An Effective Keyword Extraction Method for Videos in Web Pages by Analyzing their Layout Structures

Jongwon Lee
Chungkang College
162 Chungkang-ro Majang-myun
Ichon-si, Gyunggi-do 467-744, Korea

Giseok Choi, Juyeon Jang, and Jongho Nang
Sogang University
1 Sinsu-dong
Mapo-gu, Seoul 121-742, Korea

Abstract- **This paper proposes an effective keyword extraction method for the Web videos by analyzing the structure of the Web pages. The proposed scheme calculates the relative importance (or weights) of the text blocks to a video by analyzing the distances of the text blocks to the video. This distance, called *the layout distance*, indicates a degree of relevance of text block to video, and could be estimated by analyzing the layout structure of Web pages. Since the Web pages with several videos such as Web pages posting UCC videos have a special layout structure, this layout analysis helps to precisely estimate the relevance of text block to the video. This weight of text block is used to compute the final weights of keywords extracted from that text block by analyzing their HTML tags and other well-known techniques such as TF/IDF. Some experiments with 1,087 Web pages that have total 2,462 videos show that the precision of the proposed extraction scheme is 17% higher than ImageRover[1].**

I. INTRODUCTION

There have been a lot of researches [1, 2, 6, 7, 8] that tried to automatically extract relevant keywords for images from text blocks of Web pages by analyzing their HTML tags and some well-known techniques such as TF/IDF. However, these techniques could not be directly applied to the videos in Web pages because the roles and desired keywords of videos are different from each other. From the authoring perspective, the images tend to play the assisting role of describing the texts, while the videos on Web sites are assisted by the texts. That is, the text is usually the main objects in the Web pages with images, whereas the video is usually the main objects in the Web pages with videos. From the search perspective, the image search is limited to find the certain objects in the text, while both objects and certain events are the subjects of the search for the video. Therefore, the techniques used to find the relevant keywords from the text block for images are not appropriate for keyword extraction for videos in WWW. In addition, since many previous studies [3, 4, 5] on the development of video annotation tools merely focus on indexing and annotation techniques based on domain-specific ontology, their effectiveness is simply limited to videos in a specific domain so that they could not be applied to a very large number of videos in WWW.

This paper proposes a new keyword extraction method for videos in Web pages by analyzing their layout structures. In the proposed scheme, the Web pages with videos are classified into four types according to the number of videos in the Web page and *the layout distances* between the video and the surrounding text blocks. For each Web pages types, we propose a new weighting scheme for the text blocks to videos based on their layout distance. Basically, their weights are inverse proportional to the layout distances to the video, however, they are adjusted by reflecting the structural characteristics of Web pages with videos. After assigning the weights to the text blocks, the keywords for the video are extracted from all text blocks in the Web pages together with their importance with some well-known techniques such as TF/IDF and HTML tag analyses. The final weights of keywords for the video are calculated by considering the importance of keywords within the text block and the layout distance of that text block to the video. Since the proposed scheme for calculating the degree of relevance of text blocks to videos reflects the structural characteristics of Web pages with videos, more useful relevant keywords could be extracted from the text blocks of Web pages. Experiments with 1,087 Web pages that have total 2,462 videos show the proposed scheme is especially effective for the videos in Web pages that have a complex layout structure, and the total precision is 17% higher than ImageRover[1, 2].

II. RELATED WORKS

The researches on video annotation are mainly divided into the manual annotation and automatic annotation methods. VAnnotator[3] is a simple manual annotation tool for which the user enters the annotation, and EVA[4] is also a Web based manual annotation tool. Although these manual annotation methods could produce a lot of useful annotations, it is labor intensive and requires a lot of time. It leads that this approach is not feasible for annotating the videos in WWW. The method by M. Bertini[5] builds an ontology for soccer by clustering visual information of the videos during the training phase and uses it for comparison of the input video. Although it is an automatic annotation method, it could be applied for extracting keyword of videos for a specific area such as soccer videos.

The researches on the automatic keyword extraction algorithm and system (for example, WebSeer [7], AMORE [8], ImageRover [1, 2]), have been focused mainly on the images in Web pages. They usually calculate the importance of the keywords for the image by analyzing the information in Web pages such as URL, HTML tags, image file name, the name used in the hyperlink, the distance between the image and the text blocks, the frequency of the keyword (TF/IDF), and additionally some visual information in images. Although these researches could extract some useful keywords for images in Web pages, they could not directly applied to extract the keywords for the videos in the Web pages since the roles of multimedia data in Web pages and

1-4244-1272-2/07/$25.00 ©2007 IEEE

the layout structures of Web pages are different from each other as shown in the following section.

