual information (TF-IDF), and the clustering for video shots is based on the mid-level or high-level visual concepts. Sivic *et al.* [6] extended their object grouping framework for clustering the video shots in the movie. First, an object is extracted by a sequence of feature extraction, feature tracking, homography estimation and object grouping. The 3D structure of the object is computed and used for searching the same object in other shots. Ngo *et al.* [4] has proposed a two-level hierarchical clustering method for grouping the shots. Both color and motion information are used as features. One color histogram in the YUV space is computed for each shot from its discrete cosine (DC) images and used in the first level clustering. Temporal slice analysis is used to compute the tensor histogram, which is a motion feature, for the second level clustering. Cheng *et al.* [2] proposed a structure called *Shot Cluster Tree*. First, the shots that are visually similar and are adjacent in time are grouped into shot groups. The shots groups are later merged into shot clusters based on their content similarity. The color histogram of the key-frame of each shot is used as

the similarity feature.

Videos have the common hierarchy of [frames] → [shots] → [scenes/stories] → [video]. Many of the above mentioned grouping methods consider the video shots as their basic computational units, and most of them are only based on the low-level visual information. However, in news videos, audiences are more likely to retrieve storylines with complete semantic topics, which cannot be covered by a single shot. Stories, on the other hand, contain more semantic contents than video shots and often provide the complete storylines. In this sense, story level matching is more relevant in the semantic retrieval task.

In this paper, we present a framework for the semantic linking of the news topics. Unlike the conventional video content linking methods based only on the video shots, the proposed framework links the news stories across different sources. Another advantage is that the proposed method uses more semantic features compared with other methods, such as face related features and the textual information. The semantic linkage between the news stories is computed based on their visual and textual similarities. The visual similarity is carried on both of the story key-frames with or without faces detected. The textual similarity is computed using the automatic speech recognition (ASR) output of the video sequences. The proposed method is tested on a large open bench-mark dataset. Furthermore, the output of the story linking method is applied in a news ranking task. The matched stories are modelled in a bipartite graph. The graph is segmented into sub-graphes using the connect component method, and the story ranking is performed by analyzing the corresponding component's size. The rest of this paper is organized as follows: Section 2 describes the proposed framework in detail, including the computation of the visual similarity and the textual correlation between stories; Section 3 presents the system evaluation results for the tasks of the story linking and news ranking; finally, Section 4 concludes our work.

2. PROPOSED FRAMEWORK

Let us consider how humans link the stories on the same topic and distinguish the ones that are different. Given two news stories, our vision system gives us the first impression about the common contents of the two. It could be the same person of interest, the same action/activity, or the same physical site. For example, when the General Secretary of the United Nations proposes a peace plan, all the news channels broadcast the same picture containing his face. On the other hand, in a riot reporting, there may not be any particularly person of interest. Instead, the physical scene is emphasized. In this scenario, global matching between the images is needed. Besides the visual information, we also acquire the language information of the video, which provides the most direct semantic cue in the news videos. Often the end-form of the language information is the speech and/or the closed caption (CC) on the screen. They contain the text form of the spoken words in the video. The proposed method tries to imitate the human logic and constructs the semantic linkage between news stories in a similar way using both visual and textual information. It computes the visual similarity based on both of the facial and non-facial key-frames of the stories, and establishes the textual correlation using the automatic speech recognition (ASR) output.

(a). Sample key-frames from the testing videos with faces labeled.

(b). The "body" regions extracted from the above detected faces.

Figure 1: (a). The sample key-frames with the detected faces; (b). The body regions extended from the faces. Global feature comparison or face correlation fails to link the same person in these examples, while the comparison of the "body" regions provides the meaningful information.

