MINING TEXT AND VISUAL LINKS TO BROWSE TV PROGRAMS IN A WEB-LIKE WAY

Xin Fan[1], Hisashi Miyamori[2], Katsumi Tanaka[2,3], Mingjing Li[4]

[1]Dept. of EE and IS, Univ. of Science and Tech. of China, P.R. China
[2]National Institute of Information and Communications Technology, Japan
[3]Graduate School of Informatics, Kyoto University, Japan
[4]Microsoft Research Asia, P.R. China

ABSTRACT

As the amount of recoded TV content is increasing rapidly, people need active and interactive browsing methods. In this paper, we use both text information from closed captions and visual information from video frames to generate links to enable one to explore not only the original video content but also augmented information from the Web. This solution especially shows its superiority when the video content cannot be well represented only by closed captions. A prototype system was implemented and some experiments were carried out to prove the effectiveness and efficiency.

1. INTRODUCTION

Nowadays, millions of homes enjoy interactive television. Web-like technologies such as pay-per-view, personal video recording, video-on-demand etc. are revolutionizing our way of watching television. Watching television is becoming an active entertainment and the viewers are evolving from a passive to an active role.

Although television programs are well-edited by professionals, the details and scope of the information are limited to meet public taste. More additional information and related materials are required for the diversity of personal information needs. In some web-based TV recommendation systems [1, 2], users can view their personalized TV guides in specially customized HTML or WML pages. Furthermore, some efforts [6, 8] have been initiated to provide links to further information and data inside television programs.

In [7], the concept of "webified video" is proposed, which helps users to acquire value-added content from Web content when they view the original TV programs. The TV program is divided into different levels of segments using close captions attached to it. Then the Web content relevant to these structured text data is retrieved and hyperlinks are created to associate the TV program with the Web pages. Accordingly, the videos and links to further information are integrated in a Web-based browser. However, this solution is based on the assumption that the topic of a video scene can be represented by word distribution of closed captions. Many TV programs, such as news videos, can conform well to the assumption. However, in some cases such as animations or dramas, the video structure does not directly correspond to the closed captions and the relationship between visual content of the scene and the closed captions is weak. Therefore, it is often difficult to generate video structures and retrieve the correct related information only by closed captions

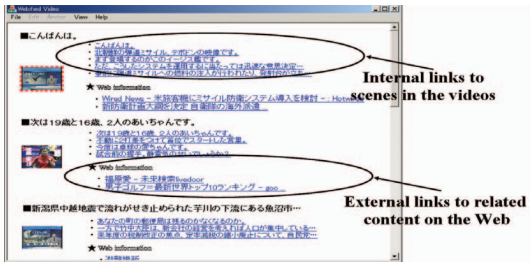

Figure 1. The Web-like TV browsing interface

In this paper, we utilize both the text information from closed captures and visual content from video frames to produce external links to Web content and internal links to other scenes in TV programs. Considering that pure text information may not well describe the TV content, we first combine text information and visual information to generate a hierarchical video structure. Based on the structured video, relevant Web content is respectively retrieved based on similarity of text and visual appearance. Then the Web content is fused in the TV program display in a zooming and adaptive manner. The external links to related Web content and internal links in TV programs are listed on the side storyboard as Figure 1. Viewers can easily access related information in detail and from different viewpoints. They can also look through the video content from the information provided by fast forwarding. Since the links can be built by the visual similarities among the videos, the viewer can even switch to another scene with an identical character or an identical prominent object.

The system mainly consists of three modules: hierarchical video structure generation, creation of internal and external links and adaptive information display.

2. HIERARCHICAL VIDEO STRUCTURE GENERATION

We first carry out hierarchical segmentation to extract three level structures in the TV program, which are defined as segments, topics and sub-topics sequentially in our system. As mentioned above, the closed captions may not synchronize well with the video content in some types of videos. In this case, the word distribution in closed captions is not so well-regulated that it is difficult to judge video boundaries by pure text information. Therefore, a two-step process is adopted in our solution. First, the video is initially divided into a series of segments according to the statistical differences of visual content between video frames. Based on the visual content-based segmentation results, the second step is to adjust the segment boundaries using the information of word contributions in corresponding closed captions, followed by iterant dividing processes into topics and sub-topics.