III. A New Keyword Extraction Algorithm

A. *Differences between Web Pages with Images and Videos*

The text blocks of Web pages with images often contain the words that cannot be directly related to the images although they are used frequently in the text blocks. It is because the image is usually used as a supplementary object that helps to explain the concepts (or information) that the Web pages want to show. On the other hands, if there is a video in the Web page, its aim is mainly to explain the contents of the video so that the text blocks contains a lot of words that explain the contents of the video. Typical examples are Web pages posting UCC videos that are recently widespread explosively. Fig. 1 shows an example of Web page from CNN (http://www.cnn.com) in which both of image and video are included. The keywords related to the image and the video are marked manually in Fig. 1-(a) and Fig. 1-(b), respectively. As shown in these figures, there are only some keywords such as "military" and "soldier" that are directly related to the image, whereas there are a lot of words that are directly related to the videos because the text blocks are used to explain some situations in Iraq and the video is used to also explain this situation visually. This difference comes from the fact that the roles of image and video in Web pages are different from each other. The image is usually supplementary object in Web page, whereas the video is main one. This difference could be found clearly in the Web pages with UCC videos. Furthermore, the Web pages for posting UCC video are usually structured specially and there are some keywords that represent the metadata of the video such as "actor", "title", and "category" in the nearest text blocks. It leads that the keyword extraction algorithm for Web videos should reflect their structural characteristics and should utilize these video specific metadata.

Figure 1. Example of Web Page with Image and Video

B. *Layout Distance between Text Blocks and Video in Web Page*

The number of videos in the Web pages is varying with respect to the aims of Web pages. If there is only one video, the surrounding text blocks would have the same relevance to the video. In this case, the keywords extracted from these text blocks would have the same weights. On the other hands, if there are more than one videos in the Web page, the (visual) positions of text blocks are very important to measure the relevance to the video. The text blocks that are (visually) close to the video would contain some explanations on the contents of that video. The text blocks that are equally far from all videos would include some common descriptions on all videos in the Web page.

The simplest way to calculate the layout distance (or visual closeness on the screen) between the text block and video would be to measure the distance of two nodes at the VIPS DOM tree [9]. However, a direct application of VIPS DOM tree structure to measure the relevance of text blocks to video would cause some problems because of the structural characteristics of Web pages with videos. Since the text blocks that contain the descriptions on the video might be located direct top/bottom/sides of the video, basically the distance at the VIPS DOM tree could be used to measure the relevance of text blocks of video. However, the meanings of the distance at breadth and depth directions in DOM tree are different in the measuring the distance. Since the visual distance on the screen becomes larger as the text blocks is located further at the breadth direction of DOM tree, the layout distance should be proportional to the distance at breadth direction of DOM tree. However, the distance at the depth direction does not mean the visual distance, but the complexity of the layout structure of Web page. Therefore, the distance at the depth direction should not be considered in the relevance of the text blocks to the video. This structural analysis is especially important for the Web pages posting several UCC videos together.

Let us formalize the mechanism to measure the layout distance between the text blocks and videos. Let N_v and N_t be the video node and text node in VIPS DOM tree, respectively. Furthermore, let $P=\{p_0,p_1,...,p_n\}$ be a set of ancestor nodes of N_v, $Q=\{q_0,q_1,...,q_m\}$ be a set of ancestor nodes of N_t, and p_i (or q_j) be the nearest common ancestor of N_v and N_t. Then, the layout distance between N_v and N_t, $d(N_v,N_t)$, is defined as follows;

$$d(N_v,N_t)=L(N_v)-L(p_i)-1 \qquad (1)$$

where

$$L(N_i)=\begin{cases}0, \text{if } N_i \text{ is the root node}\\ L(N_k)+1, \text{where } N_k \text{ is the parent node of } N_i\end{cases}$$

This equation guarantees that the layout distance is proportional to the distance only at the breadth direction and irrelevant to their distance at the depth direction.