2.1 Visual Correlation

The first step in the computation of visual correlation is to detect faces in the key-frames of the stories. If a face is detected in a key-frame, that key-frame is classified as a facial key-frame; otherwise, it is classified as a non-facial key-frame. Given a story S_i, we have a set of its key-frames, which is composed of two disjoint sets, the facial key-frames, $K = \{k_{(i,1)}, \cdots, k_{(i,m_i)}\}$, and the non-facial key-frames, $\Phi = \{\phi_{(i,1)}, \cdots, \phi_{(i,n_i)}\}$, where m_i and n_i are the numbers of facial and non-facial key-frames in S_i, respectively. Computation of the visual similarity between two stories is carried on K and Φ separately. Here, the video shots containing anchor person(s) are not considered in the visual similarity computation. The reason is that it does not provide meaningful linkage, since no anchor person works for two stations, and stories broadcasted by the same anchor person do not imply they are similar. We use a graph-based method to remove the anchor shots [9]. The underlying mechanism for the anchor removal technique is to analyze the frequencies of the video shots in the news program. The shots are classified as anchor, sub-anchor and non-anchor shots according to their frequencies based on the fact that the anchor and sub-anchor shots appear much more often than other non-anchor shots.

2.1.1 Facial Key-Frame Matching

Many times, the news networks broadcast events that involve a particular person or a group of persons. In these types of news stories, since the person is performing the action (e.g., a political leader giving speech), or the persons constitutes the major part of the event (e.g., meeting of foreign leaders), he/she becomes the focus of the interest. The images often reveal the person's face. In these situations, the best linkage between stories is provided by the correlation of the persons by their facial information. Common face correlation methods are known to have some drawbacks, such as being sensitive to the pose of the face, lighting conditions, and the sizes of the faces. This is because that traditional face correlation methods use the local information of the face patch. To overcome the aforementioned problems, we utilize

(a). Image1 (b). Image2 (c). Overlap Image (d). Residual Map

(e). Image1 (f). Image2 (g). Overlap Image (h). Residual Map

Figure 2: Example demonstration on the Affine matching between non-facial key-frames. (a). One key-frame from ABC video; (b). One key-frame from CNN video; (c). The overlap image between (a) and Affine-warped (b); (d). Residual map between (a) and (b); (e,f). Key-frames from ABC and CNN videos, respectively; (g). Overlapping image between (e) and (f); (h). Residual map between (e) and (f). Darker pixel represents lower residual. Key-frames (a) and (b) match well comparing to key-frames (e) and (f). In the residual map, brighter colors represent higher residual values, while darker pixels represent lower residual values.

the global properties related to the detected faces. An extended region, "body", is used. The procedure for obtaining the "body" region is as follows: first, the face in the key-frame is detected by the face detector [7]. The detected face region is then extended to cover the upper body of the corresponding person. The idea behind this is that in the news stories involving the important person, the person usually wears the same clothes. Therefore, this can be taken as the cue for the similarity. All the body regions in story S_i are collected to provide the body set, $B = \{b_{(i,1)}, \cdots, b_{(i,\beta_i)}\}$, where β_i represents the total number of body patches in the story. Note that $\beta_i \geq m_i$, because there might be more multiple faces detected in a single key-frame. Some of the facial key-frames and their body patches are shown in Figure 1.

We compute the 3D color histogram, denoted by $h_{(i,j)}$, of each body patch $b_{(i,j)}$. The Bhattacharya distance between two histograms is used in the similarity measure. The similarity between two body patches $b_{(i,j)}$ and $b_{(p,q)}$ is defined as,

$$SimF(b_{(i,j)}, b_{(p,q)}) = e^{-d_B\left(b_{(i,j)}, b_{(p,q)}\right)}, \quad (1)$$

where $d_B\big(b_{(i,j)}, b_{(p,q)}\big)$ is the Bhattacharya distance computed as $d_B\big(b_{(i,j)}, b_{(p,q)}\big) = -ln\big(\sum_{r \in allbins} \sqrt{b^r_{(i,j)} b^r_{(p,q)}}\big)$. The visual similarity between two stories S_i and S_p over the facial key-frames is computed as follows,

$$\Gamma_F(i,p) = max\Big(SimF(b_{(i,j)}, b_{(p,q)}\Big), \quad (2)$$

where $j = \{1, \cdots, \beta_i\}$ and $q = \{1, \cdots, \beta_p\}$. Based on Eqns. 1 and 2, the range for Γ_F is bounded in $[0, 1]$.

2.1.2 *Non-Facial Key-Frame Matching*

Some stories do not contain human faces. For instance, in a reporting of a riot, no particular human face can be detected due to various reasons. Another case is the stories with special format, such as weather news and sport reporting. In these stories, only non-facial key-frames are available. The visual linkage here is defined by the similarity between the non-facial key-frames.