Since it is not necessary to achieve very high accuracy in the initial segmentation, a common color histogram method is adopted in the initial visual content based segmentation. A running histogram method similar to the algorithm described in [10] is performed and two thresholds are used. The high threshold is for declaring a cut and the low threshold is assumed for the gradual shot transition.

By the above segmentation method based on visual features, the original TV program S can be divided into a series of initial segments $S_1S_2 \ldots S_n$. Here we assume the prior probability of segmentation Si to be normal with respect to the distance of frame numbers $D(N_i, N_x)$ between the detected boundary frame i and a certain frame x, that is:

$$P(S_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{D(N_i, N_x)^2}{2\sigma^2}} \quad (1)$$

Then we consider the statistical word distributions in corresponding closed captions. Let $W = W_1W_2 \cdots W_m$ be the text span attached with the video $S = S_1S_2 \cdots S_n$ consisting of n words. Therefore the posterior probability of segmentation S_i is presented as:

$$P(S_i \mid W) = \frac{P(W \mid S_i)P(S_i)}{P(W)} \quad (2)$$

According to the Maximum A Posterior (MAP) principle, the most likely segmentation $\hat{S}_i$ can be denoted as:

$$\hat{S}_i = \arg\max_{S_i} P(W \mid S_i)P(S_i) \quad (3)$$

since $P(W)$ is a constant for a given text span W. $P(W \mid S_i)$ can be represented by the number of each different word in the text span W [9]. We update the segmentation $S = S_1S_2 \cdots S_n$ using the dynamic programming approach in [7].

Similar text based segmentation in [7] is done to the updated video segments $S_1'S_2' \cdots S_l'$ recursively.

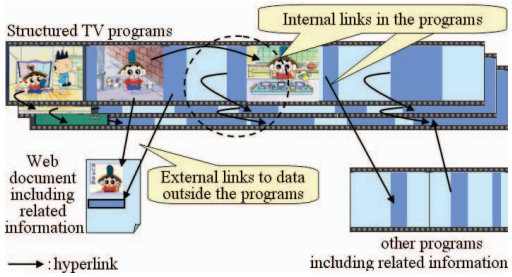

Figure 2. Illustration of links within and out of TV programs

Consequently, the segments are divided into topics and the topics into sub-topics. The original TV program is divided into three-level structured data hierarchically.

3. EXTERNAL AND INTERNAL LINK CREATION

As shown in Figure 2, we produce two kinds of links to be displayed on the storyboard beside the video shows: external links to the related Web pages and internal links to the related video topics. Both are created respectively by text-based retrieval and visual content matching.

3.1. Creation of text-based links

In this text based processing step, a complementary information retrieval method [6] is used to find related Web content and similar topics in the TV programs. First, topic structures are extracted from hierarchically segmented TV programs. Four types of queries are generated using the topic structures: content-deepening, subject-deepening, subject-broadening and content-broadening queries. These queries are issued to the Web search engine such as Google. The top results are collected and their URLs are integrated as text-based external links. Similarly, the internal links are built up according to the similarity of extracted topics.

3.2. Creation of visual content-based links

The visual video content is also used to update the external and internal links on a storyboard. It can be described in an example scenario. If the viewer is interested in an object in the current scene when watching TV, he/she can simply pause the show and the current picture will be matched in the image database to find online resources and other possible scenes including an identical object.

Generally, the task of finding similar scenes by visual appearance turns to the traditional Content Based Image Retrieval (CBIR) method. However, systems built this way generally aim at finding images with some similar visual features to the query image. This coarse-grain matching method is not suitable for our task because of its limited accuracy. We expect to find pictures containing the same prominent object or scene as the query image.

Therefore, a fine-grain matching scheme is designed based on local descriptors of key points in the image as shown in Figure 3. The image database can be composed of

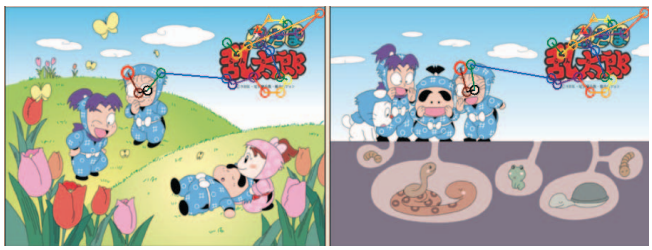

Figure 3. Example of image matching with local features of key points (Circles represent the matched key points and lines can be regarded as adding sequences of matched key points in the RANSAC algorithm.)