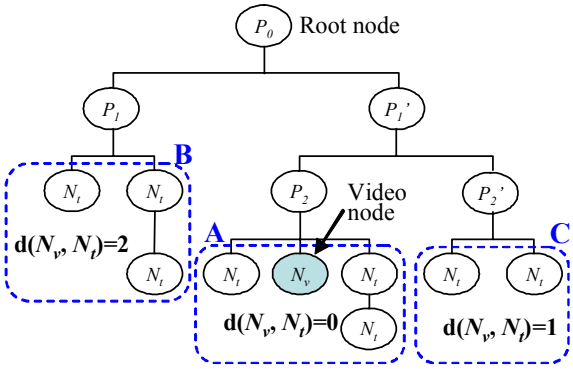

Figure 2. Example of Layout Distance in DOM tree

Fig. 2 shows an example of calculating the layout distance between the video node N_v and the text node N_t in DOM tree, in which P_0 is the nearest common ancestor node of video node and all text nodes in B branches, and P_1' is the

nearest common ancestor node of video node and all text nodes in C branches. As shown in this example, the layout distances of the video node between the text nodes in B are calculated as all 2's, whereas the ones between text nodes in C are calculated as all 1's. Using this layout distance calculation scheme, we could identify the relative relevance of text blocks to the video effectively, and it helps to extract the useful keywords from Web pages.

We have analyzed the distribution of Web pages with videos with respect to the number of videos (n) and the maximum layout distance of text blocks in that Web page ($d^{\max}$), and classified into four types. Table 1 shows the distribution of 1,087 Web pages with total 2,462 videos for each type. From this Web page distribution analysis, we could find out that a small number of Web pages (about 6.5%) have embedded a lot of videos (about 58.7%). It means that the structural characteristics of Web pages should be carefully analyzed in order to extract the useful keywords, especially in the case of Type 3 and 4.

TABLE I
CLASSIFICATION OF WEB VIDEO PAGE (T=1)

Type	*Number of Videos*	$d^{\max}$	Distribution of Web pages	Distribution of videos
Type 1	$n = 1$	$d^{\max} \le T$	37.0 %	16.3 %
Type 2		$d^{\max} > T$	56.6 %	25.0 %
Type 3	$n > 1$	$d^{\max} \le T$	3.6 %	28.8 %
Type 4		$d^{\max} > T$	2.9 %	29.9 %

If the maximum layout distance of text block to the video is smaller than T (Type 1 and 3), all text blocks would be relevant to the video and the weights should be the same regardless of their layout distances. However, if the maximum layout distance of text block to the video is larger than T (Type 2 and 4), the degree of the relevance of the text blocks whose distance are larger than T would be inverse proportional to the layout distance because the text blocks would be belonging to other video or they would be common text blocks. Of course, if there are more than one videos and several text blocks (Type 3 and 4), the text blocks whose layout distances are less than T should be associated with only one video and do not used to extract the keywords of other videos. These structural analyses lead following equation for computing the weight of k-th text block to the j-th video in the Web page A, W_j^k, where $d(N_j, N_k)$ represents the layout distance between the video node N_j and the text node N_k,

$$W_j^k = \begin{cases} 1, \text{if } GetType(A) \in \{\text{Type 1, Type 3}\} \text{ or } d(N_j, N_k) \le T \\ 1/d(N_j, N_k), \text{otherwise} \end{cases} \quad (2)$$

where $GetType(\,)$ is a function returning type of Web page.

C. Calculating the Weights of Keywords

In the proposed keyword extraction scheme, the keywords are extracted from each text block with their own weights, and these weights are combined with the weights of text block in order to assign the final weights of the keyword, as shown in Fig. 3.

