Visualizing Multimedia Content on Paper Documents: Components of Key Frame Selection for Video Paper

Jonathan J. Hull, Berna Erol, Jamey Graham, and Dar-Shyang Lee
Ricoh Innovations, Inc.
Menlo Park, CA
{ hull, berna_erol, jamey, dsl}@rii.ricoh.com

Abstract

The components of a key frame selection algorithm for a paper-based multimedia browsing interface called Video Paper are described. Analysis of video image frames is combined with the results of processing the closed caption to select key frames that are printed on a paper document together with the closed caption. Bar codes positioned near the key frames allow a user to play the video from the corresponding times. This paper describes several component techniques that are being investigated for key frame selection in the Video Paper system, including face detection and text recognition. The Video Paper system implementation is also discussed.

1. Introduction

Searching for information in a video recording can be intensely frustrating. After several attempts to find something with repeated instances of fast-forward, play, rewind, etc., all but the most motivated users will easily give up and save this task for another day. Browsing is an alternative to active search that is used by many online video retrieval interfaces [2, 3, 9, 10]. However, these methods are still limited by what can be displayed on a PC monitor. On the other hand, it's well known that people can *read* three times faster than they can listen which suggests that a paper-based interface could provide a more efficient search mechanism than an online media viewer.

Video Paper is a paper-based interface for accessing digital video [5]. The closed caption transcript that's often provided with a television program is formatted and printed together with key frames selected from the video. Bar codes are also included so that with a simple scan, a user can start playing the video at the corresponding point. The objective is an interface that's as easy to use as a newspaper, taking advantage of the high resolution and portability of paper, and which provides a means for random access to a medium (video) that has traditionally required tedious searching for effective utilization.

An example of using Video Paper for analyzing a news broadcast, like the one shown in Figure 1, would be looking for information about the stock price of the Martha Stewart Company. A quick glance over the document lets a user notice the graph on the left side and focuses his attention on that section, thus eliminating the need to look elsewhere.

The key frame selection algorithm is an important part of the Video Paper system. There is only a limited amount of space available on each page and we would like the user to understand as much of the recording as possible by merely glancing at the images. In contrast to the prior art, we have access to several sources of information that could help us select key frames. These include not only the text transcript, but also the user's interest profile and the characteristics of the key frames that the user would like the system to associate with specific topics.

The Video Paper system records television programs, converts the data to MPEG2, and stores the file on a server. The Video Paper formatting software is automatically applied to the MPEG2 file and a pdf file is output. A remote control device is comprised of a PDA (e.g., Compaq iPAQ) with a bar code reader and a wireless interface (e.g., 802.11b). Software on the PDA decodes scanned bar codes and sends commands to the server that controls replay of the video on a television attached to the server. In addition to the bar codes nearby key frames, meta control bar codes are included on the bottom of each page that pause the replay, rewind, fast forward, or display the closed caption on the television.

An alternative version converts the video file to a format suitable for display on the PDA. This provides a portable implementation that can be used when the server is inaccessible.

The rest of the paper describes components of a key frame selection algorithm that are being investigated. An objective of this work is to determine whether the selection of key frames based on image analysis can provide a more effective browsing strategy than simple time-based sampling.

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 IEEE

Figure 1. Example Video Paper document.

2. Key Frame Selection Algorithm

In the proposed key frame selection algorithm, image analysis routines, such as face detection and text recognition, are applied to the video data and their results are combined with an analysis of the closed caption transcript to suggest key frames that could be useful to someone who browses a Video Paper document.

The face detection algorithm first finds skin pixels in the normalized RG-space [12]. Small holes in skin-colored regions are removed by a morphological closing and then connected component analysis is used to identify face region candidates. In environments with complex backgrounds, many objects, such as wood, clothes, and walls, may have colors similar to skin. Therefore, further analysis of the skin-colored regions, using luminance variation and geometric feature analysis, can eliminate non-face regions. A confidence score is associated with each image based on the size of the face region and the percentage of skin pixels detected in the best-fitting ellipse. Some example face detection results are shown in Figure 2 (a).

The text recognition system applies an adaptive binarization and OCR procedure. The objective is to find the parameters for the binarization algorithm that yield the highest OCR accuracy. Some example binarization and text recognition results are shown in Figure 2 (b) and (c).

The effect of a minor difference in thresholding on OCR accuracy can be significant, especially for video data, since there is usually little text displayed on the TV screen at any one time. This is taken into account by a text-tracking algorithm such as [8]. Also, the presence of the same text in the closed caption and a video frame is evidence that the corresponding frame should be placed on the paper document. An example of the utility this could provide a reader is shown in Figure 2 (b) – (d).