Web image collections or the video frames from TV programs. For the former image database, as shown in Figure 4, key frames are extracted from video segments to search the matched images. Cached web pages containing those matched images would be selected to build up external links for reference. For the latter database, current scene can be used to find the matched video frames to create the internal links.

The SIFT algorithm [5] is employed to detect local key points for all the images in the database. The 128-D SIFT features around the key points are extracted to be represented by local descriptors. The SIFT descriptor deals well with image scaling, crop, shearing, rotation, partial occlusions, brightness and contrast change etc. The descriptors of images in the database are computed as feature vectors and sequentially stored in a feature database.

The picture for query is also processed in a same way as above and produces a set of local feature vectors. The scenes with identical objects are assumed to be proportional to the similarity of the feature vectors. Since a single image may generate hundreds to thousands of key points and features, an individual query may need to match millions of high-dimensional local features in the image database. The method of sequential scan for matching is computationally prohibitive in a large image set. The common high-dimensional indexing solution is to regard a local feature as a point in high-dimensional feature space. Then a similarity search, such as a nearest-neighbor search or a ε-range search, is performed within the feature space to achieve effective retrieval of similar local features.

In the image matching process, we adopt an index structure based on unsupervised clustering with a Growing Cell Structures (GCS) [4] artificial neural network, which has good performance in higher dimensional space. Suppose there are N feature vectors in the feature database, we choose $[\sqrt{N}]$ cells, a two-dimensional neighborhood relationship and the KNN cluster algorithm here. For each local point in the query picture, we select n nearest local points with neighborhood relationships in feature space as the matched correspondences. Moreover, if there are only m ($m < n$) points with neighborhood relationship to the query point, n equals m. The value of n depends on the whole point number in the image database. We experimentally bestow $n = 20$ in the prototype implementation.

Figure 4. Generating visually related external links from video frames

A direct way to judge the similarity of the scene can be to vote on the amount of matched points for each image in the image database. However, there would be many outliers in the matched point pairs by this method because the pair is from a similarity search. In order to eliminate such false matches, we validate the matched points in each image using spatial constraints.

Generally transform changes for the identical object in different scenes can be modeled by the combinations of the following three transforms: rotation transform, scale transform and shear transform. The changes can be represented in an affine transform [3] between different coordinate systems as follows:

$$X_c = X_a - \overline{X_a},\ Y_c = Y_a - \overline{Y_a}, \tag{4}$$

$$\begin{pmatrix} X_d \\ Y_d \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} X_c \\ Y_c \end{pmatrix} \tag{5}$$

$$= \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} \alpha & 0 \\ 0 & \delta \end{pmatrix} \begin{pmatrix} 1 & \beta \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_c \\ Y_c \end{pmatrix} \tag{6}$$

where X_c, Y_c are the normalized coordinates of the query point a, and X_d, Y_d are the normalized coordinates of a matched point d. The three coefficient matrices in Equation 6 respectively correspond to rotation transform, scale transform and X-shear transform of the coordinates.

Considering the presence of many data outliers, the RANdom SAmple Consensus (RANSAC) algorithm [3] is adopted to judge if the pairs can be aligned by a certain affine transform. The process is described as follows. If there are some matched point pairs between a query image and a database image, two of the pairs will be selected to estimate the coefficients in Equation 5 as an initial model. Some other point pairs will be sampled and added to fit the model. After a given K times iterations, if the probability that a given trial is a failure is below a specified value, the image pair is judged as a matched pair. Thus, the visual links can be established by the validated image matches with high precision.

4. ADAPTIVE INFORMATION DISPLAY

A similar zooming user interface in [7] is adopted to enable viewers to access more information in different levels of detail. Zoom-in and zoom-out functions can switch a display containing only one TV program and one containing

several programs. The search range for internal links also varies within the displayed content.

