Multimedia Simplification for Optimized MMS Synthesis

Wei-Qi Yan
Department of Computer Science
University of California, Irvine
CA 92697-3425

Mohan S. Kankanhalli
Department of Computer Science
National University of Singapore
Singapore 117543

We propose a novel transcoding technique called *multimedia simplification* which is based on experiential sampling. Multimedia simplification helps optimize the synthesis of MMS (Multimedia Messaging Service) messages for mobile phones. Transcoding is useful in overcoming the limitations of these compact devices. The proposed approach aims at reducing the redundancy in the multimedia data captured by multiple types of media sensors. The simplified data is first stored into a gallery for further usage. Once a request for MMS is received, the MMS server makes use of the simplified media from the gallery. The multimedia data is aligned with respect to the time-line for MMS message synthesis. We demonstrate the use of the proposed techniques for two applications namely soccer video and home care monitoring video. The MMS sent to the receiver can basically reflect the gist of important events of interest to the user. Our technique is targeted towards users who are interested in obtaining salient multimedia information via mobile devices.

Categories and Subject Descriptors: H.5.1 [**Multimedia Information Systems**]: Miscellaneous; H.4.0 [**Information Systems Applications**]: General

General Terms: Documentation, Languages

Additional Key Words and Phrases: Multimedia Simplification, MMS Synthesis, Soccer Video, Home Care monitoring, Hypermedia Coherence, Mobile Phone, Experiential Sampling.

1. INTRODUCTION

Mobile phones have become an integral part of our lives. Nowadays, they come integrated with multimedia devices such as a camera, speaker, radio and a microphone. While primarily facilitating tele-conversations, it also offers additional services such as text communication, games, audio/video playing, radio, image/video capture and transmission, alarm, calculator and calendar. More recently, sending and receiving of MMS (Multimedia Mes-

Author's address: School of Computing, National University of Singapore, 3 Science Drive 2, Singapore 117543

saging Service) messages, which has a substantially higher information-carrying capacity than the SMS (Short Message Service) text messages, has attracted millions of enthusiasts. A MMS with pictures and audio is more impactful in delivering messages than mere words and sentences. Text and captions can additionally help in conveying the content of a MMS message. MMS messages are like an automated powerpoint slide-show. The multimedia presentation, which consists of elements such as music, voice, images, text, and graphics are synchronized along the same time-line. When the MMS message is displayed, the presentation – which looks like a choreographed slide show with images (both photographic and animation) and sound, starts running. Thus the basic element of a MMS message is a slide and the entire message consists of a sequence of such slides. Slides have image, text and audio components, organized in the manner of the spatio-temporal layout of a SMIL presentation. A MMS message can currently embed multimedia components of size up to 300 Kilobytes per slide (though there is no limit prescribed by the standard) which is sufficient for most purposes of query, communication, and entertainment. Further details about the MMS standard can be obtained from http://www.openmobilealliance.org.

3G mobile phones are still not widely in use. As a result, the standard mobile phone is the main platform for sending and receiving media data from the MMS communication network. It can only serve as a playing device for the audio and picture components of MMS messages and it does not allow for interaction. Moreover, the huge size of multimedia data makes its transfer through the wireless channel slow. We need to make full use of the limited memory to embed the most crucial information that could convey our intent. The limitations of mobile phone usage include:

—Narrow bandwidth due to wireless communication;

—Limited screen resolution and audio capabilities;

—Relatively small amount of available memory space;

—Slow speeds for data storage & transmission via secondary mechanisms – such as bluetooth and infrared ports;

These problems are unlikely to disappear with 3G phones since the permissible MMS message size will also increase correspondingly. Therefore, a video is not allowed to be inserted into a MMS – it needs to be broken up into a demultiplexed sequence of still images and audio clips before being inserted into a MMS. On the other hand, digital images with a high quality, such as photographs, graphics and electronic maps have to be compressed, resized or cropped so as to conform to the device limitations. Reducing the data size with minimal concomitant loss of information is extremely valuable for MMS message synthesis.

Compression is an obvious way of reducing the data size. However compression techniques are not able to selectively identify the most important content in a message. Compression techniques tend to work by reducing the statistical, coding and perceptual redundances. We therefore introduce the notion of simplification. We define the process of selecting the semantically most important part of the media data as **multimedia simplification**. It crucially differs from compression in the sense it aims to minimize the "semantic redundancy".

Experiential sampling [Kankanhalli et al. 2006] generalizes the notion of *attention* such that it can be applied to multiple correlated media streams. It introduces the notion of 'attention saturation' which is a measure of importance in a media frame. It has been

successfully employed for video presentation on tiny devices [Wang et al. 2004]. The strength of the experiential sampling technique lies in its ability to perform summarization/simplification on data from different modalities using a unified process. The experiential sampling system is an adaptive sampling system which differs from techniques used in video summarization [DeMenthon et al. 1998]. In this paper, we utilize this technique to perform simplification of text, audio and motion pictures, and retain the most important information of each stream. This could be key sentences in text, key regions of a picture and non-silent regions of an audio clip.

Multimedia simplification is related to but is very different from summarization. It attempts to pick the semantically most salient aspects of the media streams. *Simplification focuses on extracting the semantic saliency of the multimedia data while summarization attempts maximal data reduction with minimal loss in semantics*. Thus, simplification is always semantically lossy which preserves only the most salient aspects for a given data size constraint. It helps obtain the most compact representation to convey the semantic saliency. On the other hand, summarization tries to minimize this semantic loss while maximizing the data compactness. In other words, the summary tries to provide a gist of the entire content – the salient aspects as well as some idea of the non-salient aspects. It must be noted that a drastic amount of summarization will asymptotically reduce to simplification since the salient aspects will be preserved at the cost of the non-salient aspects. However, there is another crucial difference between simplification and summarization in the case of multimedia data. For multiple correlated streams, summarization faithfully tries to preserve the individual stream semantics without trading off one against the other. On the other hand, simplification on correlated streams can potentially do cross-modal tradeoffs. For instance, the video stream could be totally discarded if the audio stream can completely convey the salient aspect (e.g. the current score of a live game).

Audio simplification is equivalent to experiential sampling for audio data wherein the low information-content signals within a window will be gradually ignored and dropped based on the attention saturation [Kankanhalli et al. 2006]. For example, silence regions will be considered having low information and thus less important, and will end up being discarded. Text simplification retains the high frequency or highly weighted keywords after the stop words removal (such as prepositions, conjunctions, and articles etc) [Xie et al. 2004]. Based on the attention saturation of these words, highly attended sentences and paragraphs are retained. Image simplification will discard plain regions of a picture having a lot of color variation and edges. The contour shape after edge fusion will be thought as the shape of an object where the attention saturation of different objects will be used to identify the relative importance.

For MMS synthesis, a further optimization step is required for multimedia simplification which explores the tradeoff between the media size and the semantically salient content so as to effectively convey the message. The most important content can then be transmitted via MMS based on the end user's requirement. In order to present the multimedia message on a tiny device, media data reflecting the same event are collected from different sources. The sources of MMS components and their usages are listed in Table I. Since in one MMS slide, we can only provide one image, one audio clip and one paragraph of text, we need to select the most salient parts.

In other words, in order to create the optimal MMS message for the small form-factor device, we need to select the semantically most salient media data from the gallery (repos-

Table I. Multimedia sources and the usages

Media	Sources	Usages
Audio	phone, video camera, microphones, web, etc.	MMS, broadcasting, Personal entertainment, etc.
Text	authored, newspapers, web pages, SMS, emails, etc.	SMS, emails, etc.
Pictures	phone camera, digital camera, web pages, etc.	MMS, web pages, Personal entertainment, etc.
Video	video camera, phone camera web camera	MMS, broadcasting, Personal entertainment, etc.

itory) and combine them in the most effective manner. Such an optimally composed MMS message will best convey the semantic intent. To the best of our knowledge, this is the first time that multimedia simplification is used for automatic composition of MMS messages. We present the theoretical background, algorithms as well as the implementation results.

