Maximizing Information Throughput for Multimedia Browsing on Small Displays

Xing Xie, Wei-Ying Ma, Hong-Jiang Zhang
Microsoft Research Asia
5F Sigma Center, No. 49, Zhichun Road, Beijing, 100080, P.R. China
{xingx, wyma, hjzhang}@microsoft.com

Abstract

As a great many of new devices with diverse capabilities are making a population boom, their limited display sizes become the major obstacle that has undermined the usefulness of these devices for information access. In this paper, we introduce our recent research on adapting multimedia content including images, videos and web pages for browsing on small-form-factor devices. A theoretical framework as well as a set of novel methods for presenting and rendering multimedia under limited screen sizes is proposed to improve the user experience. The content modeling and processing are provided as subscription-based web services on the Internet. Experiments show that our approach is extensible and able to achieve satisfactory results with high efficiency.

1. Introduction

In the PC+ era, a variety of new computing devices, such as SPOT watch, Smartphone, Pocket PC, Tablet PC, etc, are making a population boom. These devices are becoming more and more powerful in both numerical computing and data storage. However, low bandwidth connections and small displays are still the two serious obstacles that have undermined the usefulness of these devices in people's everyday life. With the rapid and successful development of 2.5G and 3G wireless networks, the bandwidth factor is expected to less constraint in the near future. At the same time, however, the limitation on display size is likely to remain unchanged for a certain period of time.

Since most of the information on the Internet is presented by multimedia (web pages with embedded images and audios can be considered a composite multimedia document), improving the experience of information access and browsing on small displays is critical to unleash the power of these mobile devices. Existing research directions to address this problem can be classified into the following four categories:

- *Trivial methods*. For example, direct down-sampling of image or video in the spatial domain. This approach often decreases the user experience since the results may be unreadable or unacceptable.
- *Authoring multiple versions*. For example, building separate, dedicated mobile web sites for small devices. This approach results in burden on content management. Also, it is hard to predict what devices will emerge in the market and the solution could be transient.
- *Re-authoring the content offline or on-the-fly*. This approach depends on the extraction of the original semantic structure of the content. Some success has been achieved in some areas but generally it is a hard problem because of the nature of reverse engineering.
- *New formats which are scalable by themselves*. This is the most promising direction and has been adopted in many areas, such as scalable image and video coding. However, the current research effort is less focused on the problem of diverse and small displays, and there is much space for improvement on multimedia browsing techniques.

In this paper, we focus on latter two approaches since they are more preferred by content authors or consumers. In fact, these two schemes are related to each other. The intermediate representation used in content re-authoring should be adaptive and flexible to the display size. Therefore, it will be referential when standardizing a new scalable format.

2. Related Work

So far only a few efforts have addressed the problem of browsing large web pages on small terminals and little has been done for images or videos. In the following, we will give a brief introduction to the prior art based on the media type that the content adaptation technique is designed for.

Typical web pages are designed for desktops with large displays. When they are browsed on small devices, the user experience is unacceptable. Current approaches for adapting web pages can be divided into two categories: the first one is to transform existing web pages such as [3], while the other attempts to introduce new formats and mechanisms. As we notice, few of current approaches considered the priorities of different parts in a page. What's more, none of them let authors control the final layout conveniently. That is to say, the final presentation is usually unpredictable during the designing phase.

Current digital cameras usually can take photos with more than 2M pixels. These photos should be down-sampled in order to be viewed on small devices like Smartphones. However, people might hardly catch the information, e.g., the human faces and texts in these down-sampled versions. Quite a few efforts have been put on image adaptation including JPEG and MPEG standards while proxy based image transcoding has also been studied for many years. Most of them focused on compressing and caching contents in order to reduce the data transmission time. Hence, the results are often not consistent with human perception because of excessive resolution reduction.

0-7803-8603-5/04/$20.00

Though more and more mobile devices are capable of playing videos, the limited bandwidth and small window sizes remain to be two critical obstacles. Currently, most video adaptation efforts only focus on bandwidth constraints. None of them has studied the impact of display resolution on the video browsing experience.

3. The Theoretical Framework

The following two observations are important to the development of our framework for optimizing viewer's browsing experience on small displays:

Information Asymmetry: Different parts of content have different importance values. Thus, there exists an optimal set of content blocks when a screen constraint is given. This observation has its root in psychology community. It has become clear that not all but only a small part of incoming visual information can reach short-term human memory for further processing, i.e., the Attention as Filter Metaphor. Attentional selection allows only attention-getting parts be presented to the user without affecting much user experience. For example, human faces in a home photo are usually more important than the other parts. Generally, most perceptible information can be located inside a handful of objects and at the same time these objects catch most attentions of a user. As a result, the rendering of content can be treated as manipulating objects to provide as much information as possible under resource constraints.

