

# Multimedia Thumbnails for Documents

Berna Erol

RICOH California Research Center  
2882 Sand Hill Road,  
Menlo Park, CA, USA  
berna\_erol@rii.ricoh.com

Kathrin Berkner

RICOH California Research Center  
2882 Sand Hill Road,  
Menlo Park, CA, USA  
berkner@rii.ricoh.com

Siddharth Joshi

Stanford University  
Department of Electrical Eng.  
Packard 243, 350 Serra Mall,  
Stanford, CA, USA  
sidj@stanford.edu

## ABSTRACT

As small portable devices are becoming standard personal equipments, there is a great need for the adaptation of information content to small displays. Currently, no good solutions exist for viewing formatted documents, such as pdf documents, on these devices. Adapting content of web pages to small displays is usually achieved by complete redesign of a page or automatically reflowing text for small displays. Such techniques may not be applicable to documents whose format needs to be preserved. To address this problem, we propose a new document representation called Multimedia Thumbnail. Multimedia Thumbnail uses the visual and audio channels of small portable devices to communicate document information in form of a multimedia clip, which can be seen as a movie trailer for a document. Generation of such a clip includes a document analysis step, where salient document information is extracted, an optimization step, where the document information to be included in the thumbnail is determined based on display and time constraints, and a synthesis step, where visual and audible information are formed into a playable Multimedia Thumbnail. We also present user study results that evaluate an initial system design and point to further modification on analysis, optimization, and user interface components.

## Categories and Subject Descriptors

H.5.2 [Inf. Interfaces and Presentation]: *User Interfaces*

I.7 [Document and Text Processing]: *General*.

## General Terms

Algorithms, Human Factors

## Keywords

Document representation, multimodal processing, mobile, user interface, automated browsing, summarization.

## 1. INTRODUCTION

With the increased ubiquity of wireless networks and personal mobile devices, more people are required to browse and view web

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'06, October 23–27, 2006, Santa Barbara, California, USA.

pages, photos, and even documents on mobile devices. These devices usually have small displays and limited navigation capabilities which make viewing of high resolution images/documents very difficult. One solution for web page viewing using small displays is the common way of designing simpler and low-graphics versions of web pages. Another solution involves user interaction for selection of page elements to be zoomed in or collapsed [1]. For a different information source, namely digital photos, the browsing and viewing problem is partially solved by simply showing a low resolution version of photos, giving the user ability to zoom in and scroll [2], or automatically showing the salient content of photos [3].

Unlike web pages, documents are strictly formatted with page-breaks, line-breaks, column layouts, margins, etc. Layout in formatted documents communicates semantic information about the document. As shown in [2], changing layout for reading of documents on small devices may change the level of understanding of the document content. Another difference between web pages and documents is that web pages usually contain links to support navigation. In documents, navigation is typically supported by layout, not by links. Therefore, in many cases it is desired that the layout of documents is preserved.



**Figure 1. Browsing of high resolution, multi-page documents on mobile devices is difficult.**

Compared with photos, document images may have much higher resolution and form multi-page image collections. Therefore, much more zooming and scrolling at the user's side is required in order to observe the content of a document. Moreover, documents have highly distributed information. Focus points on a photo can be only a few people's faces or an object, whereas a typical document may contain many focus points, such as title, authors, abstract, figures, and references. Another important difference between photographic pictures and documents is semantic content of documents. Whereas photos include in- and out-of-focus objects, documents contain a variety of logical units and

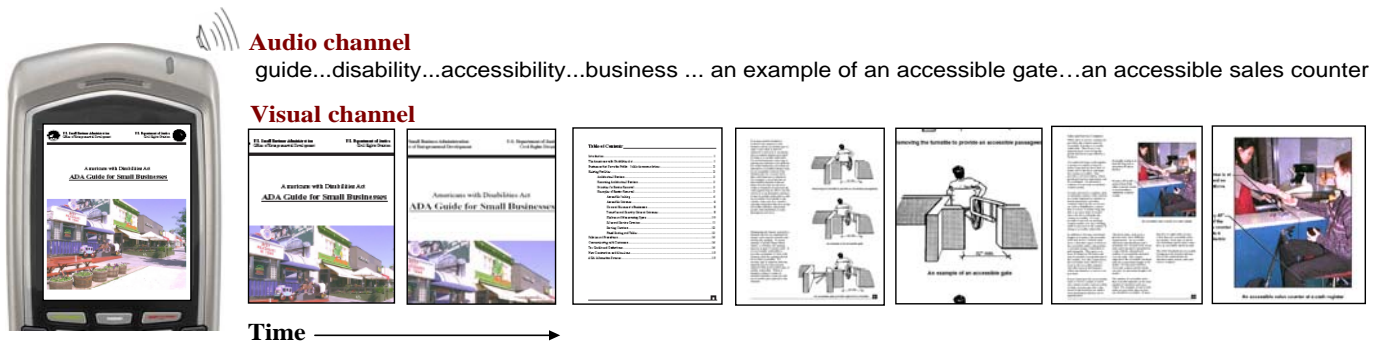


Figure 2. Multimedia Thumbnail use both visual and audio channels to communicate document information.

relationships between those units. Most dominantly, there is text which is intended for reading. Text can vary in font characteristics, a reading order may exist, and image and text units may be linked, e.g. a figure picture and a figure caption. A document viewer needs to consider these characteristics when assisting a reading task to make it as natural as possible to the reader.