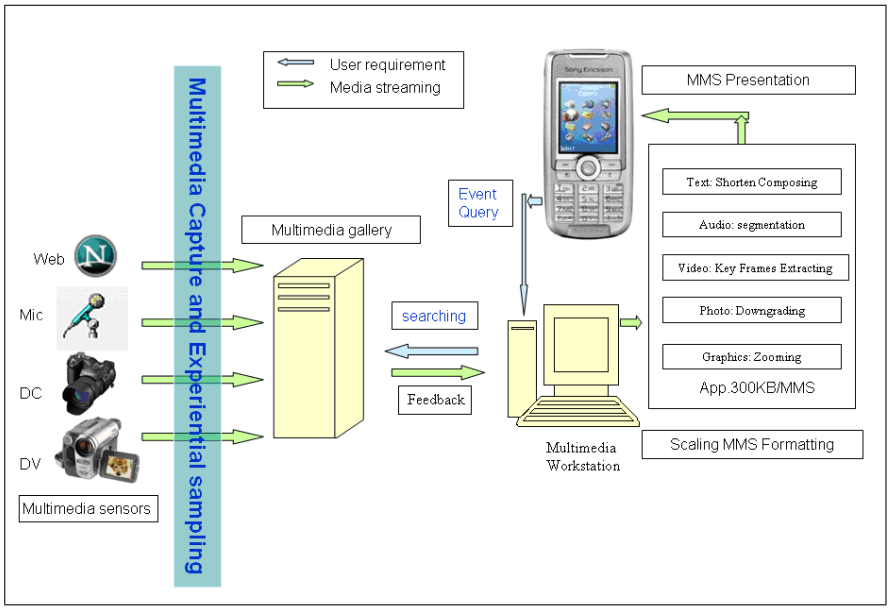


Fig. 1. Multimedia presentation on mobile phones

Figure 1 describes the flowchart for multimedia presentation on mobile phones. Once the MMS workstation (server) receives a request from a cellphone user, the server computes the relevant data for this event and collects the most representative information automatically. Once the data have been aligned on the time line, it is re-organized to conform to the MMS format via splitting and packing. Then the MMS sequence is sent to the cellphone user for display.

This paper is organized as follows: section 2 introduces the related work, section 3 describes our techniques for multimedia simplification, section 4 presents our proposal

for MMS synthesis, section 5 demonstrates the results, and the conclusion is drawn in section 6.

2. RELATED WORK

Experiential sampling has been utilized for video surveillance with multiple cameras [Wang et al. 2003]. The algorithm is built on importance sampling and linear dynamical system [Cochran 1977]. Experiential sampling selects the relevant data based on the context and past history. The successful application of this technique motivates us to explore its utility for MMS message synthesis.

During the MMS synthesis process, synchronization and fusion among different types of multimedia are necessary. The framework in [Wickramasuriya et al. 2004] utilizes different modalities of sensor data in providing information that assists the video subsystem in detecting and classifying anomalies. Cross-media correlation in [Chu and Chen 2002] proposes two types of multimedia correlation: explicit and implicit relations. The proposed synchronization techniques include speech-text alignment processing in temporal domain, automatic scrolling process in spatial domain and content dependency check process in content domain. The SnapToTell system [Chevallet et al. 2005][Lim et al. 2004] provides information directory service to tourists based on pictures taken by the camera phones and location information. A method for generating metadata for photos using spatial, temporal, and social context is provided in [Davis et al. 2004]. MMG (Mixed Media Graph) [Pan et al. 2004] is a graph-based approach to discover the cross-modal correlations. The model can be applied to diverse multimedia data to find multi-modal correlations. However, the application of this approach has constraints in the sense it needs to build a big multimedia database for the collections and also needs to extract the features of these multimedia data, with the similarity function for each medium properly defined.

Assuming that each of the media stream has a priori probability of achieving the goal and their underlying correlations are known, the assimilation framework in [Atrey and Kankanhalli 2004] fuses the individual probabilities using the quantitative correlation based on a Bayesian approach and a heuristic function. The cross-media adaptation strategy in [Boll et al. 1999] provides models for the automatic augmentation of multimedia documents by semantically equivalent presentation alternatives. The cross-media adaptation strategy allows for flexible reuse of multimedia content in many different environments and maintains a semantically correct information flow of the presentation. The concepts, techniques and issues of video adaptation are described in [Chang and Vetro 2005]. Video adaptation transforms the input video(s) to an output in video or augmented multimedia form by utilizing manipulations at multiple levels (signal, structural, or semantic) in order to meet diverse resource constraints and user preferences while optimizing the overall utility of the video.

Xie et al [Xie et al. 2004] present an approach for discovering meaningful structures in video through unsupervised learning of temporal clusters and associating them with metadata using co-occurrence analysis and models similar to machine translation. The patterns in video are modelled with Hierarchical Hidden Markov Models (HHMM) [Duda et al. 2000], with efficient algorithms to learn the parameters, the model complexity, and the relevant features; the meanings are contained in words of the speech transcript of the video. The pattern keyword association is obtained via co-occurrence analysis and statistically machine translation models.

[Duygulu et al. 2003] combines speech, image and natural language understanding to automatically transcribe, segment and index video for intelligent search and image retrieval with automatically generated metadata and indices for retrieving videos from the library with thousands hours of video with over two terabytes of data. The integration of visual and textual data is proposed to solve the correspondence problem between video frames and associated text [Duygulu and Wactlar 2003].

About MMS related works, Coulombe and Grassel [Coulombe and Grassel 2004] provide an overview of the multimedia messaging service. It also addresses the interoperability challenges this new service brings as mobile terminal capabilities evolve at a very fast pace. It explains how server-side multimedia message adaptation technologies can provide smooth format and service evolution while ensuring interoperability. IBM demonstrated a interactive MMS development platform in 2003 [Jun et al. 2003]. Recently, [Lin and Tseng 2005] described a user-oriented video semantic filtering and abstraction method. Although it appears similar to our proposal, it does not cater for multiple media (audio and text) processing.

In contrast with all of the past work, in this paper, we propose the use of multimedia simplification for the purpose of MMS synthesis. It reduces the redundancy of multimedia content while exploiting its correlated nature.

3. MEDIA SIMPLIFICATION

For MMS message synthesis, we need to consider all kinds of media types including text, image or picture, video, and audio. Since our aim is to send MMS messages from a server workstation to the cellphone via the wireless channel, the MMS message is composed of three types of media: audio, picture and text. All the media will have to be converted to the three formats before being transmitted and played. In this section, we will discuss on how to simplify the various kinds of media from different resources to the above three types based on experiential sampling.

3.1 The Experiential Sampling (ES) technique

The experiential sampling technique computes the generalized goal-oriented attention based on a dynamical system framework [Jacobs 1993][Welch and Bishop 2001]. It analyzes the multimedia streams to select the most important portions and discards irrelevant data. Highly attended data are captured by a sampling representation. Based on the task at hand, the important portions of the media gain a high attention which is probed by the sensors samples randomly distributed among the multimedia content. The high attention regions are then obtained based on the past history and the current context. The high attention regions are described by the density and distribution of the attention samples. The computed attention samples are captured as a dynamical system which can use them to predict the attention samples in the next time instant [Kankanhalli et al. 2006]. The predicted attention samples are re-sampled to take into account the current context. The experiential sampling algorithm is described in Algorithm 1.