A profile matching routine detects the presence of topic-related keywords and phrases in the closed caption transcript and optionally annotates them in the paper-based interface. Topics and keywords are stored in a user's profile (this is an application of the Reader's Helper profile[4] that was originally developed for improving the readability of html documents). A profile

thus provides a convenient means for representing information that a user would like the system to find in a multimedia document.

The document generation part of the system uses a template to choose key frames with the corresponding spatial constraint. For example, with 14 key frames per page (see Figure 1) and a typical news broadcast, about six minutes of video can be shown on each sheet. A single key frame is selected from each 25 seconds of video and placed near the corresponding portion of transcript.

3. Experimental Results and Applications

A complete working prototype for the Video Paper system has been developed. Running on a 2 GHz Windows 2000 system, less than one minute of run-time is required to produce a Video Paper PDF file for a one-hour recording.

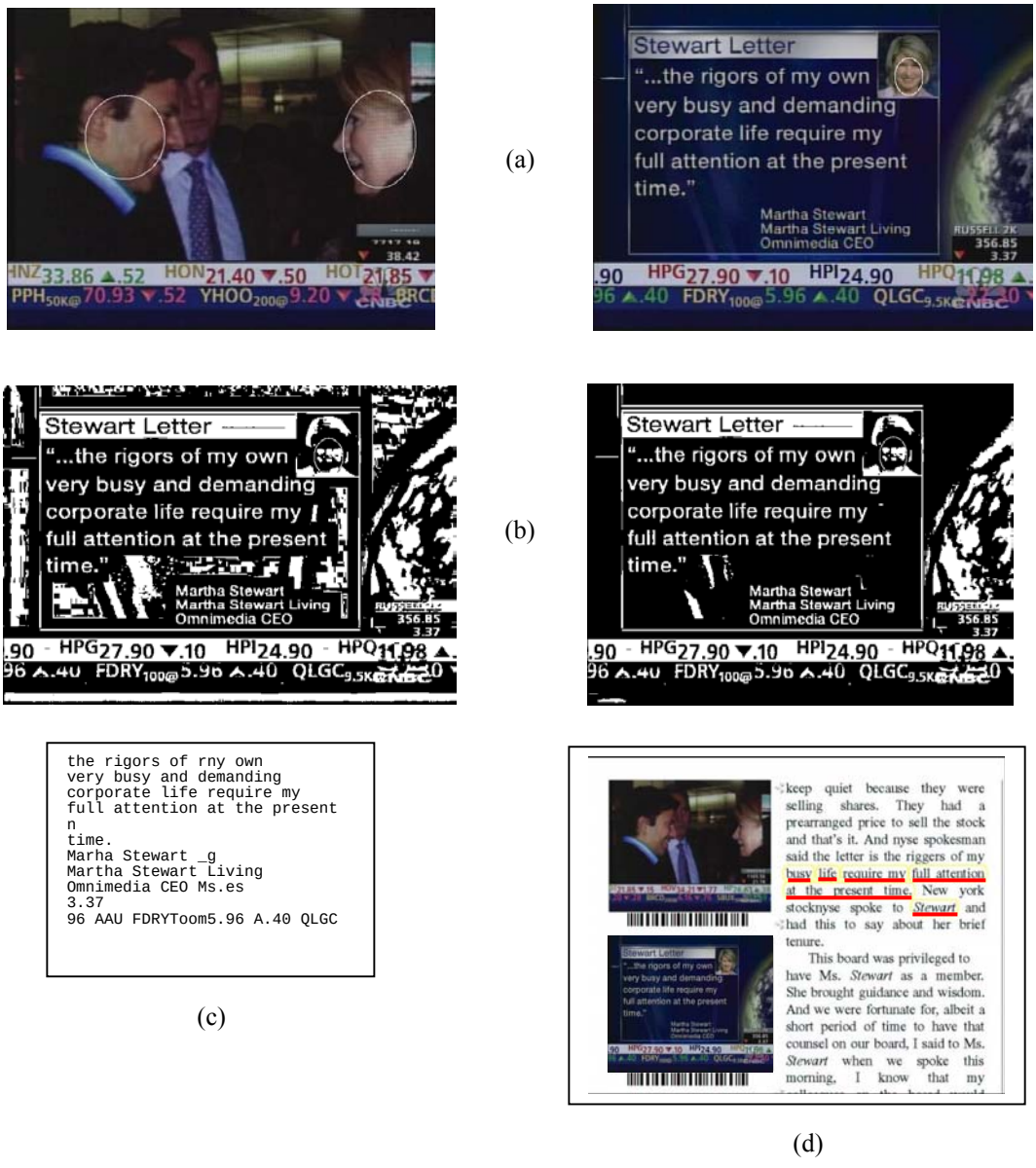

Figure 2. Face detection examples (a), binarization results (b), OCR results (c), and matching text in the closed caption (d).