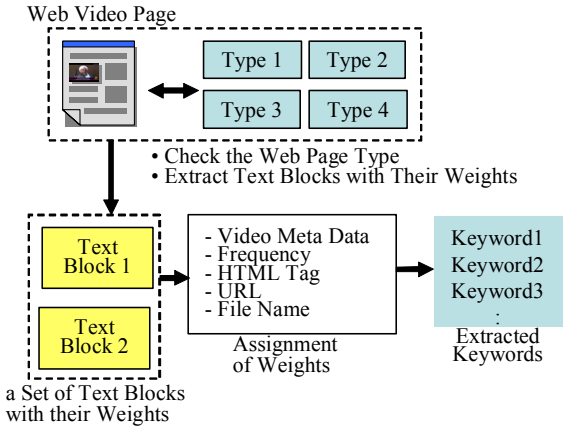

Figure 3. Overall Flow of the Keyword Extraction Method

The keywords extraction methods from the text blocks are studied intensively. Some well-known techniques [1, 6, 7] are to give a high weight to the keyword when it is marked with special HTML tags such as **<HEAD>**, **<ALT>**, and **<Anchor>**, when it is included in the file name, and when it appears more frequently than other keywords (TF/IDF). In addition to these techniques, we have used some heuristics that are useful to extract the keywords for the video. As shown in Table 1, there are some Web pages that embed a lot of videos, and these pages are usually used to post the UCC videos. In this case, the text block whose layout distance to the video is less than T usually contains the metadata for that video. Examples of metadata are "title", "author", "tags", and "category". The keywords used for the attribute values of these metadata could be very good keywords for the video, and the proposed scheme gives an extra weights to these kinds of keywords. Let tb^k be k-th text block of Web page A, and $c^{k,i}$ be i-th keyword of tb^k. Then, the final weight of $c^{k,i}$ to j-th video, $W_j^{k,i}$, is calculated with following equation;

$$W_j^{k,i} = W_j^k * weight(tb^k, c^{k,i}) \quad (3)$$

where $weight(tb, c)$ is a function returning the weight of keyword c in text block tb using some well-known techniques such as HTML tag, file name, and our video specific metadata analyses.

D. Keyword Extraction Algorithm

An overall keyword extraction algorithm proposed in this paper is shown in Fig. 4. It first extracts the videos and text blocks from the Web page, and decides its type using the criterion presented in Table 1. Then, for each pair of text block and video, it computes the degree of relevance using (2). If the layout distance between a text block and a video is less than T, this text block is used to extract the keywords for that video only, as mentioned before. For all keywords in the text block, their weights within the text block are computed with the well-know techniques such as used in ImageRover [1, 2] (the procedure $weight(tb^k, c^{k,i})$). Finally, the most weighted keywords for each pair of text block and video are selected and returned.

```
// TB : a set of text blocks in A , TB = {tb_1, tb_2,...,tb_m}
// V : a set of videos in A, V = {v_1, v_2,...,v_n}
// C^k : a set of keywords in tb_k , C^k = {c^{k,1}, c^{k,2},...,c^{k,l}}
// Tag^j : a set of selected keywords for v_j
// VTag : a set of pairs of video and keywords;

Procedure KeywrodExtractionfromPage(A) {
  TB = ExtractTextBlock(A) ; // Get text blocks from A
  V = ExtractVideo(A) ; // Get videos from A
  A^type = GetType(A) ; // Get the Web page type of A
 VTag = {};
  for ∀v_j (v_j ∈ V) do {
      for ∀tb_k (tb_k ∈ TB) do {
          Compute W_j^k with (2);
          for ∀c^{k,i} (c^{k,i} ∈ C^k) do {
              W_j^{k,i} = W_j^k * weight(tb^k, c^{k,i}) ;
          }
      }
      Tag^j =GetKeywordswithMaxWeights() ;
      VTag = VTag ∪ {(v_j, Tag^j)};
  }
  return(VTag ) ;
}
```

Figure 4. Proposed Keyword Extraction Algorithm

IV. Experimental Evaluation

We have implemented the proposed keyword extraction algorithm using VIPS DOM tree [9], and experimented with 1,087 Web pages that have total 2,462 videos. The ground truths of keywords for each video are generated manually, and they are compared with the keywords extracted with our algorithm and ImageRover [1, 2]. The parameters used to extract the keywords from the text blocks are set to the same as ImageRover except giving the extra weights to the keywords related to the video specific metadata. Fig. 5 shows the precisions of the proposed method and the conventional method (actually, the algorithm used in the ImageRover [1, 2]) for four Web page types while varying the number of desired keywords. As shown in this experiment, the precisions are lowered as the number of desired keywords is increased. The precision of the proposed method is similar to the conventional one when the Web pages belong to Type 1 because of their simple structures, as shown in Fig. 5-(a). For the Web pages belong to Types 2, the proposed method produce a somewhat higher precision over the conventional method as the number of extracted keywords is growing as shown in Fig. 5-(b). It is caused by the fact that the conventional method does not take account into the structural characteristics of the page sufficiently, while the proposed method extracts the keyword by evaluating the relationship between the video and surrounding texts. This superiority is shown clearly when the Web page types are Type 3 and 4, as shown in Fig. 5-(c) and (d). Since the proposed scheme reflects the structural characteristics of the Web pages posting a lot of videos in a bulletin board style (that is a typical Web page style of UCC site), the precision could be improved properly. The overall precision of the proposed scheme for 1,087 Web pages that have total 2,462 videos is 17% higher than that of ImageRover[1 ,2].