In algorithm (1), $Random(M,t)$ randomly samples the data of media M at time t. $\tau_1 > 0$ and $\tau_2 > 0$ are the experiential sampling thresholds for the sampling and re-sampling. $SS(t)$ is the set of uniform random samples at any time t which constantly sense the environment. $AS(t)$ is the set of dynamically changing attention samples which essentially represents the data of interest at time t. $AS(t)$ is calculated, adjusted and filtered from $SS(t)$, after discarding the irrelevant data. A_{sat} is the total amount of attention at a time instant, $N_A(t) =$

Input : Source media M
Output : Attention Samples $AS(t)$

Procedure:
1. Initialization: $t = 0$;
2. $SS(t) \leftarrow Random(M,t)$;
3. $SS(t) = \{SS_i(t) : i = 1,2,\cdots,|SS(t)|\}$;
4. $A_{sat}(t) = \{SS_i(t) : SS_i(t) > \tau_1; i = 1,2,\cdots,|SS(t)|\}$;
5. $N_A(t) = |A_{sat}(t)|$;
6. if $(N_A(t) = 0)$, $t = t+1$; goto step 2;
7. $AS(t) = \{A_{Sat_i}(t) : A_{Sat_i}(t) > \tau_2; i = 1,2,\cdots,N_A(t)\}$;
8. $SS(t) \leftarrow AS(t) + Random(M,t+1)$;
9. $t = t+1$; goto step 3;

Algorithm 1: The Experiential Sampling Algorithm

$|A_{sat}|$ is the number of the attention samples. The number (possibly zero) of the attention samples aroused depends on the context. The exact location of the attention samples also depends on the sensed environment. In this paper, we use experiential sampling to select the semantically most salient information.

3.2 Text simplification

Text is regarded as a one-dimensional signal. Suppose the vocabulary set of a text is given as $\Gamma = \{\omega_1, \omega_2, \cdots, \omega_n\}$, $\omega_i, i = 1,2,\cdots,n$ are the words in the context, the set of stop words is $\Xi = \{\theta_1, \theta_2, \cdots, \theta_m\}$ whose examples are shown in Table II, which normally has a high frequency of occurrence. Thus, θ_i, $i = 1,2,\cdots,m$, $m < n$ are the stop words and so the relevant words are $\Xi \in \Gamma$. The application of the ES algorithm in text processing is to find the important sentences and paragraphs using the relevant keywords set $\Lambda = \Gamma - \Xi = \{\lambda_1, \lambda_2, \cdots, \lambda_{n-m}\}$ appearing in the context, where λ_i, $i = 1,2,\cdots,n-m$ are the relevant words. Only the frequencies of relevant words in the context $F_\Gamma = \{f_1, f_2, \cdots, f_{n-m}\}$ are used to compute the salience of the objects, where f_i, $i = 1,2,\cdots,n-m$ is the frequency of the i^{th} relevant word. The vocabulary set of relevant words in this ES algorithm is treated as the basic unit for the determining the saliency of sentences and paragraphs.

Given a sentence t, the stop words are initially discarded. The remaining words are regarded as the sensor samples $SS(t)$. The frequency of each word in the whole context is used to determine the attention samples $AS(t)$ of this sentence.

$$AS_s(t) = \frac{\Sigma_{k=0}^{C_w^s} f_k}{C_w^s} \tag{1}$$

where C_w^s is the total number of words in this sentence inclusive of the stop words, f_k is the frequency of the k^{th} word. Building upon on the notion of the sentence attention, the attention samples of a paragraph at position t is given by:

$$AS_p(t) = \frac{\Sigma_{k=0}^{C_s^p} AS_s(k)}{C_s^p} \tag{2}$$

Table II. Stop words removed from context

Type	Samples
Pronouns	I, you, he, she, it, we, they, them,
	us, me, him, her, its, my, your, etc.
Prepositions	for, from, of, up, down, in, out, or, and,
	to, below, with, by, how, what, etc.
Conjunctions	and, so, therefore
Articles	a, an, the
Auxiliary	will, would, shall, should,
	must, may, might, can, could, etc.
...	...

where C_s^p is the total number of sentences in the paragraph, $AS_s(k)$ is the attention of the k^{th} sentence in the paragraph and $AS_p(t)$ is the attention of this paragraph. Essentially, we use the frequency of relevant words as a basis for measuring saliency and build up the saliency of sentences and paragraphs.

We now present some results from the Singapore prime minister's speech in 2003. The speech has a total of 12204 words, 744 sentences, and 28 paragraphs. Some high attention keywords are: 'singapore', 'singaporeans', 'many', 'jobs', 'SARS', 'work','CPF', etc.; some key sentences are: "i graduated from the university of Singapore in 1964"; "their departure would throw 100,000 out work"; "dessmon explained to her that he was trained to do job";"many people died";"now it rising again";"leaves wages CPF", and so on. The fifth paragraph is found to be most important using our system. From these results of text simplification, one can surmise that the speech was about the 'SARS' breakout and its impacts on Singapore in 2003.

Our proposal for text simplification can thus capture the key words, sentences and paragraphs at different levels based on the statistical frequency measure. With more contextual information and higher order correlations, the salient information can be captured more accurately.

3.3 Picture simplification

Pictures are two-dimensional signals. Experiential sampling was first used to reduce the redundancy in surveillance video [Kankanhalli et al. 2006]. We employ the experiential sampling technique on pictures to obtain the most salient region of the picture. With the availability of high resolution cameras, images tend to be detailed and huge. Therefore, they cannot be fitted into a MMS message directly. We could possibly reduce the resolution via sub-sampling, which can however result in the loss of details, which may not be desirable. Hence, it would be better to have a cropped but detailed rendition of the most salient region of the image to be included in the MMS message.

—*Simplification of Still Pictures*

We first calculate the gradient of the image using [Rosenfeld 1969]:

$$\overrightarrow{g}I = \left(\frac{\partial I(x,y)}{\partial x}, \frac{\partial I(x,y)}{\partial y}\right) = (I_x, I_y) \tag{3}$$

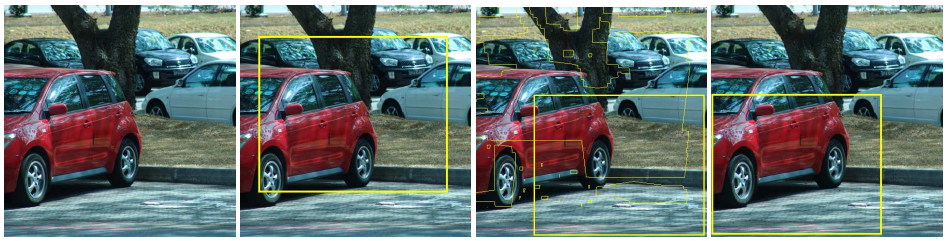
where $I(x,y)$, $x = 1,2,...,W$; $y = 1,2,...,H$; W and H are width and height of the image respectively. We compute the intensity changes to discard the low-gradient regions:

$$\overrightarrow{g}I(x,y) = (I_x(x,y), I_y(x,y))$$
$$= (\frac{I(x+\triangle x,y) - I(x,y)}{\triangle x}, \frac{I(x,y+\triangle y) - I(x,y)}{\triangle y}) \quad (4)$$

$\triangle x$ and $\triangle y$ are the step lengths in the X and Y direction respectively. The sensor samples of the ES algorithm are $SS(t) = \{\overrightarrow{g}I(x,y), (x,y) \in W_t\}$ where W_t is a candidate window. If the $|\overrightarrow{g}I(x,y)| > \tau_1$, then $I(x,y)$ belongs to the salient region. The attention saturation is $A_{sat}(t) = \{\overrightarrow{g}_t I(x,y) : |\overrightarrow{g}_t I(x,y)| > \tau_1, \tau_1 > 0\}$. The attention samples are determined by the ratio between the salient regions and the non-salient regions, namely:

$$AS(t) = \{\overrightarrow{g}_t I(x,y) : \frac{No_S(t)}{No_N(t)} > \tau_2, (x,y) \in W_t\} \quad (5)$$

where $No_S(t)$ and $No_N(t)$ are the numbers of salient and the non-salient regions. Only the important region $\Omega = \{W_t : \frac{No_S(t)}{No_N(t)} > \tau_2, \tau_2 > 0\}$ will be selected and presented on the mobile device, W_t is t^{th} window on the picture. Basically, this technique is based on image segmentation which utilizes image block fusion and multi-level selection based on thresholds.