Given two non-facial key-frames $\phi_{(i,j)}$ and $\phi_{(p,q)}$ from stories S_i and S_p, their similarity is computed based on the Affine transformation between them. The Affine transformation is expressed as,

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & b_1 \\ a_3 & a_4 & b_2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \quad (3)$$

where vector $[b_1\ b_2]^T$ captures the global translation between two images, and $[a_1\ a_2;\ a_3\ a_4]$ captures the scaling, rotation and shearing. The above equation can be expressed in a shorter form $X' = AX$. This transformation results in $\phi_{(i,j)}(x,y) = \phi_{(p,q)}(x', y')$, where $\phi(\cdot,\cdot)$ refers to the image pixel value at the corresponding location. The goodness of the transformation is expressed as the residual $\delta(x,y) = |\phi_{(i,j)}(x,y) - \phi_{(p,q)}(x',y')|$. The entire process is as follows:

1. Compute the Affine parameters, A, between key-frames $\phi_{(i,j)}$and $\phi_{(p,q)}$;
2. For each pixel (x,y) in $\phi_{(i,j)}$, compute the residual $\delta_{(i,j)}(x,y)$ by applying A;
3. The overall residual between $\phi_{(i,j)}$ and $\phi_{(p,q)}$ is com-

Starting Time: 11:35:59
Key-Words: federal, grand, jury, meet, investigate allegation, president, clinton, sexual, relationship, former, white, house, interned, urged, lie, deputy, counsel, bruce, lindsey, chief, staff, john, ...
Ending Time: 11:36:34

Figure 3: The key-frame of an example story in a v accompanied by the key-words extracted from that s The starting and ending times are based on the anal version of the video (tape).

puted as $ResN(\phi_{(i,j)}, \phi_{(p,q)}) = \frac{1}{N_\phi} \sum_{all\ (x,y)} \delta_{(i,j)}$

The visual similarity between stories S_i and S_p based o non-facial key-frames is defined as,

$$\Gamma_N(i,p) = exp\Big(- \frac{\big[min\big(ResN(\phi_{(i,j)}, \phi_{(p,q)})\big)\big]^2}{\sigma^2}\Big),$$

where $j = \{1, \cdots, n_i\}$ and $q = \{1, \cdots, n_p\}$. Parar σ is the scaling factor. In the experiments, we have $\sigma = 20$. The similarity values for Γ_N are also bou in the range of $[0, 1]$. Two pairs of the images and corresponding residual maps are presented in Figure 2. pair of images is related to the same scene, and the in in the other group are un-related. The residual bet the related images is much lower than the one betwee un-related images.

2.2 Textual Correlation

Sometimes, visual information is insufficient to distinguish the difference. Consider the following story: the Congress is passing a bill. One news source is showing the debate among the senators, while another source is showing the comments from the political activists. The content in each of these stories is focusing on the same topic, but they are visually different. In this type of situation, the text information plays more important role in the semantic linking process.

The text information is the automatic speech recognition (ASR) output of the video. The ASR output contains the recognized words from the audio track of the news programs with their starting time and duration. For each candidate news story S_i, we extract the key-words between its time lines by applying a filter to prune the stop words, such as "the", "and", "or", etc. The story time line covers all the video shots, including both anchor and non-anchor shots. The extracted key-words form the *sentence* of the story, which is denoted by Sen_i and has length of $L(Sen_i)$. One example of the story *sentence* is shown in Figure 3. If two stories are focusing on the same topic, there usually is a correlation in the narration of the video. In our approach, this textual linkage between stories S_i and S_p with *sentences* Sen_i and Sen_p is computed by the normalized text similarity (NTS),