When viewing documents on portable devices, such as the one shown in Figure 1, convenient user interaction with the device may not often be possible. For example user may have only one hand available for viewing and navigation. Moreover, even when interaction with the device is possible, the amount of zooming in, switching between document pages, and scrolling required to navigate a high resolution document is not easily achievable on devices such as phones and PDAs with limited navigational capabilities. A better solution for document browsing and viewing is needed.

In this paper, we introduce a new document representation called *Multimedia Thumbnail* (MMNail). MMNails can be seen as movie trailers or automated guided tours of documents. An MMNail representation automatically animates the document pages by zooming into and panning over the most important document elements, such as title and figures. While document contents are shown in the visual channel, the audio channel is used to communicate some of the textual document information, so called *audible* information. Audible information includes synthesizable document information such as keywords and figure captions. In MMNail representations, dense spatial information of documents is distributed among limited viewing area, time dimension, and the audio channel.

An example of an MMNail is shown in Figure 2. In this example, the MMNail representation shows the first page, then automatically zooms into the title, shows the consecutive pages while automatically zooming into the figures. 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 be read on the screen.

MMNail representation is automatically generated by analyzing the contents of a document, computing important audible and visual document information, optimizing the selection of document elements based on time and display constraints, and finally synthesizing the document elements into the MMNail representation. More details of automatic MMNail generation are presented in Section 3.

After our initial implementation of MMNails, we performed an observational user study to evaluate our system. The results of our user study are given in Section 4. Outcome of the study led to adaptation of a new analysis and optimization framework. These ideas are illustrated in Section 5. In the next section, we give an overview of the prior research.

## 2. REVIEW OF RELEVANT RESEARCH

Automatic re-flowing of text in documents and web pages is suggested by some researchers as a solution to fit text in small displays [4][5]. However, these solutions either do not support multi-page document images, or require changing the layout and appearance of the document. Web page and document summarization research in general focuses heavily on text [6][7] and usually does not utilize information channels (e.g. audio) that are not used in the original document representation.

SmartNail technology [5] creates an alternative image representation for a single document page by scaling, cropping, and reflowing page elements, given display size constraints. Even though a SmartNail representation handles images as well as text, the output is a static visual representation and it does not incorporate information from multiple pages. In Multimedia Thumbnails, the output consists of document information from multiple pages represented in a dynamic way using animation and audio.

The prior research on conversion of documents to audio mostly focuses on aiding visually impaired people. Adobe provides a plug-in to Acrobat reader that synthesizes PDF documents to speech [8]. Also in [9], guidelines are given on how to create an audiocassette from a document for blind or visually impaired people. As a general rule, it is mentioned to include any information that is included in tables and picture captions. Graphics in general should be omitted. Moreover, some work has been done on developing Web browsers for blind and visually impaired users. The focus in [10] is to map a graphical HTML document into a 3D virtual sound space environment, where non-speech auditory cues differentiate HTML documents. In all the applications for blind or visually impaired users, the goal is to transform as much information as possible into the audio channel and giving up on the visually channel completely. In contrast, MMNails optimize the communication of document information in both channels. Our framework also allows setting preference for one of the two channels.

One of the most relevant prior art to our work is described in [3] and [11], where authors propose a method for non-interactive picture browsing on mobile devices. Their method is to find salient, face and text regions on a picture automatically and then use zoom and pan motions on this picture to automatically provide close ups to the viewer. Their method concentrates on representing photos, where our method focuses on representing high-resolution multi-page document content. Moreover, the method in [3] is image based only, where we employ visual and audio channel for document thumbnails.

### 3. AUTOMATIC MULTIMEDIA THUMBNAIL GENERATION

Multimedia Thumbnails are created from electronic or scanned documents with a three step algorithm shown in Figure 3. In the analysis step, document content is analyzed in order to identify important visual and audibly presentable document elements. Also, information and time attributes are computed for each of these elements. In the optimization step, document elements to be included in the representation are selected based on a time constraint. In the last step, selected visual and audible information is synthesized into the audiovisual representation of a document.

#### 3.1 Analysis

A multi-page 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 layout analysis and optical character recognition via commercial software. The software also automatically determines a reading order based on the layout. The output of the preprocessor, which is a collection of document elements, is further analyzed in the visual focus points determiner to assign logical labels to visual document elements, such as page thumbnails, title, and figures. Note that these visual focus points may be application dependent. Besides visual information, the analysis step also determines audible document information from the document image and metadata. Examples of audible information include figure captions and keywords that can be converted to synthesized speech. We compute the keywords of a document with TF-IDF analysis [12].

Optimization problems related to documents generally involve some spatial constraints, such as optimizing layout and size for readability and reducing spacing [5][13]. 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.

Let  $E$  denote the set of all document elements. This set is composed of the set of visual elements  $E_v$  and the set of audible elements  $E_a$ . Some visual and audible document elements, such as figures and captions, should be presented synchronously in the MMNail representation, where some other visual and audible document elements, such as thumbnails of pages and keywords, can be presented asynchronously.