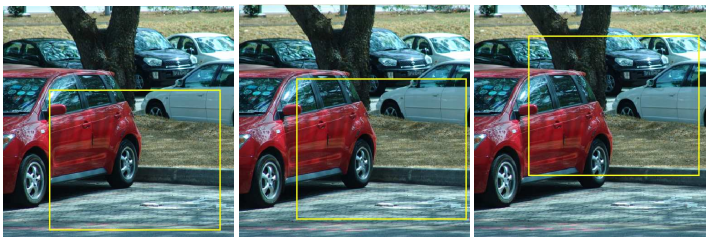


(a) Original picture

(b) The salient region (manually selected)

(c) Picture with extracted edges

(d) Picture with importance at 70%



(e) Picture with importance at 80%

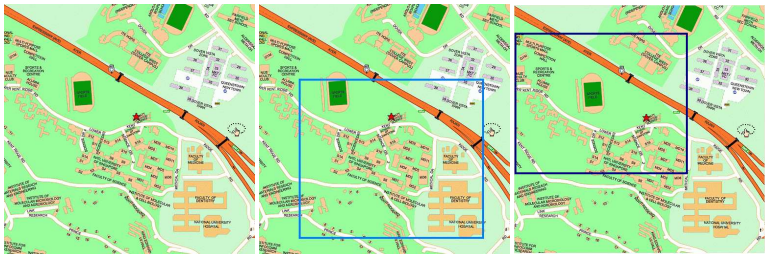
(f) Picture with importance at 90%

(g) Picture with importance at 100%

Fig. 2. Picture simplification of an image for different ratios

Figure 2 shows an example of still picture simplification. Fig. 2(a) shows the original picture, Fig. 2(b) shows the salient region (manually marked). Fig.2(c) is the image

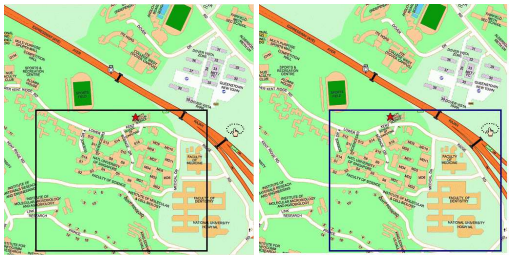
after automatic edge extraction and adjacent block fusion. Fig. 2(d), Fig.2(e), Fig. 2(f) and Fig. 2(g) show the automatically cropped figures at 70%, 80%, 90% and 100% saliency ratio. We calculate the ratio between the number of the salient regions to the number of the non-salient regions within a window (352×288). Once a user decides his requirement for the saliency ratio, it can quickly provide the corresponding picture (which is not necessarily unique).



(a) A map of NUS.

(b) The most salient region of the map(manually selected).

(c) A simplified map of NUS at 90% ratio



(d) A simplified map of NUS at 95% ratio

(e) A simplified map of NUS at 100% ratio

Fig. 3. Picture simplification for a map

Figure 3 presents the results on a map image. A map of an university campus shown in Fig. 3(a) is used to find the most salient part. The star marks the position of destination. Fig. 3(b) is the most salient region that we manually selected. From Fig. 3(c), Fig. 3(d), and Fig. 3(e), we can see that the desired destination is included in the automatically simplified pictures.

—*Simplification of Motion Pictures*

The experiential sampling technique has been used to extract the important frames from a video [Wang et al. 2003]. For our problem, video data cannot be directly inserted into the MMS slides. We have to split the video into the audio track and motion pictures first. For the motion pictures, we need to extract the salient portions of the important frames. We assume that we have the entire set of full pictures for an event [Medioni et al. 2001][Zhang et al. 1993]. We then select the most important frames from the entire sequence. For a single picture, we can use the earlier technique to perform still picture

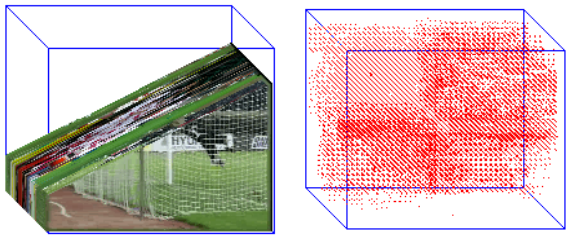
simplification. But this will lead to a problem. For a sequence of pictures in a video, different frames may be cropped in different places with different aspect ratios. So if we run a slide-show of such sequence of individually simplified frames, we are likely to observe a jitter. Hence, for motion pictures, we need to consider the entire sequence of frames in order to eliminate jitter. For this purpose, we first compute the motion trajectory of the single-frame simplification results with respect to the original frames. We then perform a trajectory smoothing operation based on either the mean trajectory technique or the Bezier curve fitting technique as in [Yan and Kankanhalli 2002]. Given a motion picture set $M_P = \{P_1, P_2, \cdots, P_n\}$, the gradient of the motion pictures is given as follow:

$$\overrightarrow{g} P_t = (\frac{\partial P_t(x,y)}{\partial t}, \frac{\partial P_t(x,y)}{\partial x}, \frac{\partial P_t(x,y)}{\partial y}) \tag{6}$$

For discrete frames, it reduces to:

$$\overrightarrow{g} P_t(x,y) = (\frac{P_{t+\triangle t}(x,y) - P_t(x,y)}{\triangle t}, \frac{P_t(x+\triangle x,y) - P_t(x,y)}{\triangle x}, \frac{P_t(x,y+\triangle y) - P_t(x,y)}{\triangle y}) \tag{7}$$

where $t = 1, 2, \cdots, n$; $\triangle t$, $\triangle x$ and $\triangle y$ are the step length in different directions. For the salient regions of each picture $P_i = \{O_{i,1}, O_{i,2}, \cdots, O_{i,m}\}$ selected from Eq. (5), the importance of each object is $|O_{i,j}|$, $i = 1, 2 \cdots, n$; $j = 1, 2, \cdots, m$; $|P_i| = max_{j=1,2,\cdots,m}(|O_{i,j}|)$. The importance of that frame among motion pictures is measured by $|P_i|$. Using this information, we can pick the salient frames as well as the salient regions of every picked frame. The motion trajectory can then be computed which is then smoothed. The final set of salient regions is then obtained.

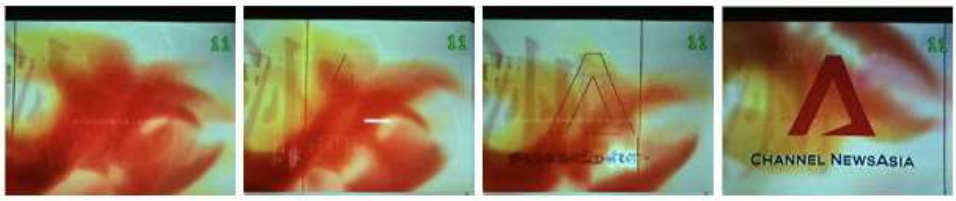


(a) Video volume cross-section (b) Sampling of the video

Fig. 4. Analysis of motion pictures

Figure 4 depicts the uniformly extracted figures for simplification analysis. Fig. 4(a) is the cross-section of a picture sequence in which we can visualize the event of a soccer goal being scored. Fig. 4(b) shows the attention samples for this picture sequence. It can be seen that it captures the salient event in the motion pictures [Bennett and McMillan 2003].

Figure 5 is an example of simplification for motion pictures, Fig. 5(a) shows some selected motion pictures for the television advertisement of the *Channel News Asia* broadcast channel. Fig. 5(b) is the extracted key frames with over 90% importance. These



(a) Original motion pictures

(b) Simplified key frames

Fig. 5. The simplification for motion pictures

frames basically reflect the main content of this set of motion pictures, which is the Ad for the channel.

