COMPUTING A MULTIMEDIA REPRESENTATION FOR DOCUMENTS GIVEN TIME AND DISPLAY CONSTRAINTS

Berna Erol[(1)], Kathrin Berkner[(1)], Siddharth Joshi[(2)], and Jonathan J. Hull[(1)]

[(1)] Ricoh California Research Center
2882 Sand Hill Rd, S.115, Menlo Park, CA, USA
{berna_erol, berkner, hull}@rii.ricoh.com

[(2)] Department of Electrical Eng., Stanford University
Packard 243, 350 Serra Mall, Stanford CA, USA
sidj@stanford.edu

ABSTRACT

It is difficult to view multipage, high resolution documents on devices with small displays. As a solution, we introduce a *Multimedia Thumbnail* representation, which can be seen as a multimedia clip that provides an automated guided tour through a document. Multimedia Thumbnails are automatically generated by taking a document image as input and first performing visual and audible information analysis on the document to determine salient document elements. Next, the time and information attributes for each document element are computed by taking into account the display and application constraints. An optimization routine, given a time constraint, selects elements to be included in the Multimedia Thumbnail. Last, the selected elements are synthesized into animated images and audio to create the final multimedia representation.

1. INTRODUCTION

Devices with small displays, such as MFPs, PDAs, cellular phones, and digital cameras are increasingly being used to access documents, web pages, and images. Browsing and viewing of documents on such devices, however, is still very difficult. Currently this problem has limited solutions. For example, often web pages are re-designed for viewing on small displays. In digital cameras, the problem of browsing photos is usually solved by simply showing a low resolution version of photos and expecting the user to zoom into the picture for more details. Document viewers on PDAs employ a similar method, allowing user to zoom into the document and scroll to see the details. These solutions require interaction (zoom in, pan, etc.) with a device that has limited navigation capability, e.g., a cellular phone. Automatic re-flowing of text in documents and web pages is suggested by some researchers as a solution to fit them into small displays [1][2]. Moreover, automatic navigation of photos is presented in [3]. However, these solutions either do not support multipage document images or require changing the layout and appearance of the document.

We introduce a new document representation called *Multimedia Thumbnail* (MMNail) that is suitable for viewing documents on small displays. Input to the MMNail generation algorithm is a 2D document image and output is a multimedia clip that can be seen as a guided tour through a document. We animate the document pages, zoom into and pan over the most important visual elements, such as title and figures, automatically. This way, we utilize both spatial and time dimensions for presenting the documents. Moreover, the audio channel is used to communicate some of the textual information, so called *audible* information. While document contents are shown in the visual channel, the audio channel is used to speak important keywords, figure captions, etc. As a result, an MMNail utilizes both the visual and audio channel of the browsing device in order to present an overview of the document on a limited display and in a limited time-frame, while keeping the interaction required by the user to a minimum. An example of an MMNail of a two page document is shown in Figure 1. In this example, the MMNail representation shows the first page, then automatically zooms into the title, shows the second page, and then automatically zooms into the figure. The audio channel on the other hand, first communicates the important keywords from the document and then reads out the figure captions that are too small to read on the screen.