$$NTS(i,p) = \frac{M_{i\to p} + M_{p\to i}}{L(Sen_i) + L(Sen_p)}, \quad (5)$$

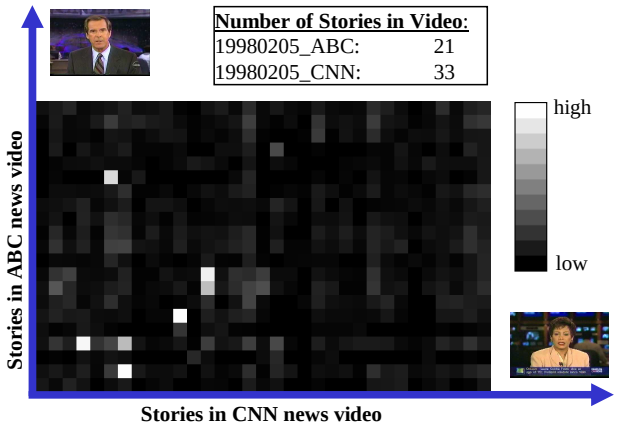

Figure 4: The similarity between two videos. The horizontal and vertical axis represent the stories from a CNN and an ABC video respectively. The axis are represented by the selected anchor images. In this example, brighter cells correspond to higher story similarity values.

where $M_{i\to p}$ is the total number of key-words in Sen_i that also appear in Sen_p, and $M_{p\to i}$ is the number of key-words in Sen_p that also appear in Sen_i. The textual similarity Γ_T between stories S_i and S_p is defined as,

$$\Gamma_T(i,p) = NTS(i,p). \quad (6)$$

Based on the definition, it is easy to see that the range for the textual similarity values Γ_T is also $[0, 1]$.

2.3 Correlation Fusion

Up to this point, the visual and textual similarities have been determined. The semantic linkage between the news stories is the fusion of these similarities. To determine the form of the fusion, the relationship between the similarities must be defined. The final fusion of the visual and textual similarities is defined as,

$$SSim(i,p) = \alpha_V \times \Psi\big(\Gamma_F(i,p), \Gamma_N(i,p)\big) + \alpha_T \times \Gamma_T(i,p), \quad (7)$$

where α_V and α_T are constants to balance the importance of the visual and textual similarities, respectively, and $\Psi(\cdot)$ is the fusion function between Γ_F and Γ_N.

The visual similarities Γ_F and Γ_N are computed from two disjoint sets: facial key-frames K and non-facial key-frames Φ, therefore, these two measures are independent from each other. Thus, the one that has a higher value is dominant over the other. In our formulation, the fusion function $\Psi(\Gamma_F, \Gamma_N)$ is defined as $max(\Gamma_F, \Gamma_N)$. On the other hand, no conclusion of independence can be drawn between the visual and textual similarities. Therefore, we use a linear fusion to combine these two measurements. The constants α_V and α_T balance the importance of the visual and textual effects. The simplest way to select them is to let

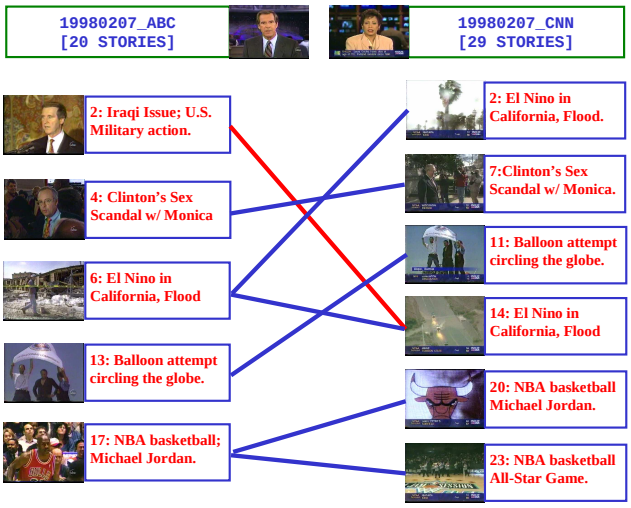

Figure 5: One example of the story matching. Two news videos from ABC and CNN for the same date are used. In total seven matches were detected, six of them are labelled as "Relevant" (blue matches), and one is labelled as "Irrelevant" (red match). The matched stories are displayed by their first key-frame and brief summaries.

$\alpha_V = \alpha_T = 0.5$. However, users can tune them according to their preferences. If more effect is expected from the textual information, α_T can be increased, while α_V is decreased.