5. EXPERIMENTS

A prototype system was implemented and we carried out some experiments on it to evaluate our solution from the viewpoints of computational cost and accuracy of retrieved information. We used three animations recoded from NHK in different time. The duration of each was 721 seconds and the resolution was 640x480 pixels. An image database was set up, which consisted of 302 images from captured TV frames, 7137 images from NHK animation websites and 2264 images from Google Image Search results by eight animation names. The excessively small or banner-like images had been removed beforehand. The image dimensions ranged from 100 to 1440 pixels. We extracted the SIFT features and built up the index for all the images offline. The prototype ran on a PC with an Intel P4 3.2GHZ CPU, 2G RAM and MS Windows XP system.

First, a preliminary user study was conducted to give a comparative evaluation of current retrieval results and the results retrieved only by closed captions. All the participants found the function of listing links on the storyboard can provide useful complementary information for them. Most of them (5 out of 6) considered the results produced by the current solution to be superior to the output results based only on closed captions since more visual related information were provided.

Second, the above three animations were structurized by the proposed method. We extracted 552 key frames from them as query pictures to generate visual links. The accuracy of links by text-based and visual content-based retrieval methods are listed in Table 1.

Table 1. Accuracy of generated links

Visual content-based search method	*Text-based search method*
0.645	0.583

This indicates that more than half of the links provide related information which is judged by the users, i.e. they can easily find helpful information through the given links.

Third, we analyzed the computational cost for all the above three animations and the time consumption is recorded stage by stage in Table 2.

Table 2. Time consumption of each processing module (sec)

Visual content-based segmentation	*Text-based segmentation*	*Text- based link creation*	*Visual link creation*
2183.0	2.9	2.8	3.4

It can be seen that visual content-based segmentation is the most time-consuming procedure in the whole process. Fortunately, this segmentation can usually be processed beforehand for recoded TV programs.

6. CONCLUSIONS

We have proposed a solution to generate links for browsing recoded TV programs in a Web-like manner. Unlike previous work, a hierarchical video structure is generated by integrating both the text information from closed captions and visual information from video frames. Both the text-based search method and the visual content-based search method are employed to generate external links to related Web pages and internal links to other scenes. In future work, we plan to improve the search method, particularly the image matching algorithm.

7. ACKNOWLEDGEMENTS

The authors thank Mr. Menglei Jia and Dr. Xie Xing for sharing the tool for high-dimensional indexing and Dr. Qiang Ma for helpful discussions.

8. REFERENCES

[1] P. Baudisch and L. Brueckner, "TV Scout: Lowering the entry barrier to personalized TV program recommendation," *Proceedings of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems*, Malaga, Spain, pp. 58-67, May 2002.

[2] P. Cotter and B. Smyth, "PTV: Intelligent Personalised TV Guides," *Proceedings of the 12th Innovative Applications of Artificial Intelligence (IAAI-2000) Conference*, Austin, USA, pp. 957-964, Aug. 2000.

[3] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*: Prentice Hall Professional Technical Reference, 2002.

[4] B. Fritzke, "Growing cell structures - a self-organizing network for unsupervised and supervised learning," *Neural Networks*, Vol. 7, No. 9, pp. 1441-1460, 1994.

[5] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91-110, 2004.

[6] Q. Ma and K. Tanaka, "Topic-Structure-Based Complementary Information Retrieval and Its Application," *ACM Transactions on Asian Language Information Processing*, to appear.

[7] H. Miyamori and K. Tanaka, "Webified Video: Media Conversion from TV Programs to Web Content for Cross-Media Information Integration," *Proceedings of the 16th International Conference on Database and Expert Systems Applications*, Copenhagen, Denmark, pp. 176-185, Aug. 2005.

[8] K. Sumiya, M. Munisamy, and K. Tanaka, "TV2web: Generating and browsing Web with multiple LOD from video streams and their metadata," *Proceedings of the 13th international World Wide Web conference*, New York, NY, USA, pp. 398-399, May 2004.

[9] M. Utiyama and H. Isahara, "A statistical model for domain-independent text segmentation," *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Toulouse, France, pp. 491-498, July 2001.

[10] H. J. Zhang, A. Kankanhalli, and S. Smoliar, "Automatic Partitioning of Full-motion Video," *ACM Multimedia Systems Journal*, Vol. 1, No. 1, pp. 10-28, 1993.