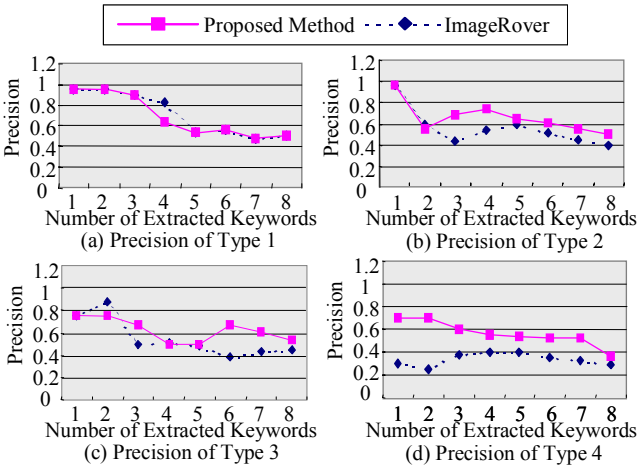

Figure 5. Experimental Comparison

V. Conclusion

As the videos are posted widely and the demands for searching the videos in WWW are increased rapidly, an automatic keyword extraction scheme for the videos in WWW is required. Some previous schemes developed for the images could not be applied directly to video annotation because the roles and posting style of the images and videos in Web pages are different from each other. This paper proposed a new keyword extraction method for the videos embedded in the Web pages by analyzing the structural characteristics of the Web pages. It first classified the Web pages into four types, and assigns different weights to the text blocks according to their layout distances and Web page types. Experimental results showed that the proposed method could extract the useful keywords more precisely than ImageRover because it sufficiently reflected the structural characteristics of Web pages posting a lot of videos in a bulletin board style. The proposed method could be used to build a powerful video search system for WWW.

References

[1] S. Sclaroff, L. Taycher and M. L. Cascia, "ImageRover: A Content-Based Image Browser for the World Wide Web," in *Proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries*, pp. 2~9, 1997.

[2] M. L. Cascia, S. Sethi and S. Sclaroff, "Combining Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web," in *Proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries*, pp. 2~9, 1998.

[3] M. Costa, N. Correia and N. Guimaraes, "Annotations as Multiple Perspectives of Video Content," in *Proceedings of the 10th ACM International Conf. on Multimedia*, pp. 283-286, 2002.

[4] T. Volkmer, J. R. Smith, and A. Natsev, "A Web-based System for Collaborative Annotation of Large Image and Video Collections," in *Proceedings of the 13th annual ACM International Conf. on Multimedia*, pp. 892-901, 2005.

[5] M. Bertini, A. Bimbo and C. Torniai, "Automatic Video Annotation using Ontologies Extended with Visual Information," in *Proceedings of the 13th annual ACM International Conf. on Multimedia*, pp. 395-398, 2005.

[6] S. Rui, W. Jin, and T. Shua, "A Novel Approach to Auto Image Annotation Based on Pair-wise Constrained Clustering and Semi-naïve Bayesian Model," in *Proceedings of IEEE 11th International Conf. on Multimedia Modeling*, pp. 322-327, 2005

[7] C. Frankel, M. J. Swain, and V. Athitsos, *WebSeer: An Image Search Engine for the World Wide Web*, Technical Report 96-14, University of Chicago Computer Science Department, 1996.

[8] S. Mukerjea and J. Cho, "Automatically Determining Semantics for World Wide Web Multimedia Information Retrieval," in *Journal of Visual Languages and Computing*, vol.10, pp. 585~606, 1999.

[9] D. Cai, S. Yu, J. R. Wen and W. Y. Ma, *VIPS:a Vision-based Page Segmentation Algorithm*, MSR-TR-2003-79/Microsoft Research, 2003.