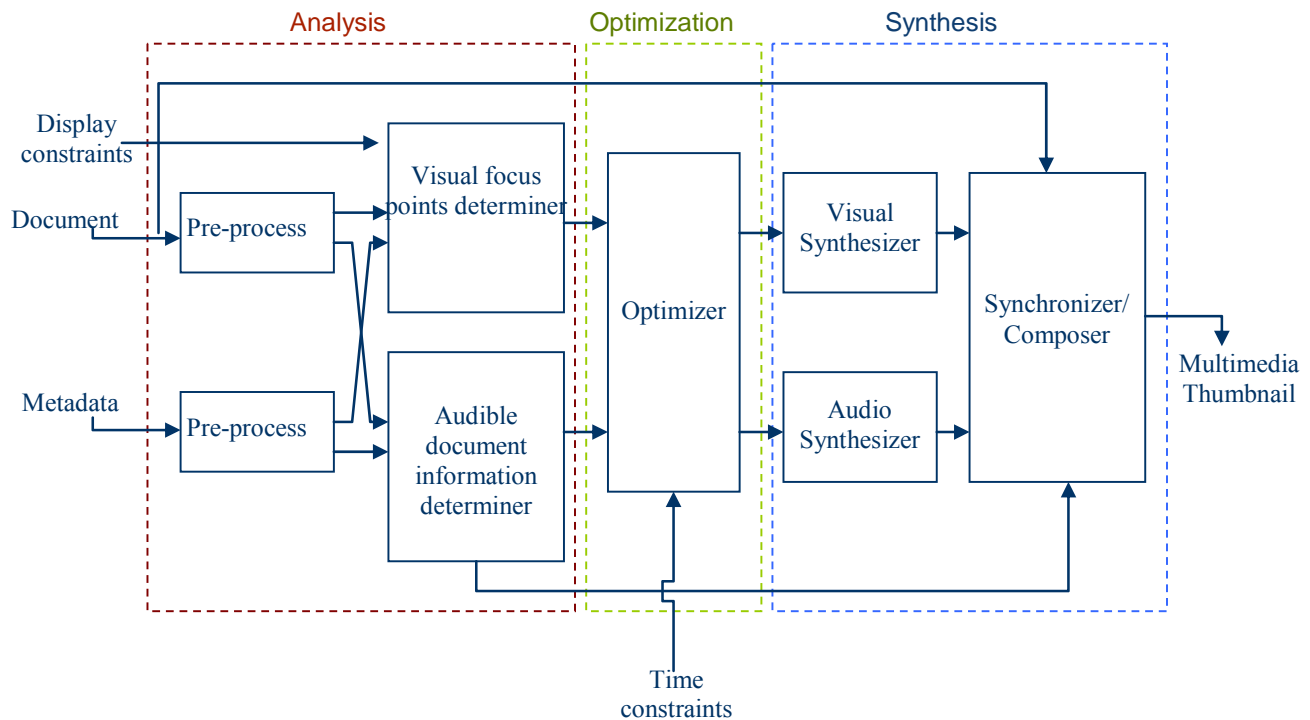


Figure 3. Overview of the Multimedia Thumbnail generation algorithm.

Given a visual document element  $e \in E_v$ , its time attribute,  $t(e)$ , is the approximate duration that is sufficient for a user to comprehend the document element. The time attribute for a text document element (e.g., title) is determined to be the duration of the visual effects necessary to show the text segment to the user at a readable resolution. In previous experiments, text was determined to be at least 7 points high in order to be readable on a CRT monitor [5]. Although our target display device is small portable devices, not CRT monitors, we use 7 points as a good approximation of the minimum text height in our implementation. If text is not readable when the whole document is fitted into the display area (i.e. thumbnail view), a zoom operation is performed. If even zooming into the text is not sufficient for readability, then zooming into a part of the text is performed. 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 a zoom factor  $Z(e)$  is determined as the factor that is necessary to scale 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:

$$t(e) = \begin{cases} SSC \times n_e, & Z(e) = 1 \\ SSC \times n_e + Z_c, & Z(e) > 1 \end{cases},$$

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 (3) compute  $SSC = T/K$ . The  $SSC$  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 an 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  $m_e$  out of a total of  $n_e$  characters, and keeping the focus on the text for  $SSC \times m_e$  seconds. Then the remainder of the time, i.e.  $SSC \times (n_e - m_e)$ , is spent on the pan operation.

The time attribute for an audible text document element  $e \in E_a$  is computed in a similar fashion:  $t(e) = SSC \times n_e$ , where  $SSC$  is the speech synthesis constant and  $n_e$  is the number of characters in the document element.

The information attribute of an element  $e \in E$  is denoted by  $I(e)$ . For a keyword element  $e \in E_a$ ,  $I(e)$  is the TF-IDF importance score that is normalized into [0-1] range. Information attributes of all other document elements are set to a constant value  $D$ .

### 3.2 Optimization

The optimizer receives the time constraint  $L$ , i.e. the duration of the Multimedia Thumbnail, and the output from the analyzer,

which includes the characterization of the visual and audible document information. Then a combination of visual and audible information is determined that meets the time constraint.

An overview of the optimization algorithm is shown in Figure 4. Even though the only constraint accepted as input is the time constraint, the display constraint is indirectly considered when computing the time attributes for text elements.

The main functionality of the optimizer is to first determine how many pages,  $n_j$ , can be shown to the user, given each page is shown as a still image on the display for  $t_1$  (currently set to 0.5) seconds. Given a number of pages in a document  $n_j$ , if there is time left after showing all the pages, i.e.  $L - t_1 \times n_j \geq 0$ , optimizer allocates time  $t(e_T)$  for zooming in and displaying the title. If there is any time left after page showing and title zooming, the optimizer sorts the figure caption elements  $e_c \in E_a$  based on their time attributes  $t(e_c)$ . The optimizer then applies a linear packing/filling order approach to the sorted time attributes to select which figures will be included in the MMNail. Still-image holding is applied to the selected figures of the document. During the occupation of the visual channel by image holding, the caption is recited in the audio channel.