3.4 Audio simplification

Audio simplification aims to find the most salient content in an audio clip. The saliency is determined by the energy and spectrum of the audio content. It is based on intra- and inter-frame analysis. We use the experiential sampling technique to discard the low attended audio frames. Consider an audio clip: $\Upsilon = \{\upsilon_i, i = 1,2,\cdots,n\}$, where υ_i are audio frames; for each frame, the samples are $\upsilon_i = \{\psi_{i,1,1}, \psi_{i,2,1}, \psi_{i,\cdots,1}, \psi_{i,j,1}, \psi_{i,1,2}, \psi_{i,2,2}, \psi_{i,\cdots,2}, \psi_{i,j,2}, \cdots, \psi_{i,j,m}\}$, i is number of this frame, j is the number of channels and m is number of samples per frame. The size of each frame in bytes is given by:

$$L = \frac{m \times j \times w}{8} \tag{8}$$

where w is the number of bits allocated for each sample. The sensor samples $SS(t)$ for audio experiential sampling (using the ES algorithm) are calculated as:

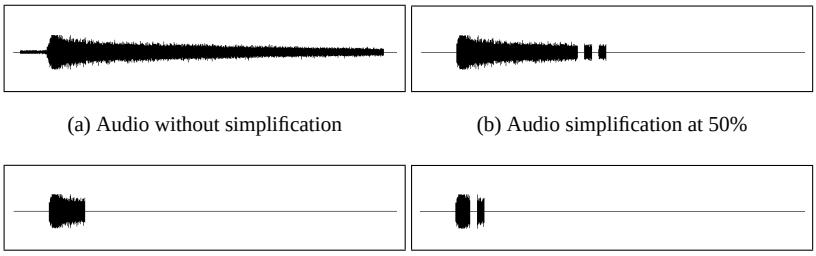
$$SS(i) = \{\upsilon_i : \Sigma_{i=1}^{n} |\psi_{i,j_0,k} - \psi_{i-1,j_0,k}| > \tau_1\} \tag{9}$$

$$A_{sat}(i) = \{\upsilon_i^1 \in SS(i) : \Sigma_{k=1}^{m} |\psi_{i,j_0,k} - \psi_{i,j_0,k-1}| > \tau_2\} \tag{10}$$

The attention samples can now be computed:

$$\begin{aligned} AS(i) \;=\; & \{\upsilon_i^2 \in A_{sat}(i) : |max_{1\le k\le m}(\psi_{i,j_0,k}) \\ & -min_{1\le k\le m}(\psi_{i,j_0,k})| > \tau_3\} \end{aligned} \tag{11}$$

where $\tau_i > 0$, $i = 1, 2, 3$ are the thresholds for determining the different degrees of salience.



(a) Audio without simplification (b) Audio simplification at 50%

(c) Audio simplification at 80% (d) Audio simplification at 90%

Fig. 6. Audio simplification at various saliency levels

Figure 6 illustrates results for audio simplification at different levels. Fig. 6(a) shows the original audio, Fig. 6(b), Fig. 6(c), and Fig. 6(d) are the waveforms after the audio simplification at 50%, 80% and 90% saliency respectively. From these figures, we can clearly see that only the important frames are retained. In general, silence, simple repeats and low volume frames will be discarded as they are considered having low attention. For a specific type of audio such as music, algorithms segmentation and event detection [Gao et al. 2004] can be employed to refine the notion of saliency.

3.5 Time-line synchronization information

For multimedia communication, semantic coherence between multiple types of multimedia is critical for properly conveying the intent. The various media components need to carefully be synchronized along a common time-line for the story to properly unfold as the presentation is played. While performing individual media simplification, the critical attributes of time-stamps, frame-rates and spatial resolutions should be carefully recorded. This will be the basis for the MMS message synthesis. For example, figure 7 illustrates the time line for video editing using the Microsoft MovieMaker. It ensures coherence among motion pictures, audio and titles by synchronizing them with respect to this common time-line.

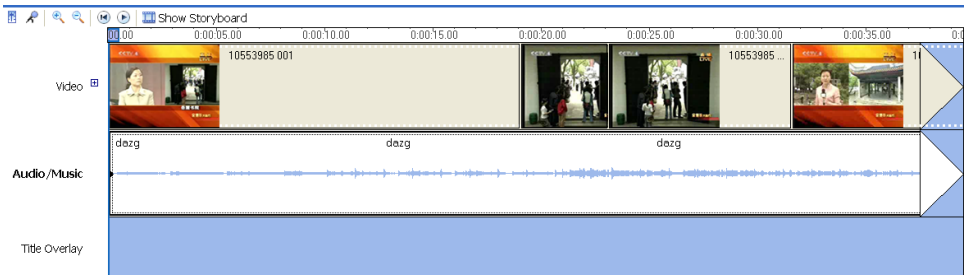


Fig. 7. Time-line based media coherence

4. OPTIMIZED MMS MESSAGE SYNTHESIS

In this section, we will discuss on how to synthesize a MMS file from the media data in a gallery. We consider the coherent composition of video, audio, image and text. There is an optimization issue in MMS message synthesis because each MMS slide can accommodate a limited amount of data S_{Slide} (300KB currently). We need to simplify and synchronize the media components based on the constraint:

$$S_{audio} + S_{pic} + S_{txt} = S_{Slide} \tag{12}$$

where S_{audio}, S_{pic}, and S_{txt} are the data sizes of audio, picture and the text. Moreover, only one picture is allowed per slide. Once the size of the embedded picture is fixed, the rest of the space is shared by audio and text. Normally, text does not take up much space. Also, the words in MMS messages are written in an abbreviated form, e.g. long words are replaced shorter homonyms. For example, 'to' is replaced by the numeral '2', 'for' by '4', 'before' is shortened to 'b4', and 'you' is replaced by 'u'. This is also referred to as the 'SMS vocabulary'. We apply these transformations while synthesizing MMS messages. Audio uses up the remaining space quota. The coherence requirement necessitates the coordination of the audio clip with the content of the picture and the main idea of the text to be conveyed.

Ideally, one picture, one audio clip and one paragraph of text should add up to exactly S_{Slide} bytes. However, most of times, we need to perform some trade-off in order to achieve this. We now introduce the optimal synthesis techniques for video MMS and web-page MMS.

4.1 Optimized MMS synthesis from a video

The pictures are first ordered based on the time-line. We then add in the audio such that it is synchronized with the picture. If text exists (which does not in many cases), it is used as the title or as a description. We split the time line into many segments, each of which can contain a maximum of S_{Slide} bytes inclusive of a picture, an audio clip and some text. The segments are then saved in the MMS format. The constraint for each MMS slide is:

$$S_{audio} + S_{pic} = S_{Slide} \tag{13}$$

For synchronization in MMS synthesis, we would like that there is only one picture and one audio clip within the S_{Slide} limit in the gallery. However, in most cases, the total size exceeds S_{Slide} bytes. Thus, we need to explore trade-offs. If the key frames of the motion pictures are selected $M'_P = \{P'_i, i = 1,2\cdots,k\}$ from $M_P = \{P_i, i = 1,2\cdots,k\}$ at time $T_P = \cup_{i=1}^{k-1}[T_i, T_{i+1}]$, and the corresponding audio track is $\Omega = \{\omega_i, i = 1,2,\cdots,l\}$, where $l > 0$ is the duration in seconds, then $\forall P'_i \in M'_P$, the duration of the candidate audio clip will be within $R_i = [T_i - \Delta T_i,\ T_i + \Delta T_i]$, where ΔT_i is the half duration of $S_{Slide} - S_{pic}$ audio data, namely:

$$\Delta T_i = \frac{S_{Slide} - |P'_i|}{2 \times L} \tag{14}$$

where $|P'_i|$ is the size of the i^{th} picture; L is the audio frame length in Eq. (8). The potential problem is $T_i - \Delta T_i < 0$ or $T_i + \Delta T_i > l$. In this case, the audio duration can be adjusted

as: $R_i = [T_i - \Delta T_i - \delta T_i,\ T_i + \Delta T_i + \delta T_i]$ or $R_i = [T_i - \Delta T_i + \delta T_i,\ T_i + \Delta T_i - \delta T_i]$, where $0 \leq \delta T_i \leq \Delta T_i$, $T_i - \Delta T_i - \delta T_i \geq 0$, $T_i + \Delta T_i - \delta T_i \geq 0$ and $T_i - \Delta T_i + \delta T_i \leq l$, $T_i + \Delta T_i + \delta T_i \leq l$. It is best if $\cup_{i=1}^{k} R_i = T_P$ and $\cap R_i = \emptyset$.