A few situations need special attention. Some news stories occur only in the anchor shots. Therefore, only textual information is available. Similarly, one of the visual similarities might be missing due to the absence of the facial or non-facial key-frames. To deal with these cases, we have following rules:

- If the facial key-frame set K is empty, and the non-facial key-frame set Φ is not empty, set $\Gamma_F = \Gamma_N$;
- If Φ is empty, and K is not empty, set $\Gamma_N = \Gamma_F$;
- If both Φ and K are empty, replace $\Psi(\Gamma_F, \Gamma_N)$ by Γ_T. This means that if no visual information is available, the textual similarity plays the dominant role.

Given two news videos containing multiple stories, a story similarity map can be constructed. One example is shown in Figure 4. In this similarity map, brighter cells represent higher similarity values.

3. SYSTEM PERFORMANCE

We have experimented our method on a large dataset. This dataset is provided by the U.S. National Institute of Standards and Technologies (NIST). It is a open-benchmark for the content extraction evaluation tasks. The dataset contains 104 videos in MPEG-1 format from two news sources: ABC *World News Tonight with Peter Jennings* and CNN *Headline News*. The videos are distributed over 52 days, with each day having a video from ABC and CNN. Each video is around 30 minutes long, covering both of the regular news programs and non-news segments in between the

Date	Matches	Relevant	Somehow Relevant	Irrelevant
19980204	7	6	1	0
19980205	7	6	0	1
19980207	7	6	0	1
19980208	9	7	1	1
19980209	8	8	0	0
19980211	7	5	0	2
19980212	8	6	2	0
19980213	9	9	0	0
19980214	6	5	0	1
19980215	7	6	0	1
19980216	5	5	0	0
19980217	10	9	0	1
19980218	7	6	0	1
19980219	8	7	0	1
19980220	4	4	0	0
19980221	8	7	1	0
19980223	7	7	0	0
19980225	10	10	0	0
19980226	3	3	0	0
19980302	4	3	0	1
19980304	9	8	1	0
19980305	6	6	0	0
19980306	9	9	0	0
19980307	6	4	0	2
19980308	6	4	1	1
19980309	7	7	0	0
19980310	6	5	1	0
19980311	8	8	0	0
19980312	8	7	1	0
19980313	8	7	1	0
19980315	1	1	0	0
19980316	10	9	0	1
19980317	7	6	0	1
19980318	7	6	0	1
19980319	7	7	0	0
19980320	4	4	0	0
19980321	5	5	0	0
19980322	1	1	0	0
19980323	11	8	2	1
19980325	5	5	0	0
19980326	8	8	0	0
19980327	5	5	0	0
19980328	3	3	0	0
19980329	5	4	0	1
19980330	10	8	0	2
19980413	13	12	0	1
19980414	7	7	0	0
19980416	12	10	0	2
19980417	3	3	0	0
19980418	5	4	0	1
19980419	4	3	0	1
19980420	6	5	0	1

Figure 6: Table Summarizing the Story Linking Results.

stories. The TDT2 [1] has provided the ground truth for the news stories generated by manual annotation. Accompanying with the video data, NIST also provided the ground truth data for the common shot boundaries, key-frames and the automatic speech recognition (ASR) outputs by LDC [3]. Each video contains around 20-30 news stories.

Given two news videos from different sources, assume video1 contains stories $\{S_{1,1}, \cdots, S_{1,n_1}\}$, and video2 contains stories $\{S_{2,1}, \cdots, S_{2,n_2}\}$. We first compute their similarity map $SimMat$ based on Eq. 7. To classify a match between $S_{1,i}$ and $S_{2,j}$, the value of cell $SimMat(i,j)$ is verified against the pre-defined threshold. In our experiment, only videos from the same date are matched with each other. One reason for that is because the news stories are interesting only to the audience in their proposed time periods. Stories that are apart in time do not tend to match. However, the proposed method has the capability to match stories across videos, regardless of their time difference. A full set of matches for a pair of example videos is shown in Figure 5, and one pair of the matched stories is demonstrated in detail in Figure 7. In Figure 7, the key-frames and the extracted key-words of

Figure 7: An example of the matched stories from two different sources. The left block contains the key-frames and key-words extracted from a story in video [19980204_ABC], and the right block contains the key-frames and key-words extracted from a story in video [19980204_CNN]. The key-frames that are bounded by red-boxes provide the visual similarity between these two stories, since both stories are captured at the same presidential palace. The key-words in blue boldface are the common words that appear in both of the two stories. From the figure, the reader can easily draw the conclusion that both stories deal with the issue of weapons inspections of the Iraqi presidential palaces.

the stories are shown. The key-frames providing the vis similarity is red-boxed, and the common key-words in b stories are in blue boldface.