Flexible Rendering: The content layout should not be fixed to a specific display size. In other words, the layout should be optimized for each specified screen size. Therefore, we need not to design multiple versions for the same content. In addition, we should not restrict us to have exactly the same browsing experience as on desktop PCs. More advanced user interface technologies can be employed to improve the usability.

In summary, a scalable content model and a flexible rendering algorithm are two essential issues that we would like to address in our framework. Though different media types may need some customization, it is possible to develop a common content model to provide the basic operations for the optimization process. We will discuss this content model in the following.

3.1 A Content Model for Small Screen

A piece of media content P usually consists of several information objects B_i. An information object is an information carrier that delivers the author's intention and catches part of the user's attention as a whole. For example, it may be a human face, a flower or a text sentence.

Since each information object has different importance values, we introduce property *IMP* as a quantified value of author's subjective evaluation on an information object. It is also an indicator of the weight of each object in contribution to the whole information. This value is used when choosing less important objects for summarization under small displays. The importance values in the same content should be normalized so that their sum is 1.

As mentioned before, the information delivery of an object is significantly relying on its area of presentation. If an information object is scaled down too much, it may not be perceptible enough to let users catch the information that authors intend to deliver. Therefore, we introduce minimal perceptible size (*MPS*) to denote the minimal allowable spatial area of an information object. They are used as thresholds to determine whether an information object should be shrunken or summarized when rendering the adapted view.

As regards to those information objects of less importance, it is desirable to summarize them in order to save display space for more important objects. Instead of deleting contents or showing imperceptible adapted version, we introduce alternative (*ALT*) as a substitute of the original content. It should occupy less space than the original information object.

Our proposed content model for small screen presentation is defined as below.

Definition 1: The basic content representation model for a piece of media content P is defined as an unordered set of information objects:

$$P = \{B_i\} \quad 1 \le i \le N \tag{1}$$

and

$$B_i = (IMP_i, MPS_i, ALT_i) \tag{2}$$

where

B_i,	the i^{th} information object in P
IMP_i,	importance value of B_i.
MPS_i,	minimal perceptible size of B_i
ALT_i,	alternative of B_i.

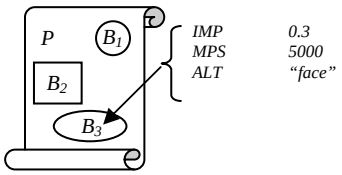

Fig 1. An example of the content representation model. In the example, three information object, *B1*, *B2* and *B3* are contained in the media content *P*. Each information object has three properties, *IMP*, *MPS* and *ALT*. For object *B3*, they are 0.3, 5000 and "face", respectively.

The representation can be in a form of XML descriptions and saved as metadata within original content. An example of the content representation model is shown in Figure 1 where three information objects are contained in the media content *P*.

3.2 Presentation Optimization

We introduce *Information Fidelity (IF)* as an objective comparison of a modified version of media content with the original version. The value of information fidelity is confined between 0 (lowest, all information lost) and 1 (highest, all information kept). It is defined as a sum of importance values of existing objects in the adapted version. If an object is replaced by its alternative, its importance value will not be included. Suppose P' is the set of existing information objects in the adapted version, $P' \subset P = \{B_1, B_2, ...B_N\}$. Thus, the mission of rendering phase is to find the set P' that carries the largest information fidelity while meets the display constraints.

In order to ensure that all the information objects are possible to be included in the final presentation, following space constraint should be satisfied.

$$\sum_{B_i \notin P'} size(ALT_i) + MPS_i \le Area \quad (3)$$

where *Area* is the size of target area and *size (x)* is a function which returns the size of display area needed by ALT_i. It says that the space occupied by the information objects or their alternatives should be smaller than the target display area.

If the constraint (3) is transformed to

$$\sum_{B_i \in P'} (MPS_i - size(ALT_i)) \le Area - \sum_{B_i \in P} size(ALT_i) \quad (4)$$

the rendering problem becomes:

$$\max_{P'} \left(\sum_{B_i \in P'} \mathrm{IMP}_i \right) \quad \text{subject to}$$

$$\sum_{B_i \in P'} (MPS_i - size(ALT_i)) \le Area - \sum_{B_i \in P} size(ALT_i) \quad (5)$$

We can see that the problem (5) is equivalent to a traditional NP-complete problem, 0-1 knapsack. It can be efficiently solved by a branch and bound algorithm.