However, this is for an ideal situation. For audio track $\Omega = \{\omega_i, i = 1, 2, \cdots, l\}$, $l > 0$, after audio simplification, we get $\Omega' = \{\omega'_i, i = 1, 2, \cdots, l'\}$, the corresponding duration for each simplified audio clip ω'_i is R'_i. If $T_A = \cup_{i=1}^{l'} R'_i$ - $\cup_{i=1}^{k} R_i \neq \emptyset$, namely $T_A = \cup_{i=1}^{k'} [t_i, t_{i+1}], 0 < k' < k$, then we have to select the highly attended frames from $m_P = M_P - M'_P = \{p_i, i = 1, 2, \cdots, q\}$ for the audio clip $\Omega' - \{\omega(t), \omega \in \Omega, t \in R_i\}$ for the period T_A as given by Eq. (13).

Input : Motion pictures M_P and audio Ω from a video
Output : The synthesized MMS slide sequence O

Procedure:
1. $step = 0$;
2. $\forall P'_i \in M'_P$, obtain the audio clip $A_i = \{\omega(t), t \in R_i\}$;
3. $O_{step} = \{(P'_i \oplus A_i)\}$, $step++$;
4. If $T_A = \emptyset$, exit;
5. $\forall R'_i \in T_A$, get the important pictures p_i from $p_i \in M_P - M'_P$, $i = 1, 2, \cdots, q$ subject to Eq. (13).
6. Calculate the linkage between p_i and $a_i = \{\omega(t), t \in r_i\}$, noted as $O_{step} = \{p'_i \oplus a_i, i = 1, 2, \cdots, q\}$, $step++$;
7. Goto 4;
8. Output $O = \cup_{i=0}^{step} O_{step}$;
9. $O = \{S_i : (P'_i \oplus A_i), i = 0, 1, \cdots, step - 1\}$

Algorithm 2: Algorithm for MMS synthesis for a video

In algorithm 2, we use motion pictures simplification to obtain the most important regions of a video sequence. All pictures are cropped to the size of 352×288 pixels in the JPEG format. If the region size is less than that, there is no change in the size. Our observation is that the average size of a JPEG image with resolution of 352×288 is around 100KB. For the audio clips, we follow the rules of MMS synthesis in Eq. (13) to locate the corresponding important pictures within the duration. Actually, we could have adopted this approach at the beginning of the process itself. However it potentially can lead to picture-audio starting at the beginning of the slide sequence. The algorithm 2 can overcome this to provide a superior result. We record the timing information in script files for MMS synchronization. In the audio script file, the simplified audio durations within the MMS constraints are fused while the locations of the key frames (for the pictures track) are provided in the picture script file to obtain the suitable audio. The optimized result is saved in the MMS script file for MMS synthesis.

4.2 Optimized MMS synthesis from a web page

To compose a MMS with the data from a web page, we should simplify the pictures and put them on the time-line. For the text, we put the simplified text based on the requirements of the user. Normally, the MMS from a web page does not have audio. For the optimization

problem in MMS composition from web pages, the main content will be the text with the display of the most important paragraphs. The rest of the space will be for the simplified pictures. A user can fully utilize the quota to provide pictures with the best quality. Since only one picture is allowed for each MMS slide, the pictures can be sequenced based on the decreasing saliency. Thus, the constraint for each MMS slide is:

$$S_{txt} + S_{pic} = S_{Slide} \tag{15}$$

5. EXPERIMENTAL RESULTS

In this paper, we provide two examples of composing a MMS based on video and other information from a media gallery. One is a video for soccer game and the other is a video for family care. We list the user survey at the rear of this section.

5.1 Scenario 1: Soccer Video

Soccer fans are interested in the latest updates on the crucial ongoing matches, such as the matches for the FIFA world cup. Soccer video simplification for MMS synthesis provides a solution to the soccer fans who are not able to watch the live-cast of these games. This application can keep them updated of the progress of the game with well-organized pictures, audio and comments in text.

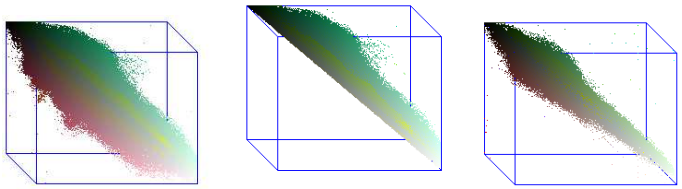
The detection and filtering of the important events in live broadcast video programs and adapting the incoming streams based on the content importance and user preference are mentioned in [Xu et al. 2001]. An algorithm for parsing the structure of produced soccer programs is proposed in [Xie et al. 2002]. A fully automatic and computationally efficient framework for analysis and summarization of soccer videos using cinematic and object-based features is available at [Ekin et al. 2003]. These techniques could be used in soccer video simplification and automated MMS synthesis.

For soccer video simplification, we input one clip into the MMS composing system. The audio channel is extracted first and its length is dependent on the selection of the importance degree. For each audio clip, its resolution and redundancy are reduced to the minimum quality according to the degree required. The cropped pictures having the designated width and height will be put on the time line according to the requirements. Then the corresponding texts from web pages, which are simplified by experiential sampling, will be put on the time line if the corresponding web page can be found. Each piece of upto S_{Slide} bytes will be packaged as a MMS file and sent to the end user cellphone.

In soccer video simplification, we remove the less salient video frames by using the field color (green) as the first cue. We assume if a frame includes enough grass color, it is likely to be a soccer match, or else that clip should be deleted. The grass field color is automatically learned by training. We compute the statistics for amount of the green grass color, and the dominant color in a frame. Then those frames under the given threshold will be removed. Fig. 8 shows some examples after non-grass color removal.

We then analyze the picture and motion details in order to remove the still frames. We only keep the frames that have both vivid color and strong motion. Thus a large number of frames will be discarded from the sequence. Furthermore, we use dynamic cropping to reduce the frame size so that the frame is suitable for transfer over the wireless channel and for playing on the small screen.

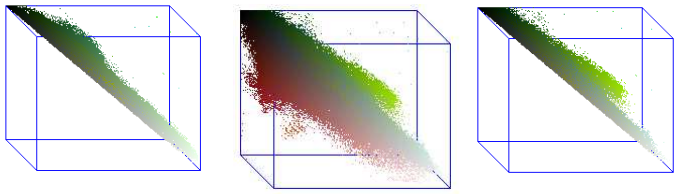
Here we provide a group of results for soccer video simplification in Table III. Column



(a) Case 1: Color distribution before non-grass color based frame removal in a soccer video

(b) Case 1: Color distribution after non-grass color based frame removal in a soccer video

(c) Case 2: Color distribution before non-grass color based frame removal in a soccer video



(d) Case 2: Color distribution after non-grass color based frame removal in a soccer video

(e) Case 3: Color distribution before non-grass color based frame removal in a soccer video

(f) Case 3: Color distribution after non-grass color based frame removal in a soccer video

Fig. 8. Color distribution before and after non-grass color based frame removal in a soccer video

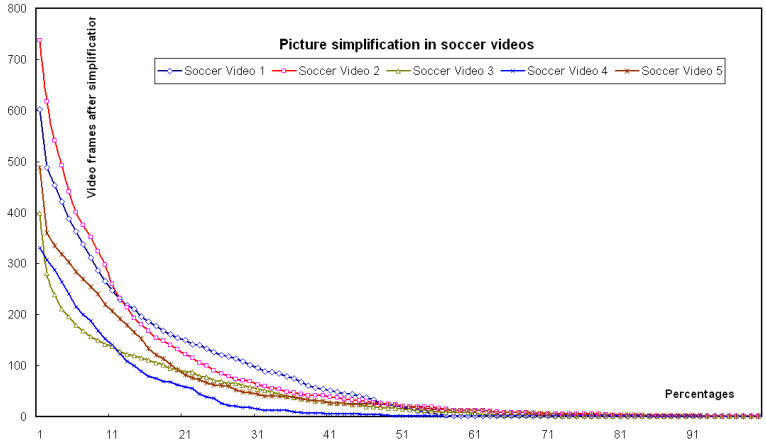
one lists the video names and their size, the corresponding frame numbers and the event numbers of these videos are shown in column two, column three describes the frame numbers and event numbers after motion based video simplification, column four provides the event numbers and frame numbers after grass color based redundancy reduction.