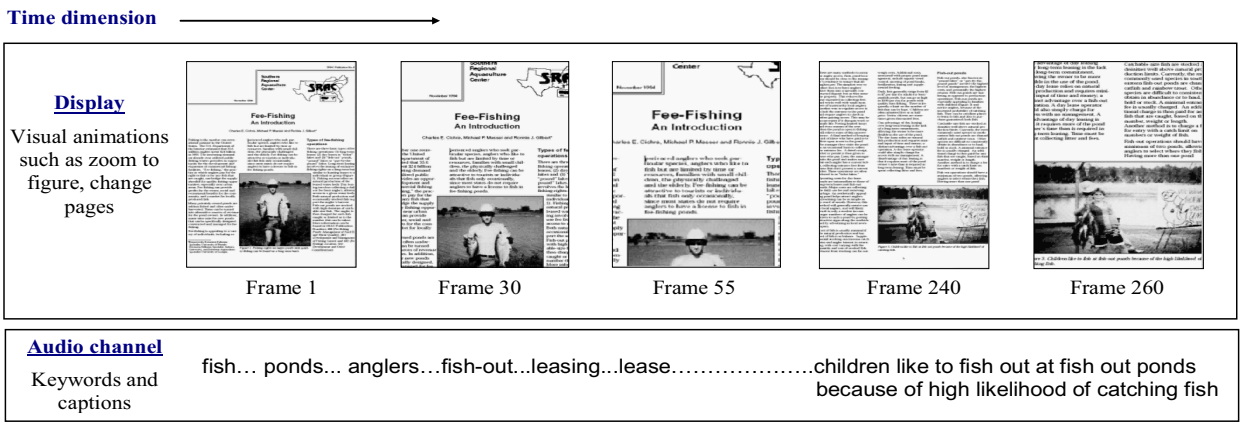

Figure 1. Multimedia Thumbnail Example.

Given that display devices have limited resolutions and typically limited playback time for a multimedia clip, it is often not possible to animate through an entire document or communicate all the audible information available via the audio channel. This leads to the following problem: Given the time and display constraints, which parts of the document should be included in the multimedia representation? Our paper addresses this problem with a three step algorithm for automatically generating MMNails, which consists of analysis of document image, optimization of document representation given a time constraint, and synthesis of Multimedia Thumbnails. In the next sections, we describe each of these steps in detail.

2. ALGORITHM OVERVIEW

Multimedia Thumbnails are created from electronic or scanned documents with a three step algorithm. In the analysis step, document contents are analyzed in order to identify important visual and audible document elements. Also, information and time attributes are computed for each of these elements. In the optimization step, a two stage knapsack-based algorithm is employed to determine the navigation path of MMNails, given a time constraint. In the last step, selected visual and audible information are synthesized into the audiovisual representation of documents.

3. ANALYSIS AND COMPUTATION OF ATTRIBUTES

A multipage document image and, optionally, a metadata file are input of the analysis step. Currently, the system accepts PDF and TIFF files as inputs. First, a preprocessing step is applied to the document, which includes OCR and layout analysis via commercial software. The output of the preprocessing, which is a collection of document elements, is further analyzed to assign logical labels to the document elements, such as title, sub-titles, author names, abstract, figures, and figure captions. Besides visual information, the analysis step also determines audible document information from the document image and metadata. Examples of audible information include figure captions, keywords, authors' names, publication name, etc., that can be converted to synthesized speech. We compute the keywords of a document with TF-IDF analysis [4].

Optimization problems related to documents generally involve some spatial constraints, such as optimizing layout and size for readability and reducing spacing [2][5]. In such frameworks, some *information attributes* are commonly associated with different parts of a document. In our framework, since we try to optimize not only the spatial presentation but also time presentation, we associate *time attributes* with each document element in addition to information attributes.

Document elements are divided into the following three groups: purely visual, E_v, purely audible, E_a, and synchronized audiovisual, E_{av}. Visual elements include document elements such as figures and graphs without any captions. Audible elements include elements that can be communicated easily in the audio channel without a visual representation. Examples of audible elements include keywords, and number of pages. Audiovisual elements are composed of elements that are presented on the audio and visual channel simultaneously. Examples include figures with captions.

3.1 Time attributes

Given a document element *e*, the time attribute, *t(e)* is the approximate duration that is sufficient for a user to comprehend a document element. For computing time attributes for figures without any captions, we make the assumption that complex figures take a longer time to comprehend. The complexity of a figure element *e* is measured by the figure entropy *H(e)*, which is computed using Multi-resolution Bit Distribution described in [6]. A time attribute for a figure element is computed as $t(e) = \alpha H(e)/H(P)$, where *H(e)* is the figure entropy, *H(P)* is the entropy of the entire page, and α is a time constant. Time required to comprehend a photo might be different than that of a graph, there for different α can be used for these different figure types. We do not distinguish different figure types in this paper and α is fixed to 4 seconds, which is the average time a user spends on a figure in our experiments.