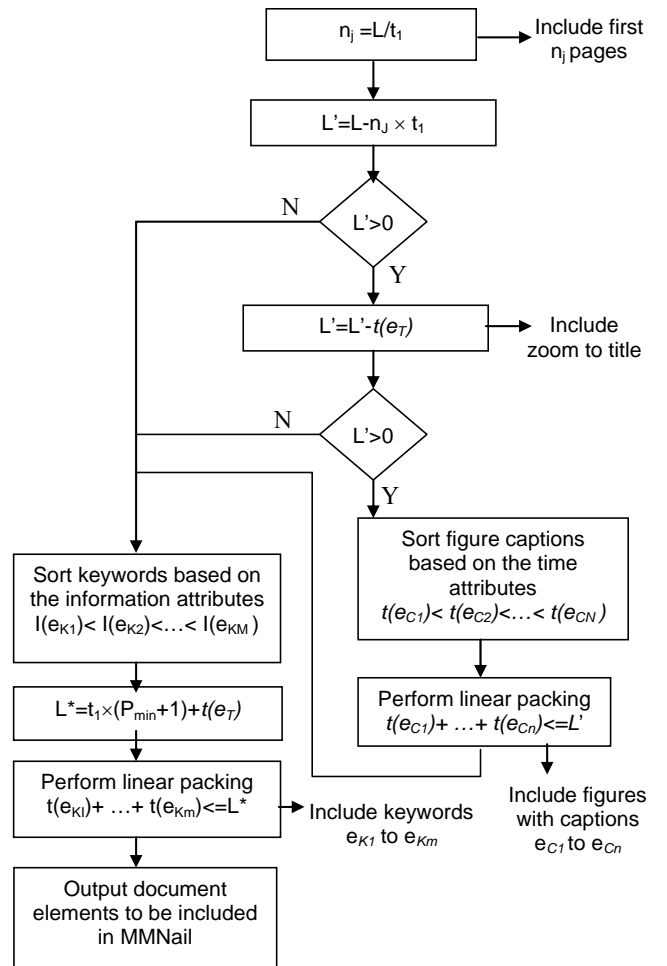


Figure 4. Overview of the optimization method.

The audio channel is occupied only when the figure captions are spoken. In order to utilize the audio channel better, the time available until the first figure caption occupies the audio channel is used to speak out keywords. This time window  $L^*$  is computed as

$$L^* = t_l \times (P_{min} + 1) + t(e_T),$$

where  $t_l$  is the time required to display one page,  $P_{min}$  is number of pages selected for the MMNail starting from the beginning of the documents that have no figure included in the MMNail, and  $t(e_T)$  is the time attribute of the title element  $e_T$ .

Keywords are found by TF-IDF analysis, and time and information attributes are assigned as discussed in the previous section. Using the time constraint  $L^*$ , a limited set of keywords is selected to be included in the MMNail as follows: First, keyword elements  $e_K \in E_a$  are sorted with respect to the information attributes  $I(e_K)$ . Then the first  $m$  keywords are selected such that  $t(e_{K1}) + \dots + t(e_{Km}) \leq L^*$ ,  $t(e_{Ki})$  is the time attribute of the element  $e_{Ki}$ .

The optimization scheme described here was used for initial prototyping of Multimedia Thumbnails. This scheme allows inclusion of limited number of document elements and provides limited flexibility. A more generalized optimization scheme is presented in Section 5, which has been developed incorporating findings and feedback obtained from user studies.

### 3.3 Synthesis

The synthesizer creates the final Multimedia Thumbnail by first ordering the document elements that are selected by the optimizer with respect to the reading order. Then the processing steps determined by the optimizer are executed. Multimedia processing steps include speech synthesis, the “traditional” image processing steps, such as cropping and scaling, and animation steps such as page flipping, panning, and zooming.

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. Audible information is converted to audio 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 in the Flash format.

## 4. OBSERVATIONAL USER STUDY

One of the challenges of automatically creating MMNails is to identify which parts of the document to be included in an MMNail representation. In the initial development of Multimedia Thumbnails, we made some assumptions about how people browse documents, which parts of the document they would like to view given a limited time window, and designed an ad-hoc optimizer for generation of MMNails.

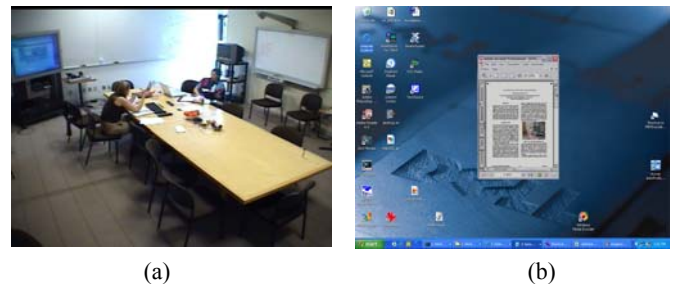
Nevertheless, for MMNail representation to be truly useful, it should be inline with the user’s viewing needs. What parts of the document users like to view if they have only 30 second overview of a document? Title? Abstract? Figures? Thumbnails?

References? How does this preference depend on the task on hand?