Table III. Frames after video redundancy reduction

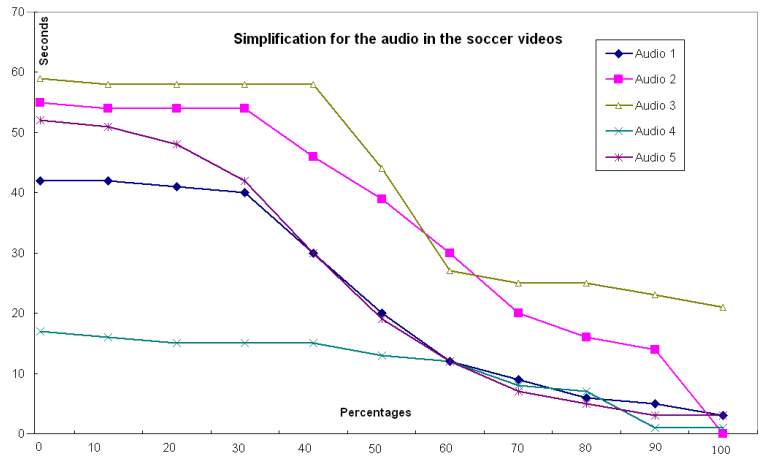
Video Name	Length (Events)	Motion (Events)	Color (Events)
1.avi(40.6MB)	1081(4)	635(4)	529(4)
2.avi(52.6MB)	1399(6)	787(4)	476(4)
3.avi(16.8MB)	1480(8)	562(5)	274(3)
4.avi(5.7MB)	448(4)	332(3)	228(3)
5.avi(16.0MB)	1318(6)	747(5)	333(3)

Figure 9 shows the results after the simplification. The relationship between frame numbers and the saliency (in percentage) is shown in Fig. 9(a); The relationship between the audio length and the saliency degrees is shown in Fig. 9(b). From Fig. 9(a), we can see the reduction procedure of frame numbers at 10%∼20% are nonlinear, other places are almost

linear from 0% to 10% with a steep decline and from 20% to 100% with a slow decline. In Fig. 9(b), the most steep range is from 30% to 60%.



(a) The relationship between picture numbers and saliency degree after pictures simplification for the soccer video



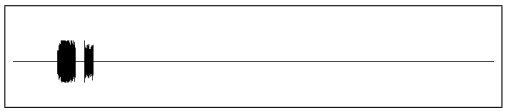
(b) The relationship between audio length and saliency after audio simplification for the soccer video

Fig. 9. Results of soccer video simplification

We list the results for this soccer video simplification at saliency of 90% in Fig. 10. Fig. 10(a) is the result after extracting the video frames and Fig. 10(b) is the corresponding simplified audio from the video clip. This video is the second goal of Chinese football team at 2002 FIFA World Cup qualifier match on August 15, 2001 at Shenyang city, China. The report on this video:"...Li turned provider on twenty minutes, floating in a delightful

cross from wide on the right to Qi Hong, who rose above two defenders to head China's second past Humaid Juma Rashid into the top right-hand corner of the net..."
(See: http://www.sinosoc.com/news/index.asp?id=665).

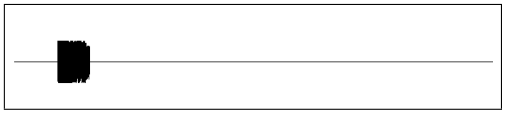
(a) Motion pictures simplified from a soccer video frames (>90%)



(b) Audio simplification from a soccer video (>90%)

Fig. 10. The simplification for the soccer video

(a) The picture corresponding to the highest portion of the audio track

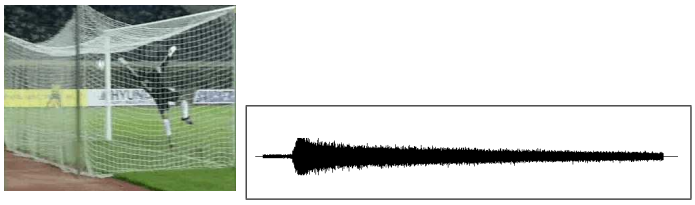


(b) The highest portion of the audio track in the soccer video

Fig. 11. The simplification to the audio track and its corresponding picture.

Figure 11 provides the important part of the audio track, the important picture on this duration is shown in Fig. 11(a), the corresponding audio clip is shown in Fig. 11(b).

Figure 12 is the MMS package including audio and picture that is sent to the receiver. Only one picture shown in Fig. 12(a) will be inserted into the MMS slide and this picture is saved in the JPEG format needing only 86KB (352×288). We can use the entire audio track for the audio and visual matching since the audio track will consume only 84.2KB (43 seconds) in MP3 (MPEG Layer-3) format. If we add the web description in the MMS, the 44 words require only 222 bytes. Therefore, the total size of this MMS slide will be 170.4KB. This is a rich package of information for 43 seconds.



(a) The only picture extracted from a soccer video at 317 frame (or at 12.20seconds).

(b) The audio clip from the soccer video

Fig. 12. MMS synthesis for the soccer video

5.2 Scenario 2: Family Care

In this scenario, the sensors installed at home will transfer the visual and audio signals to a home server. Only the relevant simplified information is integrated and transferred to the home server. Once an alarm is triggered, the home server sends the relevant data to the MMS server. The server processes the multimedia data and sends the information to the mobile device. On the mobile device, the important information will be shown. If the mobile users would like to know in detail about the alarm, he can request the MMS server to send more information.

Here we show a motion picture sequence for infant care. For the motion pictures corresponding to Fig. 13(a), we obtain two pictures having high saliency among the sequence. The obtained pictures in Fig. 13(b) are useful in monitoring the baby's actions especially to check for face laceration with their fingernails. The two important pictures appear at frame numbers 198 and 286 from the total of 366 pictures. Fig. 13(c), Fig. 13(d), Fig. 13(e) and Fig. 13(f) are the results after simplification at important degrees of 60%, 65%, 70% and 75% respectively.

5.3 Results of the User Study

We conducted a user study to evaluate the efficacy of our approach. We polled 16 subjects who are familiar with SMS usage but not MMS usage. The survey is available at *http://www.comp.nus.edu.sg/∼mohan/survey/*. The survey involved providing feedback on text simplification, picture simplification, audio simplification and two automatically composed MMS messages.

For text simplification, we asked the subjects to judge the quality of the abstracted result of the 'SARS' example in section 3.2. 66.66% of the users could correctly infer the year of the story. For image simplification, 50% of the subjects could correctly select the important region for the example in Fig. 2 while 62.5% correctly got the example in Fig. 3. Note that the presentation order in the survey was random. For audio simplification, all except one could correctly order the quality of simplified audio clips in Fig.6. Half of the subjects found the simplified audio to be satisfactory.

We give users five choices ('bad','not bad','good', 'wonderful','excellent') to evaluate the quality of the two synthesized MMS messages. The first MMS shown in Fig.9 judged to be 'not bad' by 37.5% of the subjects. For the second MMS in Fig.13, 37.5% of the subjects found it to be 'wonderful'. The combined average score for all participants for

(a) Motion pictures for the new born baby.

(b) Two important pictures among the picture sequence (>95%).