Time attribute for a text document element (e.g., title, abstract) is determined to be the duration of the visual effects necessary to show the text segment to the user in a readable resolution. In previous experiments, we determined that text should be at least 7 points high in order to be readable [2]. If text is not readable when the whole document is fitted into the display area (i.e. thumbnail view), then a zoom operation is performed where the text is fitted to the display area. If even zooming in to the text is not sufficient for readability, then zooming into a part of the text is performed. Then a pan operation is carried out in order to show the user the remainder of the text. In order to compute time attributes for text elements, first the document image is downsampled to fit the display area. Then *Z(e)* is determined as the zoom factor that is necessary to bring the height of the smallest font in the text to the minimum readable height. Finally the time attribute for a visual element $e \in E_v$ is computed as follows:

$$t(e) = \begin{bmatrix} SCC \times n_e, & Z(e) = 1 \\ SCC \times n_e + Z_C, & Z(e) > 1 \end{bmatrix},$$

where n_e is number of characters in *e*, Z_C is zoom time (in our implementation this is fixed to be 1 second), and *SSC* (Speech Synthesis Constant) is the average time required to play back the synthesized audio character. *SSC* is computed as follows: (1) Synthesize a text document with the known number of characters, *K*, (2) measure the total time it takes for the synthesized speech to be spoken out, *T*, and compute *SSC=T/K*. *SSC* constant may change depending on the language choice, synthesizer that is used and the synthesizer options (female vs. male voice, accent type, talk speed, etc). With the AT&T speech SDK that we used to prototype Multimedia Thumbnails, *SSC* is computed to be equal to 75 ms when a female voice was used. Computation of *t(e)* remains the same even if a text element cannot be shown with one zoom operation and both zoom and pan operations are required. In such cases, the presentation time is utilized by first zooming into a portion of the text, for example the first *m* characters, and keeping the focus on the text for *SCC*×*m* seconds. Then the remainder of the time, i.e. *SCC*×*(n_e-m)*, is spent on the pan operation.

Time attributes for an audible document element, $e \in E_a$, is also computed in a similar fashion: $t(e)=SCC \times n_e$, where *SCC* is the speech synthesis constant and n_e is the number of characters in the document element.

Audiovisual elements are composed of an audio component, *A(e)*, and a visual component, *V(e)*. Time attribute for an audiovisual element is computed as the maximum of time attributes for its visual and audible components:

$$t(e) = \max(t(\mathrm{V(e)}), t(\mathrm{A(e)})).$$

For example, *t(e)* of a figure element is computed as the maximum of time required to comprehend the figure and the duration of synthesized figure caption.

3.2 Information attributes

An information attribute determines how much information a particular document part contains for the user. Obviously, this very much depends on the user's viewing/browsing style and the task on hand. For example, information in the abstract could be very important if the task is to understand the document, but it may not be as important if the task is merely to determine if the document has been seen before.

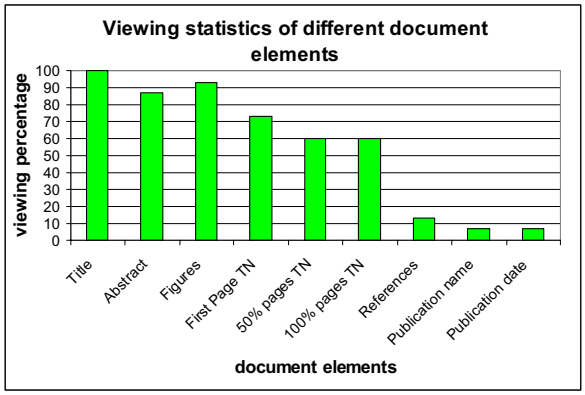

Figure 2. Percentage of users who viewed different parts of the documents.

In order to understand how important the information is in the different parts of a document, we performed an observational user study. Nine users participated in the study and three documents were shown to each of them (giving us 27 separate data points) for the task of understanding the contents of a document in a limited time. Users browsed the high resolution PDF documents on a small (PDA-size) display. The users' navigation behaviors were recorded and analyzed in order to understand which document parts were viewed during browsing. Figure 2 shows the percentage of users who viewed various document parts. This initial experiment gives us an idea about how much users value different document elements. For example, 100% of the users read the title, whereas very few users looked at the references, publication name and the date. We use these results to assign information attributes to text elements depending on the amount of being viewed. For example, the title has the information value of 1.0, where references are given the value 0.13.