4. Adapting Multimedia for Small Displays

In this section, we will show how the content model can be applied to adapt different types of multimedia content for small displays.

For web pages, we have introduced an approach [2] similar to the fisheye view. The extensions to the original representation model are mainly twofold:

- In order to let authors have controls on the final page layout, we leverage binary slicing trees, a data structure widely used in computer aided design community, instead of an unordered set to organize the information blocks.
- We add three additional properties to each information object in order to characterize their special display constraints.

Figure 2 shows the rendering results on three typical screen sizes for an example web page.

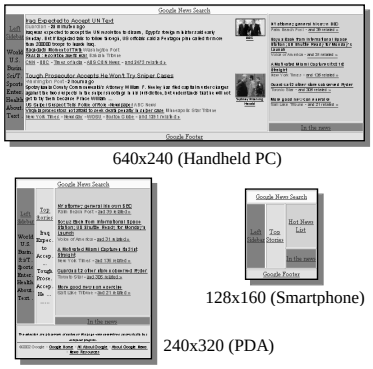

Fig 2. An example of adaptive web page rendering. The original web page comes from a Google news page and it has been adapted to three typical screen sizes: 640x240 (Handheld PC), 240x320(PDA) and 128x160(Smartphone).

For images, an attention model based adaptation and browsing scheme is developed in [1][5]. Besides that the notion of attention object is just equivalent to information object, other differences are:

- The image attention model adds a ROI property to each information object. It is borrowed from JPEG 2000 and is referred as a spatial region or segment that corresponds to an information object.
- We suppose the alternative of an object in images to be null since the information object will be cropped if it can not be put on the display.

As shown in Figure 3, the most important part of a large image is identified and cropped to fit the limited display size.

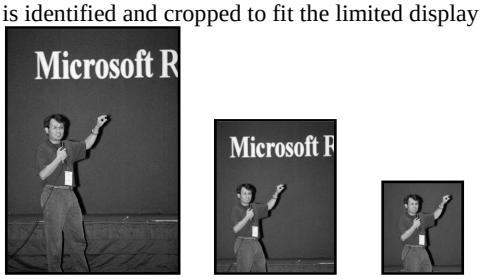

Fig 3. An example of attention based image adaptation. The original large image (768x1024) is cropped to fit two typical screen sizes: 240x320(PDA) and 128x160(Smartphone).

Video adaptation is another natural application of our content model. In [4], we proposed a solution for browsing amateur video clips such as home videos or surveillance videos. For this kind of video, it is possible to optimize the contents for different resolution conditions. Previous results on image adaptation can be easily extended to video adaptation if we simply consider each video frame as an image. However, this naïve approach will cause jitters in the video sequences since the frames will be discontinuous after cropping. To solve this problem, virtual camera control is applied to improve the quality of output stream. Figure 4 shows an example of the attention based video adaptation.

Fig 4. An example of attention based video adaptation. The human faces are detected in the video stream and only those regions are delivered to the client device. Therefore, he/she can see a more clear face image on the small screen and the bandwidth cost can also be reduced.

5. Content Services Networks

As the Internet is moving toward a service-centric model, more and more storages and computational resources are being put into the Internet infrastructure and provided as services to customers. In the content networking world, for instance, this trend can be seen on the development of content delivery networks (CDNs) which make content distribution a network infrastructure service available to content providers and network access providers. On the other hand, the progress on standardization and development of web services has marked the beginning of a new era that every computational resource and service on the Internet can be connected to provide new user experiences on accessing, sharing, and using information anytime, anywhere, from any device.

In this paper, we propose to provide content modeling and adaptation functions as subscription-based web services. It is based on our previous work named content services networks (CSN) [6] which aim to make content delivery networks (CDN) capable of delivering content adaptation services.

Figure 5 shows the overall system, which constitutes two layers of network infrastructures: content delivery overlay (i.e. CDNs) and service delivery overlay. The content delivery overlay is constituted of a network of service-enabled web caches which extend the functionalities of traditional web caches for performing value-added processing. The service delivery overlay consists of a large number of application servers which act as remote call-out servers for service-enabled web caches. These two overlays work together to provide content-oriented web services.