It is possible to find user studies in the literature that address viewing images and documents on small form factor devices [2][14]. The most relevant user study described in [2] focuses on how students review lecture material and annotate on devices with small displays. We were not able to use the results of this study directly since it does not give an evaluation of how much time people spend on different parts of a document or their specific browsing behavior. Moreover, we needed to answer questions such as “What information do users prefer to ‘view’ versus ‘hear?’” and “Would hearing some document information be annoying or useful?”. Our user study tries to shed a light on some of these questions. In the study, we also showed the users the initial prototype of Multimedia Thumbnails and asked for their evaluations.

### 4.1 Experimental Setup and Methodology

In total, 9 users participated to the study and 6 documents were given to each of them to perform 2 different tasks. The users browsed pdf documents on a small (PDA-size) viewing area as shown in Figure 5(b). Users were allowed to perform limited navigation, which included going left-right-up-down and changing the zoom factor. The initial view of the pdf document was a random location on a random page with a random zoom parameter. This way, if a user wanted to see the thumbnail (an overview of the page where the text is not easily readable), she had to do so explicitly. Users’ navigation behaviors were recorded and analyzed in order to understand which document parts they viewed during browsing.



**Figure 5. Experimental setup. (a) Interviews and user’s browsing behaviors were recorded. (b) Users browsed PDF files in a small viewing area with limited manual navigation.**

The following two tasks were given to the users:

1. Document search/browsing task

Identify whether users recognize a document as being in the stack of 15+ documents that were shown to them one week in advance. Each user is given 4 documents and up to 2 minutes each for browsing.

2. Content understanding task

Understand the content and visual appearance of a document in order to answer 10 multiple choice questions that will be given later. Each user is given 2 documents and 3 mins each for browsing. At the end no questions were asked to the users. This was because our purpose was

not to measure their understanding of the document, but to record their behavior when they are paying attention to the document content.

General-interest documents were shown to non-technical users while technical people were shown technical documents.

Afterwards, four questionnaires were given to the users:

1. For the searching/browsing task, rate the usefulness of viewing each part of the document
2. For the understanding task, rate the usefulness of viewing each part of the document
3. After seeing MMNails, rate the usefulness of animation and audio parts
4. For both tasks, rate the usefulness of hearing information from various parts of the document

An example questionnaire is given in Figure 6. Questionnaire results are analyzed by computing mean and standard deviation of scores for each document element. “Very useful”, “useful”, and “not useful” choices are assigned 10, 5, and 0 points, respectively. “I don’t know” answers were ignored in the computation of mean and variance.

	Very useful	Somewhat useful	Not useful	I don't know
Thumbnail of first page	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Title	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Thumbnails of 50% of pages	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Thumbnails of all pages	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Abstract	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Author's names	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Figures	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Figure captions	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
References	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Publication name	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Publication date	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figure 6. An example questionnaire

Besides questionnaires, recorded information was reviewed for whether or not users look at certain parts of the document and for how long. Moreover, users’ comments were collected from the audio recording.

## 4.2 Document Browsing Behavior Analysis

Below we give the results for analyzing user clicks through the document when they were performing the search/browse and understanding tasks, respectively. During the search/browse task, majority of the users viewed the title, figures, and page thumbnails as shown in Figure 7. When performing the content

understanding task, majority of the users viewed the title, abstract, figures, and page thumbnails, also as shown in Figure 7. Performing the search/browse task, users, who viewed the corresponding document element, spent on average 4 seconds on titles, 3 seconds on figures, and 17 seconds on abstracts. During the document understanding task, they spent on average 4 seconds on titles, 4 seconds on figures, and 21 seconds on abstracts.

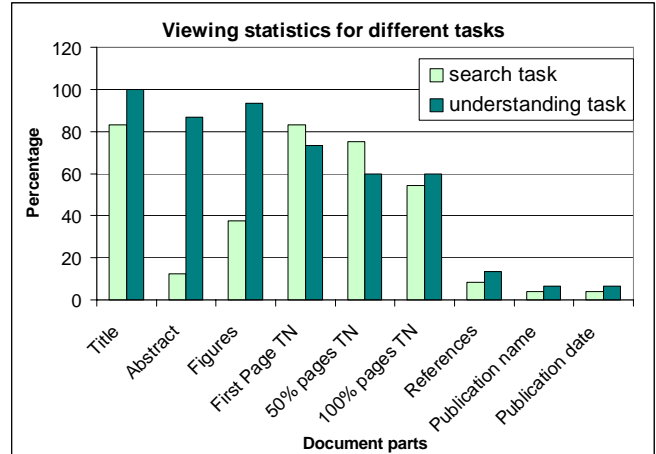


Figure 7. Percentage of users who viewed certain parts of the documents during the document search and document understanding tasks.

When we compare the percentage of users who viewed various document parts for the two different tasks, we can see that viewing “abstracts” and “figures” of documents is very much task dependent. As expected, very few users read the abstract of a document for the search task, but the majority of the users read the abstract for the understanding task. Another interesting observation was that, where almost 80% of the people viewed the thumbnail versions of 50% of the pages for the search/browse task, 60% of the people viewed the thumbnail version of 50% of the pages for the document understanding task. This should be expected since the page thumbnails are likely to play a significant role when recalling an already viewed document, where they play less of a role when the task is understanding of the document content.