4. THUMBNAIL OPTIMIZATION

Once the time and the information attributes are computed for the visual, audible, and audiovisual elements, the job of the optimizer is to produce the best thumbnail by selecting a combination of elements that can be displayed in a given time. The best thumbnail is one that maximizes the total information content of the thumbnail and can be displayed in the given time. The total information content of the thumbnail is the sum of the information content of the selected elements. Let the information content of an element e be denoted by $I(e)$, and the time required to display by $t(e)$. An element e belongs to either the set of visual elements E_v, or the set of audible elements E_a, or the set of audiovisual elements E_{av}. While displaying a visual element, an audible element can be played, and therefore overlap in time.

In this paper we give a strict priority to the visual elements in creating thumbnail. This means that we create a *partial* thumbnail by selecting elements from the set of visual elements and the set of audiovisual elements satisfying the display time constraint, such that the total information content in the partial thumbnail is maximized. The resulting optimization problem is

$$\text{maximize} \quad \sum_{e \in E_v \cup E_{av}} x(e)\ I(e)$$

$$\text{subject to} \quad \sum_{e \in E_v \cup E_{av}} x(e) t(e) \le T \qquad (1)$$

$$x(e) \in \{0,1\} \text{ for all } e \in E_v \cup E_{av}$$

where the $x(e)$ are the optimization variables, and T is the given display time. For an element e, $x(e)$=1 means it is selected to be in the thumbnail, and, $x(e)$=0 means it is not selected. Let $x^*(e)$ be the solution to the optimization problem. After the partial thumbnail is created the time for which the audio channel is free $\overline{T}$ is calculated by

$$\overline{T} = T - \sum_{e \in E_{av}} x^*(e)\ t(e).$$

The audible elements are chosen by solving another optimization problem similar to (1),

$$\text{maximize} \quad \sum_{e \in E_a} x(e)\ I(e)$$

$$\text{subject to} \quad \sum_{e \in E_a} x(e) t(e) \le \overline{T} \qquad (2)$$

$$x(e) \in \{0,1\} \text{ for all } e \in E_a$$

Thus by solving this two stage optimization problem we obtain the Multimedia Thumbnail.

The above optimization problems can be seen as a `0-1 knapsack' problem, which is a hard combinatorial optimization problem [7]. If we relax the constraints $x(e) \in \{0,1\}$ to $0 \le x(e) \le 1$, then the optimization problem becomes much easier to solve.

The solution to the optimization problem (1) -after approximation- is

- sort the elements $e \in E_v \cup E_{av}$ according to the ratio $I(e)/t(e)$ in descending order, i.e.,
$$\frac{I(e_1)}{t(e_1)} \ge \ldots \ge \frac{I(e_m)}{t(e_m)},$$
where m is the number of elements in $E_v \cup E_{av}$.
- find the integer k such that
$$\sum_{i=1}^{k} t(e_i) \le T \text{ and } \sum_{i=1}^{k+1} t(e_i) > T.$$
- select element e_i, i.e., $x(e_i) = 1$, if $i \le k$, otherwise not ($x(e_i) = 0$).

For practical purposes this approximation works well to the problem in (1), as we expect the individual elements to have much shorter display time than the total display time.

By dividing visual, audible, and audiovisual document elements into three separate sets, instead of just two for visual and audible, we can better model the optimization of synchronized visual and audible data. Also note that even though the optimization problem takes into account only time constraint,

the display and the application constraints indirectly affect the solution, as these constraints affect the information and time attributes of the elements.

5. SYNTHESIS

After the visual, audible, and audiovisual elements to be included are identified in the optimization step, visual and audiovisual elements are ordered in the reading order. Audible elements are added in the time intervals occupied only by visual elements. Visual information is rendered to create animations such as page flipping, pan, and zoom to certain locations on a page and the audible information is synthesized into audio clips.

6. IMPLEMENTATION USER FEEDBACK

The optimization routine outputs an *actions file* which contains all the information needed by the synthesis step, such as the names of the document images to be included in the MMNail, visual animations to be performed (type, coordinates, and duration), and audible document elements to be synthesized. Visual animations are implemented in Flash using ActionScript 2.0. Speech synthesis is implemented using the AT&T Natural Voices Text-to-Speech SDK. After obtaining visual and audio streams, synchronization is performed using Action Script to obtain a playable MMNail.