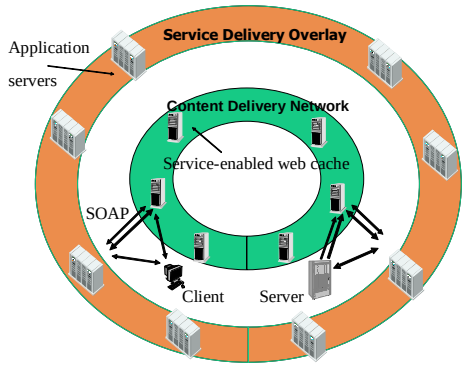

Fig 5. The architecture of content services network.

Before the content modeling and adaptation service becomes available, it needs to be registered in the UDDI (Universal Description and Discovery Integration) registry first. The received components such as service specifications and binaries from service providers are stored in the service database. In order to use the service, a mobile client needs to first find and subscribe to the service via UDDI registries. Then the service instructions are generated and transferred from the management servers to the service-enabled web caches that the subscriber is associated with. As shown in Figure 6, the service-enabled web cache determines if a message needs services according to the service instructions. In our case, the instructions may simply be type comparison, i.e., whether the content is an image or a video.

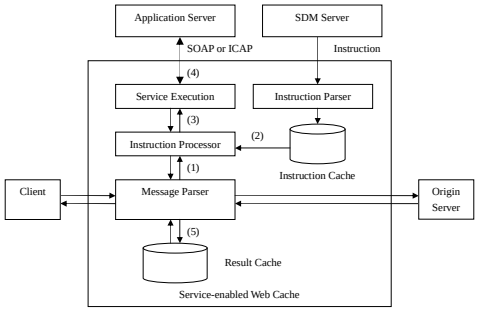

Fig 6. The service-enabled web cache.

6. Experimental Results

We have developed a service-enabled web cache based on Microsoft ISA Server 2000. A special web filter is implemented using ISAPI, which enables adaptation on HTTP messages containing HTML pages, images or videos. The processing is executed locally on the proxy which is a Windows XP system with P4 1.3 GHz CPU and 256M memory.

For web page adaptation, 16 web pages were collected from several popular websites such as MSN, Yahoo! and Google. The content model for each web page is manually created and the number of information objects in a web page varies from 5 to 20. In the experiment, the average time cost for adapting the page is 18 microseconds with variation from 2 to 56 microseconds.

Currently, the time cost for automatic image modeling process is a bit large. Averagely, it is 0.9 second for 1200x800 images on our test bed. However, as mentioned before, the automatic modeling results can be saved with the image files for reuse. Therefore, the content model is only computed once when the image is first acquired.

The performance of video adaptation can be improved if we do not process every frame in the MPEG stream. In our current implementation based on Microsoft DirectShow, we process one frame every three seconds. It can run smoothly on the test bed without causing any jitter.

7. Conclusions

In this paper, we introduced our work on adapting multimedia to small-form-factor devices. Our approaches can be easily extended to other various applications such as information summarization or thumbnail generation, since space limit is also the critical issue there. How to enable service composition, federation, and service-peering is another interesting and challenging research problem. Furthermore, the issue of data integrity also needs to be addressed as we allow content to be modified by a third-party service provider. We will continue to investigate these directions in the future.

8. References

[1] Chen, L.Q., Xie, X., Fan, X., etc. A Visual Attention Model for Adapting Images on Small Displays. ACM Multimedia Systems Journal 9(4) 2003, 353-364.

[2] Chen, L.Q., Xie X., Ma W.Y., etc. DRESS: A Slicing Tree Based Web Page Representation for Various Display Sizes. Poster Proc. WWW'03, Budapest, Hungary, May 2003.

[3] Chen, Y., Ma, W.Y., and Zhang, H.J. Detecting Web Page Structure for Adaptive Viewing on Small Form Factor Devices. Proc. WWW'03, Budapest, Hungary, May 2003.

[4] Fan X., Xie X., Zhou H.Q., and Ma W.Y. Looking Into Video Frames on Small Displays. Poster Proc. ACM Multimedia'03, Berkeley, CA, USA, Nov. 2003.

[5] Liu H., Xie X., Ma W.Y., and Zhang H.J. Automatic Browsing of Large Pictures on Mobile Devices. Proc. ACM Multimedia'03, Berkeley, CA, USA, Nov. 2003.

[6] Ma W.Y., Shen B., and Brassil J. Content Services Network: The Architecture and Protocols. Proc. WCW'01, Boston, USA, Jun. 2001.