## 4.3 Questionnaire on Document Browsing

The participants were also asked in a questionnaire to assign importance scores (0 for “not useful”, 5 for “somewhat useful”, and 10 for “very useful”) for viewing various document parts for the search and understanding tasks. The results are presented in Table 1. Users generally agreed that the title, figures, and abstract were the most important document parts for both tasks. An interesting result was that the figure captions were important for the understanding task (average score: 7.5), but not very important for the search task (average score: 3.1).

We also note that there are some inconsistencies between the user behavior analysis and the questionnaire results. For example, questionnaire show that the abstract is important for the search task, but only a few users actually stopped to read the abstract

during this task. Moreover, for the search/browse task, user ranked figures as important (9.4), but less than 40% of users viewed figures. Also, for the same task, users ranked page thumbnails middle importance (6.9) but more than 80% of users viewed page thumbnails.

**Table 1. Questionnaire results regarding to the importance of different parts of the document for the search and understanding tasks.**

Doc Part	Search Task		Understanding Task	
	Average Score	$\sigma$	Average Score	$\sigma$
Title	9.4	1.8	10	0
Figures	9.4	1.8	8.8	2.3
Abstract	7.5	3.8	10	0
Figure captions	3.1	4.6	7.5	3.1
Thumbnail of first page	6.9	3.7	5.6	4.6
Thumbnail of 50% pages	6.3	3.5	5	2
Publication name	5.7	4.5	6.1	4.1
Publication date	5	4.6	6.1	3.3
Author names	5	4.6	6.7	4.3
Thumbnail 100% pages	3.1	2.3	4.3	2.4
References	1.8	3.7	5	3.3

When reviewing the questionnaire results, the users had the following additional remarks:

- Author info is more useful if the user recognizes the person.
- Publication date is important when the user is trying to decide if she wants to read it.
- Publication name is important when the user is screening with this criteria.
- Layout of a page (usually as seen in a page thumbnail) can give the user the genre of the document.
- Showing section titles, number of pages, keywords, table of contents, conclusions, topic of each paragraph would be useful.

#### 4.4 Questionnaire on the Usefulness of Multimedia Thumbnail Representation

The participants were interviewed regarding the usefulness of the visual animation and the audio in MMNail examples. The results indicate that both elements received a score of 7 out of 10 in terms of usefulness, as shown in Table 2. When interviewed, users pointed out that animation combined with audio information, particularly combination of seeing the figures and hearing figure captions, were very useful. However, they also had some comments on the downside of this representation, such that

the synthesized speech was distracting at times, sometimes keywords were hard to understand, and that there is good information in audio but they may not like their device talking to them, especially in public places.

**Table 2. Questionnaire results for MMNails**

	Ave	$\sigma$
Animation	7.2	3.6
Audio	7.1	2.7

Some suggestions for improvement included control over skipping parts of audio and the speed of animation, and showing all the figures with RSVP. They also suggested that, in order to not to loose where you are in the document during the animation, some effects such as page flipping may be useful to include in the MMNails.

#### 4.5 Questionnaire on the Usefulness of Audible Information

In order to get some insight on what kind of information would be best suited to present in the audio channel, a questionnaire was given to the users on the importance of hearing different types of document information. User scores are presented in Table 3. Users 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.

**Table 3. Questionnaire results regarding to the importance of different audible parts of the document.**

Doc Part	Average Score	$\sigma$
Figure captions	8.9	2.2
Title	6.1	4.9
Keywords	5	4.3
Authors	4.4	3.9
Publication name	4.4	4.6
Number of pages	4.4	3.2
Publication date	3.9	4.1

Some other interesting feedback from the users includes the following:

- Number of pages can be particularly useful for printing.
- Number of pages can be converted to audio by using a low pitch sound for large number of pages and a high pitch sound for small number of pages.
- Publication name and date are usually printed small, so it may be good to give this information via audio channel.
- Some other information for the audio channel may include section headings, most important sentences (i.e. highlights) of the document, page number, and abstract.

- Pronunciation of authors' names with synthesized speech may be difficult.
- If audio is very short, there is little opportunity to get used to the voice.

## 5. GENERALIZED OPTIMIZATION FRAMEWORK

From our user study, the overall impression was that people have very different preferences about viewing documents. Also, the browsing style highly depends on the task on hand. Therefore, it is important to develop an MMNail generator which is easily expandable to include different document elements and configurable for personal preferences and target application.

In this section, we present a generalized optimization framework for generation of Multimedia Thumbnails which is developed in the light of the user studies described in the previous section. In the new framework, we redesigned some of the components of the system presented in Figure 3.

### 5.1 Computation of Attributes

In the analyzer, document elements are now divided into the following three mutually exclusive 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. 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.

Clearly, an audiovisual element can also be represented with purely visual and audible components. The decision regarding which document elements should be represented in which channel is user, target application, and task dependent.

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 visual figure element  $e \in E_v$  is measured by the figure entropy  $H(e)$ , which is computed using Multi-resolution Bit Distribution described in [15]. Time attribute for a figure element is computed as  $t(e) = \alpha H(e) / \bar{H}$ , where  $H(e)$  is the figure entropy,  $\bar{H}$  is the mean entropy, and  $\alpha$  is a time constant.  $\bar{H}$  is empirically determined by measuring the average entropy for a large collection of document figures. Time required to comprehend a photo might be different than that of a graph, therefore different  $\alpha$  can be used for these different figure types. We do not distinguish different figure types in this paper and  $\alpha$  is fixed to 4 seconds, which is the average time a user spends on a figure in our experiments.