In the areas of multimedia processing and information trieval, two accuracy measures are often used, precision a recall. They are defined as, $Precision = \frac{M}{D}$ and $Recall$ $\frac{M}{G}$, where M is the number of the retrieved docume matched with the ground truth, D is the total number of trieved documents, and G is the total number of the grou truth documents. In our application, due to the vast amou of data, it is very difficult to determine the number of grou truth G. Therefore, we concentrate on the precision m sure, which captures how precisely the method performs. our evaluation, there are three categories for the matcl stories: Relevant, Somehow Relevant and Irrelevant. A p of matched stories is said "Relevant", if there is no ambi ity in their content. For stories that are partially relat they are classified as "Somehow Relevant". If the stor are focusing on completely different topics, "Irrelevance" assigned as their label. Since there are three categories matched story pairs, we used a modified version of the p cision measurement. For each "Relevant" pair detected is assigned a satisfactory score of 1.0; for each matching p with "Somehow Relevant", it is assigned a score of 0.5; nally, if a matching pair is "Irrelevant", a score of 0.0 assigned. Higher overall scale indicates better satisfact rate. Ideally, there should be an overall scale of 100% sa faction. Within a total of 353 matching pairs, there are pairs that are the "Relevant" matches, 12 "Somehow R evant" matches and 27 "Irrelevant" matches. The syst performance, average satisfactory scale, is 0.9065. If o the "Relevant" matches are considered as the true match the overall precision is $Precision = 0.8895$. The detai results are shown in Figure 6.

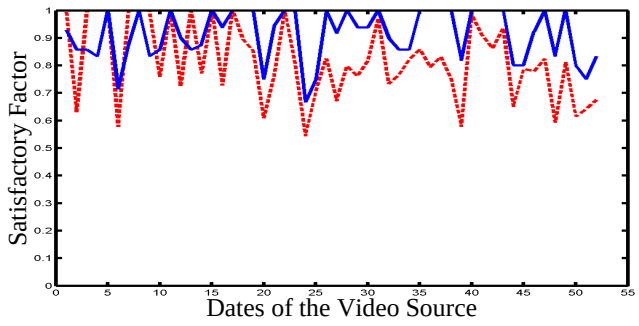

Figure 8: Comparison between the results obtained using only textual features (red dashed plot) and the results obtained using both the visual and textual features (blue solid plot). The plots are based on the matched stories from CNN and ABC in each day.

Since the textual features often provide more semantic inferences, there is a question on the impact of using visual features. In our experiments, the visual features are demonstrated to improve the satisfactory levels. If only textual features are used, a few outliers are generated because of the similarity of the key-words, even though they do not constitute semantic similarity if we form them into sentences. For example, key-words "war on drugs" refer to different topic than key-words "war in iraq". However, the textual correlation between them is large due to the common word "war". In these situations, the visual features help to further verify the results. A comparison between the results obtained using only textual features and the results obtained using both of the visual and textual features is shown in Figure 8. Based on the results, a clear improvement can be observed.

3.1 Application on Story Ranking

The results computed from the proposed semantic linking method are further used in the news story ranking. The ranking is proposed based on the repetition of the stories. In a general sense, more "interesting" or "hot" story topics appear more times and longer than other stories. In our formulation, we use the story's appearing frequency as the stories "interestingness" criteria. The story appearing the most is the most "interesting" story of that day.

Given two videos of the same date stamp, containing a and b stories, respectively, then a similarity matrix with size $[a] \times [b]$ can be constructed (Figure 4). Treating each story as a node in a graph, the similarity map is considered as a weighted bipartite graph (Figure 5). To rank the stories, we apply a breadth-first traversal technique on the bi-partite graphs to find the connected components of the stories. The stories are ranked by the sizes of their corresponding clusters. The cluster of the related stories can provide more coverage of the news topic, and it is better for further story summarization. The traversal algorithm is as follows:

1. Given videos with a and b stories, construct the semantic similarity matrix and establish the bi-partite graph by applying the preferred threshold;
2. Initialize the label $LL = 1$;
3. For node $i = 1$ to $(a + b)$, perform:
   - If the node is un-labelled, set $label(i) = LL$; Recursively label its connected neighbors with LL; Set $LL = LL + 1$.
4. Compute the sizes of the clusters with $label = \{1, \cdots, LL\}$;
5. Rank the story clusters based on their sizes, where larger size is assigned with higher ranking.