(c) Three important pictures among the picture sequence (>70%).

(d) Four important pictures among the picture sequence (>65%).

(e) Five important pictures among the picture sequence (>60%).

(f) Seven important pictures among the picture sequence at 8 ,66, 130, 269, 276, 198, 286 frames respectively(>55%).

Fig. 13. The simplified motion pictures for baby care.

both messages is 'good'. Though the study is preliminary in nature, it is encouraging.

6. CONCLUSIONS

In this paper, we have proposed novel approaches for multimedia simplification and optimized synthesis of MMS messages for mobile phones. We use experiential sampling based techniques for multiple types of media such as text, images, video and audio. We select the semantically most significant data within the constraints of a MMS slide. The simplified

multimedia data is then composed into a MMS presentation adhering to the mobile communication requirements. Finally, the composed multimedia message comprising different types of media will be displayed on user's mobile device in a proper way. This transcoding technique is tremendously useful for conveying salient multimedia information on small form-factor mobile devices.

7. ACKNOWLEDGMENTS

We thank Wei-Gang Fu of NUS and Jun Wang of the Delft University of Technology for their comments. We also thank the participants of our user study. We deeply appreciate the most constructive suggestions of the anonymous reviewers.

REFERENCES

ATREY, P. AND KANKANHALLI, M. S. 2004. Probability fusion for correlated multimedia streams. In *Proc. of ACM Multimedia'04*. New York, USA, 408–411.

BENNETT, E. P. AND MCMILLAN, L. 2003. Proscenium: a framework for spatio-temporal video editing. In *Proc. of ACM Multimedia'03*. Bekeley, USA, 177–184.

BOLL, S., KLAS, W., AND WANDEL, J. 1999. A cross-media adaptation strategy for multimedia presentations. In *Proc. of ACM Multimedia'99*. Orlando, Florida, United States, 37–46.

CHANG, S.-F. AND VETRO, A. 2005. Video adaptation: Concepts, technologies, and open issues. *Proc. of the IEEE, Special Issue on Advances in Video Coding and Delivery 93*, 1 (Jan.), 148–158.

CHEVALLET, J.-P., LIM, J.-H., AND VASUDHA, R. 2005. Snaptotell: A singapore image test bed for ubiquitous information access from camera. In *Proc. of ECIR'05*. Santiago de Compostela, Spain, 530–532.

CHU, W.-T. AND CHEN, H.-Y. 2002. Cross-media correlation: a case study of navigated hypermedia documents. In *Proc. of ACM Multimedia'02*. Juan Les Pins, France, 57–66.

COCHRAN, W. G. 1977. *Sampling techniques(3rd Edition)*. John Wiley, New York, USA.

COULOMBE, S. AND GRASSEL, G. 2004. Multimedia adaptation for the multimedia messaging service. *IEEE Communications 42*, 7 (July), 120–126.

DAVIS, M., KING, S., GOOD, N., AND SARVAS, R. 2004. From context to content: leveraging context to infer media metadata. In *Proc. of ACM Multimedia'04*. New York, USA, 188–195.

DEMENTHON, D., KOBLA, V., AND DOERMANN, D. 1998. Video summarization by curve simplification. In *Proc. of ACM Multimedia'98*. Bristol, United Kingdom, 211–218.

DUDA, R. O., HART, P. E., AND STORK, D. G. 2000. *Pattern classification (2nd edition)*. Johan Wiley and Sons, Inc., New York, USA.

DUYGULU, P., NG, D., PAPERNICK, N., AND WACTLAR, H. 2003. Linking visual and textual data on video. In *Proc. of Workshop on Multimedia Contents in Digital Libraries'03*. Crete, Greece.

DUYGULU, P. AND WACTLAR, H. D. 2003. Associating video frames with text. In *Proc. of ACM SIGIR'03*. Toronto, Canada.

EKIN, A., TEKALP, A., AND MEHROTRA, R. 2003. Automatic soccer video analysis and summarization. *IEEE Transactions on Image Processing 12*, 7 (Jul.), 796–807.

GAO, S., LEE, C.-H., AND ZHU, Y.-W. 2004. An unsupervised learning approach to musical event detection. In *Proc. of IEEE ICME'04*. Taipei, Taiwan, 1307–1310.

JACOBS, O. 1993. *Introduction to Control Theory (2nd Edition)*. Oxford University Press, Oxford, UK.

JUN, S., RONG, Y., PEI, S., AND SONG, S. 2003. Interactive multimedia messaging service platform. In *Proc. of ACM Multimedia'03*. Berkeley, USA, 464–465.

KANKANHALLI, M. S., WANG, J., AND JAIN, R. 2006. Experiential sampling in multimedia systems. *IEEE Transactions on Multimedia (to appear)*.

LIM, J.-H., CHEVALLET, J.-P., AND VASUDHA, R. 2004. Snaptotell: ubiquitous information access from camera. In *Proc. of Internatioanl Workshop Mobile HCI'04*. Glasgow, Scotland, 21–27.

LIN, C.-Y. AND TSENG, B. L. 2005. Optimizing user expectations for video semantic filtering and abstraction. In *Proc. of IEEE ISCAS'05*. Kobe Japan.

MEDIONI, G., COHEN, I., BREMOND, F., HONGENG, S., AND NEVATIA, R. 2001. Event detection and analysis from video streams. *IEEE PAMI 23,* 8 (Aug.), 873–889.

PAN, J.-Y., YANG, H.-J., FALOUTSOS, C., AND DUYGULU, P. 2004. Automatic multimedia cross-modal correlation discovery. In *Proc. of ACM SIGKDD'04*. 653–658.

ROSENFELD, A. 1969. *Picture Processing by Computer*. Academic Press, New York, USA.

WANG, J., KANKANHALLI, M. S., YAN, W.-Q., AND JAIN, R. 2003. Experiental sampling for video surveillance. In *Proc. of The first Workshop on Video Surveillance, ACM Multimedia'03*. Berkeley, USA, 319–322.

WANG, J., REINDERS, M. J., LAGENDIJK, R. L., LINDENBERG, J., AND KANKANHALLI, M. S. 2004. Video content representation on tiny devices. In *Proc. of IEEE ICME'04*. Taipei, 84–89.

WELCH, G. AND BISHOP, G. 2001. An introduction to the Kalman Filter. In *SIGGRAPH'01 course 8*. Los Angeles, CA, USA.

WICKRAMASURIYA, J., DATT, M., MEHROTRA, S., AND VENKATASUBRAMANIAN, N. 2004. Privacy protecting data collection in media spaces. In *Proc. of ACM Multimedia'04*. New York, USA, 48–55.

XIE, L., CHANG, S.-F., DIVAKARAN, A., AND SUN, H. 2002. Structure analysis of soccer video with hidden markov models. In *Proc. of IEEE ICASSP'02*. Vol. 4. 4096–4099.

XIE, L.-X., KENNEDY, L., CHANG, S.-F., LIN, C.-Y., DIVAKARAN, A., AND SUN, H.-F. 2004. Discovering meaningful multimedia patterns with audio-visual concepts and associated text. In *Proc. of IEEE ICIP'04*. Singapore.

XU, P., XIE, L., CHANG, S.-F., DIVAKARAN, A., VETRO, A., AND SUN, H. 2001. Algorithms and system for segmentation and structure analysis in soccer video. In *Proc. of ICME'01*. Tokyo, Japan, 721 – 724.

YAN, W.-Q. AND KANKANHALLI, M. S. 2002. Detection and removal of lighting and shaking artifacts in home videos. In *Proc. of ACM Multimedia'02*. Juan Les Pins, France, 107–116.

ZHANG, H.-J., KANKANHALLI, A., AND SMOLIAR, S. W. 1993. Automatic partitioning of full-motion video. *ACM Multimedia Systems 1,* 1 (Jun.), 10–28.

This manuscript was submitted to the ACM TOMCCAP on 25 Jun. 2005; revised on 15 Dec. 2005, accepted on 17 Mar. 2006.