An audiovisual element  $e$  is composed of an audio component,  $A(e)$ , and a visual component,  $V(e)$ . The time attribute for an

audiovisual element is computed as the maximum of time attributes for its visual and audible components:

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

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

An information attribute determines how much information a particular document part contains for the user. Obviously, this depends very much 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 7 shows the percentage of users who viewed various document parts for different tasks. This initial experiment gives us an idea about how much users value different document elements. For example, 100% of the users read the title in the document understanding task, whereas very few users looked at the references, publication name and date. We use these results to assign information attributes to text elements depending on the amount of being viewed. For example, if the task is document understanding, the title has the information value of 1.0, where references are given the value 0.13.

### 5.2 Optimization

The optimizer maximizes the total information content of the Multimedia Thumbnail given a time constraint using the time and information attributes of document elements. The total information content of the thumbnail is the sum of the information content of the selected elements. An element  $e$  belongs to either the set of visual elements,  $E_v$ , 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.

We give a strict priority to the visual elements in creating a 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

$$\begin{aligned} & \text{maximize} && \sum_{e \in E_v \cup E_{av}} x(e) I(e) \\ & \text{subject to} && \sum_{e \in E_v \cup E_{av}} x(e) t(e) \leq T \\ & && x(e) \in \{0,1\} \text{ for all } e \in E_v \cup E_{av} \end{aligned} \quad (1)$$

where  $I(e)$  is the information content of an element  $e$ ,  $t(e)$  is the time required to present  $e$ ,  $x(e)$  is the optimization variable, and  $T$  is the given time constraint. 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  $\bar{T}$  is calculated by

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

Using the new time constraint  $\bar{T}$ , the audible elements are chosen by solving another optimization problem similar to (1),

$$\begin{aligned} & \text{maximize} && \sum_{e \in E_a} x(e) I(e) \\ & \text{subject to} && \sum_{e \in E_a} x(e) t(e) \leq \bar{T} \\ & && x(e) \in \{0,1\} \text{ for all } e \in E_a \end{aligned} \quad (2)$$

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

The above optimization problems can be seen as a '0-1 knapsack' problem, which is a hard combinatorial optimization problem [16]. If we relax the constraints  $x(e) \in \{0,1\}$  to  $0 \leq x(e) \leq 1$ , then the resulting optimization problem becomes a linear program and it can be solved easily. The solution can be obtained by the following algorithm. First, sort the elements  $e \in E_v \cup E_{av}$  according to the ratio  $I(e)/t(e)$  in descending order, i.e.,  $\frac{I(e_1)}{t(e_1)} \geq \dots \geq \frac{I(e_m)}{t(e_m)}$ , where  $m$  is the number of elements in  $E_v \cup E_{av}$ . Then the following procedure can be used to select the elements to be included in the thumbnail:

$$\begin{aligned} & x(e_i) = 0 \text{ for all } i = 1, \dots, m \\ & T_{\text{remaining}} = T \\ & i = 1 \\ & \text{while } (T_{\text{remaining}} > 0 \text{ and } i \leq m) \{ \\ & \quad \text{if } (t(e_i) \leq T_{\text{remaining}}) \{ \\ & \quad \quad x(e_i) = 1 \\ & \quad \quad T_{\text{remaining}} = T_{\text{remaining}} - t(e_i) \\ & \quad \} \\ & \quad i = i + 1 \\ & \} \end{aligned}$$

Note that even though the optimization problem takes into account only a time constraint, the display and the application constraints indirectly affect the solution since these are employed to compute the information and time attributes of the elements.

### 5.2.1 Page Thumbnail Constraint

Page thumbnails (low resolution view of an entire page) are visual document elements. In order to give the user a better context and support easier navigation through the document, if any visual element in a page is displayed, that page's thumbnail view should also be displayed in the Multimedia Thumbnail. It is possible that the information content to display time ratio of a page thumbnail is smaller than those of visual elements on a particular page. Thus, when the optimization problem (1) is solved, the solution may include a visual element on a page but not that page's thumbnail. To solve this problem, (1) is modified as follows. Given a page's thumbnails element,  $p_0$ , and the visual (or audiovisual) elements on that page,  $p_1, \dots, p_k$ , it is desired that if any of the elements  $p_1, \dots, p_k$  is included to be in the multimedia thumbnail, then  $p_0$  should also be included in the thumbnail, i.e.,

$$x(p_0) = 1 \text{ if any } x(p_j) = 1, \quad j = 1, \dots, k.$$

In the relaxed version of the optimization problem (1), this can be done by adding the following inequality constraints:

$$x(p_0) \geq x(p_j), \quad j = 1, \dots, k.$$

The new optimization problem is also a linear program; but the simple algorithm presented earlier to solve problem (1) is not applicable in this case. Nevertheless any linear programming solver, for example [17], can be used to find a solution.

Note that adding the inequalities gives strict priority to a certain element irrespective of the information content and time to display attributes. Such priorities can be added not only for the page thumbnails but also for other objects in the documents.

## 6. USER INTERFACE



Figure 8. Interface for (a) document browsing and (b) document viewing

A document browser interface that displays the thumbnail of each document is shown in Figure 8(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 8(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.

## 7. SUMMARY AND OUTLOOK

In this paper, we address the problem of representing multi-page and high resolution documents on small form factor devices. Contrasting fix-formatted document representations with web pages and photographic images lead to a very specific characterization of the problem. User needs assistance in navigation, but navigation through a document is given implicitly through strict formatting, not via links. Furthermore, skimming over and focusing on significant text passages are essential patterns in reading behavior and need to be supported in a browsing system.