We have performed the story ranking method on the testing set and extracted the three most "important" story topics for each day. The detailed results are shown in Figure 9. In the results, the story clustering was performed automatically, while the story summary on each cluster is generated manually for presentation purpose.

4. CONCLUSIONS

In this paper, we have presented a semantic linking method for finding the similar news stories across different sources. The semantic correlation between two news stories is reflected by the visual similarity and the textual correlation. The key-frames of the stories are analyzed. The "body" regions are extracted from the facial key-frames for the persons of focus, and the non-facial key-frames are globally aligned using the Affine model to detect the repeating events. The language correlation is computed based on the automatic speech recognition (ASR) output of the videos. The visual and textual similarities are fused to provide the overall semantic linkage between the news stories.

The output results of the semantic linking task are further utilized in a news ranking task. The matched stories from the linking process are modelled as the vertices in a bipartite graph. Sub-graphs are detected using the connect-component technique, and the ranking of the stories is performed by analysis the component's size. The results for the story ranking task can be applied for better summarization of the stories, since more complete information is provided.

The proposed method has been tested on a large open-benchmark dataset, and very satisfactory results for both of the proposed tasks have been obtained.

5. REFERENCES

[1] J. Allan, J.G. Carbonell, G. Doddington, J. Yamron and Y. Yang, "Topic Detection and Tracking Pilot Study Final Report", *Broadcast News Transcription and Understanding Workshop*, 1998.

[2] W.G. Cheng and D. Xu, "Content-Based Video Retrieval Using Shot Cluster Tree", *International Conference on Machine Learning and Cybermetics*, 2003.

[3] J.L. Gauvain, L. Lamel and G. Adda, "The LIMSI Broadcast News Transcription System", *Speech Communication*, 37(1-2):89-108, 2002.

[4] C-W. Ngo, T-C. Pong and H-J. Zhang, "On Clustering and Retrieval of Video Shots Through Temporal Slices Analysis", *IEEE Transactions on Multimedia*, Vol.4, No.4, 2002.

[5] J-M. Odobez, D.G. Perez and M. Guillemot, "Video Shot Clustering Using Spectral Methods", *International Working on Content-Based Multimedia Indexing*, 2003.

[6] J. Sivic, F. Schaffalitzky and A. Zisserman, "Object Level Grouping for Video Shots", *European Conference on Computer Vision*, 2003.

[7] P. Viola and M. Jones, "Robust Real-Time Object Detection", *International Journal on Computer Vision*, 2001.

[8] W. Tavanapong and J.Y. Zhou, "Shot Clustering Techniques for Story Browsing", *IEEE Transactions on Multimedia*, 2004.

[9] Y. Zhai, A. Yilmaz and M. Shah, "Story Segmentation in News Videos Using Visual and Text Cues", *International Conference on Image and Video Retrieval*, 2005.

[10] D-Q. Zhang, C-Y. Lin, S-F Chang and J.R. Smith, "Semantic Video Clustering Across Sources Using Bipartitie Spectral Clustering", *International Conference on Multimedia and Expo*, 2004.