(a) (b)

Figure 3. Interface for (a) document browsing and (b) document viewing

A document browser interface that displays the thumbnail of the each document is shown in Figure 3.a. The interface is implemented in Flash 6.0, and is compatible with Windows and Macintosh operating systems and PDAs running the Pocket PC OS. When a user selects a document thumbnail in order to view the MMNail representation, automated navigation is activated in the interface given in Figure 3.b. The user has control over playback with the "control bar", which he can use to start, stop, go backward and forward in the MMNail timeline.

We performed a preliminary user study with nine participants in order to understand the usefulness of Multimedia Thumbnails. The users first browsed some documents on a PDA-size display with limited manual navigation. Then they were asked to watch the MMNail clips created for another set of documents. Later users were interviewed on the usefulness of MMNails, particularly the communication of information through both audio and visual channels. They were also asked to give a score between 1 (not useful) and 10 (very useful) to each communication channel. Users give an average score of 7.2 (σ=3.6) and 7.1 (σ=2.7) to the usefulness of visual and audio channels, respectively. They generally liked the fact that zooming in and page flipping operations are automatically performed. Nevertheless, they asked for more control over the playback speed of MMNails. In terms of visual animations, some users pointed out that they sometimes felt that they were lost in the document. They suggested that page flipping operations could be animated explicitly. Additional user comments pointed out that usefulness of the audio depends on the quality of the synthesized speech. Particularly for very short audio segments, such as keywords, understanding the audio content was considered to be difficult. Moreover, some text such as author's names, were incorrectly synthesized in most cases, which caused distraction to the users. They suggested that such text could be just displayed visually without the audio channel. On the other hand, the users found the use of audio channel for figure captions very useful (average score=8.9).

7. CONCLUSIONS AND FUTURE WORK

In this paper we introduced a new multimedia representation for documents call Multimedia Thumbnails and presented a method for automatically generating this representation. Our method includes the new idea of associating time attributes with different document parts and finding an optimum navigation path through the document given the display and time constraints.

The idea of representing static document images in a multimedia format using both audio and visual channels opens up many interesting research questions. For example, we employ entropy of figures as a measure of comprehension time for the user, but better complexity measures for figures can be developed which makes the distinction between photos, tables, and graphs, as they require different levels of attention from users. We employ constant information attributes for different parts of a document based on our user study. Nevertheless, more sophisticated methods can be used to assign information attributes. For example, figures can be assigned an information value based on their size, how many times they were referenced in the paper or the existence of some objects in the figure such as faces and buildings. Also, more user studies are needed to better understand the user's document browsing behaviors for different tasks (e.g., browsing, search, overview) and how Multimedia Thumbnails can be improved to be more useful for their browsing needs.

REFERENCES

[1] T. M. Breuel, W. C. Janssen, K. Popat, H. S. Baird, "Paper to PDA", Proceedings of the International Conference on Pattern Recognition, 2002.

[2] K. Berkner, E. L. Schwartz, C. Marle, "SmartNails - Image and Display Dependent Thumbnails," Proceedings of SPIE, vol. 5296, pp. 53-65, San Jose, 2004.

[3] M-Y. Wang, X. Xie, W-Y. Ma, H-J. Zhang, "MobiPicture - Browsing Pictures on Mobile Devices," International Conference of ACM Multimedia, Berkeley, Nov. 2003.

[4] G. Salton, Automatic Text Processing, Addison-Wesley, 1989.

[5] V. Eglin and S. Bres, "Document page similarity based on layout visual saliency: Application to query by example and document classification", Proceedings of ICDAR, pp. 1208-1212, 2003.

[6] R. Neelamani, K. Berkner, "Adaptive Representation of JPEG 2000 Images using Header-based Processing", Proceedings of ICIP, pp. 381-384, 2002.

[7] R.L. Rivest, H.H. Cormen, C.E. Leiserson, Introduction to Algorithms, MIT Pres, MC-Graw-Hill, Cambridge Massachusetts, 1997.