These prerequisites lead to an initial system design consisting of an analyzer, optimizer, and synthesizer. In the analyzer, the novel concept of associating time attributes with different document parts was introduced. The optimizer produces a multimedia navigation path through the document, given display and time

constraints. The synthesizer executes the instruction contained in the navigation path and creates a multimedia clip of a document – a Multimedia Thumbnail.

The observational user study helped evaluating the system and analyzing user- and task-specific browsing behavior. Although some behaviors can be generalized, people have different preferences about viewing documents. Therefore, it is important to design an MMNail generator which can be customized for personal preferences as well as for the end application. Our conclusions of the user study can be summarized as follows:

- Viewing of title and the thumbnail of first page is very important for the document search/browse task. Viewing of title, abstract, and figures is important for the document understanding task.
- Browsing styles can differ significantly from user to user. Therefore, allowing personalization of the guided document tour is useful.
- Generally users liked MMNail representation of documents.
- Synthesized audio quality is an important factor in the acceptance of the audio channel.
- Value of the audio channel is found to be more apparent in figure captions. Our interpretation is that, eyes move back and forth between figure and caption, and communication of the caption in the audio channel bridges this visual gap.

Using the results of the user studies, ideas for redesigning the technical components of the MMNail creation system were derived. Resulting modifications in the analyzer, optimizer, and interface enable a better adaptation of the MMNail representation to user and task specific needs.

The Multimedia Thumbnails system can be researched and developed in many different directions. The audio channel can be utilized to communicate non-text document information. For example, color blind people may want some information on colors in a document to be available as audible information in the audio channel. MMNail optimizations can be computed on the fly, based on interactions provided by user. For example, if the user slows down the visual channel (e.g., while driving a car), information delivered through the audio channel can be increased automatically. Furthermore, different complexity measures for figures, photos, tables, and graphs, can be developed as they require different levels of attention from users. High level content analysis, such as face detection, can be applied to assign information attributes to document elements. Moreover, better user interfaces can be designed for interaction with MMNails.

## REFERENCES

- [1] Lam, H., Baudisch, P. "Summary Thumbnails: Readable Overviews for Small Screen Web Browsers," Proceedings of the CHI, pp. 681-690, 2005.
- [2] Marshall, C.C, Ruotolo, C., "Reading-in-the-Small: a study of reading on small form factor devices," Proceedings of the JCDL'02, pp. 56-64, 2002.
- [3] Wang, M-Y., Xie, X., W-Y. Ma, H-J. Zhang, "MobiPicture - Browsing Pictures on Mobile Devices," Proceedings of International Conference of ACM Multimedia, pp. 106-107, 2003.
- [4] Breuel, T. M., Janssen, W. C., Papat, K., Baird, H. S., "Paper to PDA", Proceedings of the International Conference on Pattern Recognition, pp. 476-480, 2002.
- [5] Berkner, K., Schwartz, E. L., Marle, C., "SmartNails - Image and Display Dependent Thumbnails," Proceedings of SPIE, vol. 5296, pp. 53-65, San Jose, 2004.
- [6] Chen, F., Bloomberg, D.S., "Extraction of Indicative Summary Sentences from Imaged Documents," Proceedings of International Conference on Document Analysis and Recognition, pp. 227-232, 1997.
- [7] Alam, H., Hartono, R., Kumar, A., Rahman, A. F. R., Tarnikova, Y., and Wilcox, C., "Web Page Summarization for Handheld Devices: A Natural Language Approach", Proceedings of 7th Int. Conf. on Document Analysis and Recognition, pp. 1153-1157, 2003.
- [8] Adobe, PDF access for visually impaired, available at <http://www.adobe.com/support/salesdocs/10446.htm>
- [9] "Human Resources Toolbox," Mobility International USA, 2002, available at <http://www.miusa.org/publications/Hrtoolboxintro.htm>
- [10] Roth, P., Petrucci, L., Pun, t., Assimacopoulos, A., "Auditory browser for blind and visually impaired users," Proceedings of CHI, pp. 218-219, 1999.
- [11] Fan, X., Xie X., Ma W-Y., Zhang H-J., "Visual Attention based Image Browsing on Mobile Devices," Proceedings of International Conference on Multimedia and Exp, vol.1, pp. 53-56, Baltimore, MD, July 2003.
- [12] G. Salton, Automatic Text Processing, Addison-Wesley, 1989.
- [13] V. Eglin and S. Bres, "Document page similarity based on layout visual saliency: Application to query by example and document classification", Proceedings of International Conference on Document Analysis and Recognition, pp. 1208-1212, 2003.
- [14] Xie, X., Liu, H., Goumaz, S., Ma, W. Y., "Learning User Interest for Image Browsing on Small-form-factor Devices," Proceedings of ACM Conference on Human Factors in Computing Systems, pp. 671-680, 2005.
- [15] R. Neelamani, K. Berkner, "Adaptive Representation of JPEG 2000 Images using Header-based Processing", Proceedings of IEEE International Conference on Image Processing, pp. 381-384, 2002.
- [16] R.L. Rivest, H.H. Cormen, C.E. Leiserson, Introduction to Algorithms, MIT Pres, MC-Graw-Hill, Cambridge Massachusetts, 1997.
- [17] CVXOPT: A Python Package for Convex Optimization, available at <http://www.ee.ucla.edu/~vandenbe/cvxopt>