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 IEEE

This system has been running on a regular basis in our lab. The remote control mentioned in the Introduction is connected to our WiFi LAN and allows users to replay Video Paper documents on demand.

Video Paper is also being used for oral histories [7]. These are interviews of the observers or participants in historical events. Typically, the interviews are recorded and transcribed. Because of the difficulty in accessing the audio, subsequent analysis is usually limited to reading the transcript. Oral historians have recently started video taping their interviews, but these are even more difficult to access afterward because of the difficulty in synchronizing the video replay with the transcript. A recent experiment supplied Video Paper to 13 subjects who reviewed two 75-minute interviews. Post-hoc analysis showed they were completely at ease with the concept of Video Paper and reported that it provided substantial benefits with minimal overhead. One unexpected result was that users' multi-processing capabilities were enhanced. They could listen to one section of a document while browsing another.

4. Conclusions

A novel solution for video browsing and retrieval was described. In contrast to the prior art, which relies on manual effort by the user to create a multimedia document which links a paper representation to digital data ([1, 6, 11]), Video Paper is a fully automatic solution that creates a paper document which allows users to quickly access information in a long recording by reading the transcript or browsing key frames. Recent results confirm the intuition of the system's developers that Video Paper is an effective browsing and retrieval interface.

An automatic key frame selection algorithm, components of which were described here, should be useful for generating customized print outs with a user-specific profile. This should be true for certain types of content, such news broadcasts. It's an open question whether this will offer substantial value over and above that provided by a simple time-based sampling strategy.

References

1. Arai, T., Aust, D., and Hudson, S. E., "PaperLink: A technique for hyperlinking from real paper to electronic content," Conference on Human Factors in Computing Systems (CHI 1997), 327-334.
2. Boreczky, J., A. Girgensohn, G. Golovchinsky, and S. Uchihashi, "An Interactive Comic Book Presentation for Exploring Video," Conference on Human factors in Computing Systems (CHI 2000), pp. 185-192, 2000.
3. P. Chiu, A. Kapuskar, S. Reitmeier, and L. Wilcox, "Notelook: Taking Notes in Meetings with Digital Video and Ink," ACM Multimedia '99, Oct. 1999, Orlando, FL.
4. J. Graham, "The Reader´s Helper: A Personalized Document Reading Environment," Proc. of the CHI 99 conference on Human factors in computing systems: the CHI is the limit, May 1999, 481-488.
5. J. Graham and J. J. Hull, "Video Paper: A Paper-based interface for skimming and watching video," International Conference on Consumer Electronics (ICCE), Los Angeles, CA, June 2002, 214-215.
6. D. L. Hecht, "Printed Embedded Data Graphical User Interfaces," IEEE Computer, March 2001, 47-55.
7. S. Klemmer, J. Graham, G. Wolff, and J. Landay, "Books with voices: Paper transcripts as a physical interface to oral histories," ACM Conference on Human Factors in Computing Systems (CHI-2003), Fort Lauderdale, Florida, April 5-10, 2003.
8. H. Li, D. Doermann, and O. Kia. "Automatic Text Detection and Tracking in Digital Video," *IEEE Transactions on Image Processing - Special Issue on Image and Video Processing for Digital Libraries*, 9(1), 147-156, 2000 .
9. B. Shahraray and D. C. Gibbon, "Automated Authoring of Hypermedia Documents of Video Programs", *ACM Multimedia 95*, November 5-9, 1995, San Francisco, CA.
10. M.A. Smith and T. Kanade, "Video skimming and characterization through the combination of image and language understanding, " in Proc. IEEE International Workshop on Content Based Access of Image and Video Database, 1998, pp. 61--70.
11. L. Stifelman, B. Arons, and C. Schmandt, "The Audio Notebook: Paper and Pen Interaction with Structured Speech," ACM CHI Conference, Seattle, WA, March 31 – April 5, 2001, 182-189.
12. A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner, "Advances in automatic meeting record creation and access", Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 597-600, 2001.