Date	First Ranking	Second Ranking	Third Ranking
19980204	Crisis on Iraq.	Clinton's Investigation.	Military Sex Hurrassment Case.
19980205	Crisis on Iraq.	Clinton's Investigation.	El Nino in California.
19980207	El Nino in California.	NBA, Michael Jordan.	Clinton's Investigation.
19980208	El Nino in California.	Clinton's Investigation.	Crisis on Iraq.
19980209	El Nino in California.	Clinton's Investigation.	Crisis on Iraq.
19980211	Clinton's Investigation.	Crisis on Iraq.	Tragedy in Italy by US Airforce.
19980212	Winter Olympics Games.	Presidential Veto Law.	Clinton's Investigation.
19980213	Surgeon General Nomination.	Market and Stocks.	Valentine's Day.
19980214	War on Illegal Drugs.	Clinton's Investigation.	Crisis on Iraq.
19980215	El Nino in California.	Hari Carry in Hospital.	Winter Olympic Games.
19980216	Taiwan Plane Crashed.	Crisis on Iraq.	Winter Olympic Games.
19980217	Crisis on Iraq.	Zamora's Case.	Clinton's Investigation.
19980218	Crisis on Iraq.	Winter Olympic Games.	Clinton's Investigation.
19980219	Clinton's Investigation.	Crisis on Iraq.	US Trade Deficit.
19980220	Crisis on Iraq.	Crisis on Iraq.	American Wrestlers Visit Iran.
19980221	Crisis on Iraq.	Biological Weapon in Nevada.	El Nino in California.
19980223	Crisis on Iraq.	Tornado in Florida.	Union Worker Strick.
19980225	Clinton's Investigation.	Tornado in Florida.	Market and Stocks.
19980226	Crisis on Iraq.	Winfield Opera's Case.	Internet Sales Tax.
19980302	Crisis on Iraq.	Princess Dianna's Accident.	Uranian Bombs.
19980304	Sexual Harrassment for Same Sex.	Military Sex Hurrassment Case.	First Female Space Shuttle Pilot.
19980305	Clinton's Investigation.	Market and Stocks.	Blood Transfusion.
19980306	Unemployment Rate.	Shooting of Lottary Workers.	Clinton's Investigation.
19980307	White Superemist Suspects.	Clinton's Investigation.	Holicopter Crashed in California.
19980308	El Nino in California.	Cirsis on Kosovo.	Clinton's Investigation.
19980309	Woodward's case.	Winter Weather Across the Nation.	Clinton's Investigation.
19980310	Military Sex Hurrassment Case.	Winter Weather Across the Nation.	Clinton's Investigation.
19980311	Coffi Annan Visit US.	Clinton's Investigation.	Winter Weather Across the Nation.
19980312	Bi-Partison Legislation in Senate.	Asteroid 1997-AF-11.	Winter Weather Across the Nation.
19980313	Clinton's Investigation.	Market and Stocks.	Military Sex Hurrassment Case.
19980315	Issues on the Kosovo's crisis.	Bishop Nominated in Vantican.	Clinton in Camp David.
19980316	Clinton's Investigation.	Vatican Released WW2 Documents.	Separation of Sex in Military.
19980317	Clinton's Investigation.	Market and Stocks	House Construction.
19980318	Clinton's Investigation.	IRS Reform Plan.	Crisis on Kosovo.
19980319	Clinton's Investigation.	Murcoch's Sale.	US Trade Deficit.
19980320	Clinton's Investigation.	Breast Cancer Pill Approved.	American Policy to Cuba.
19980321	Social Security Issue.	Tornado in Southern States.	Pope Johe Paul II in Negiria.
19980322	Nine Cubans flead to Bahamas.	Tornado in Southern States.	African Related Stories.
19980323	President Clinton in Africa.	Oil Prices Up.	Prostate Cancer.
19980325	Arkensas School Shooting.	President Clinton in Africa.	Low Mortgage Rate.
19980326	President Clinton in Africa.	Arkensas School Shooting.	Crisis on Iraq.
19980327	President Clinton in Africa.	Market and Stocks.	Personal Incomes Increase.
19980328	Arkensas School Shooting.	Hospital Deaths in California.	Explosion in Arizona.
19980329	Clinton's Investigation.	Peru Plane Crashed.	Hospital Deaths in California.
19980330	Market and Stocks.	New Home Sale Increase.	President Clinton in Africa.
19980413	Bank Merging.	IRS Tax Return Filing.	Annual Parade in Northern Ireland.
19980414	President Clinton is visiting Taxes.	IRS Tax Return Filing.	South Africa President Mandella.
19980416	Tornado in Southern States.	Former Cambodian Dictator Died.	Violence on Television.
19980417	Tornado in Southern States.	Clinton's Speech in Chile.	US Trade Deficit.

Figure 9: Table Summarizing the Story Ranking Results. The three most "interesting" topics are shown for each day in the dataset.