VIDEO CONTENT REPRESENTATION ON TINY DEVICES

Jun Wang[1], Marcel J.T. Reinders[1], Reginald L. Lagendijk[1], Jasper Lindenberg[2], Mohan S. Kankanhalli[3]

[1]Department of Mediamatics, Faculty of EEMCS, Delft University of Technology, The Netherlands
{j.wang, m.j.t.reinders, r.l.lagendijk}@ewi.tudelft.nl
[2]TNO Human Factors Research Institute, The Netherlands
lindenberg@tm.tno.nl
[3]Department of Computer Science, School of Computing, National University of Singapore, Singapore
mohan@comp.nus.edu.sg

ABSTRACT

The perceptual satisfaction of a user watching video on a tiny mobile device is constrained by the display capability and network bandwidth. To maximize the user's perceptual satisfaction in this constrained environment, we propose a new method to adaptively represent the video content in real-time on tiny devices according to the user's attention. In our framework, firstly, a sampling based dynamic attention model is proposed to obtain and maintain the user's attention in the video streams. Secondly, based on the most attended regions and sequences extracted, the attention based representation is introduced to achieve a higher perceptual satisfaction on a small display. Experiments with users show the effectiveness of our proposed method in a video surveillance application domain.

1. INTRODUCTION

Adaptive multimedia content delivery for universal access has been actively explored for some time [3]. Recently, due to the increasing capability of multimedia processing on handheld devices such as PDAs, hand phones, *etc.*, multimedia information can be accessed by those mobile devices. However, limitations on the displays size, the bandwidth, and the processing capability of the mobile devices prevent that one can represent multimedia information in the same way as is done for normal displays such as PCs, TVs, *etc*. Therefore, for multimedia presentation on tiny devices, it is non-trivial to develop an automatic content analysis method to represent the content according to the situated environment in order to maximize the users' perceptual satisfaction under the constrained resources.

To this end, many efforts have been put on multimedia content adaptation. For example, MPEG [4] allows to describe multimedia content in different scales to be able to adapt to various clients. Recently, image based attention models [5,6,8] are proposed, and they have been used to generate display paths on small displays when the original size is larger than that of the display device [5,7].

In a surveillance domain, however, display adaptation is essential to be done in real-time when one wants to make use of mobile devices in this setting. In this paper, we , therefore, propose a new method to re-create video content in real-time for small mobile devices. This is achieved by exploiting the user's attention, so that irrelevant information can be discarded. By doing so, the perceptual quality is improved and additionally the amount of data to be transmitted is reduced in a more efficient manner.

2. ATTENTION BASED VIDEO CONTENT REPRESENTATION

Our framework is illustrated in Fig. 1(a). Firstly, based on the user's attention, a sampling based dynamic attention model is defined. Secondly, by applying the dynamic attention model, "sequences of interest" (SOIs) and "regions of interest" (ROIs) are obtained from the video stream. Finally, based on the obtained SOIs and ROIs, the adapted stream is created and transmitted to the users. In order to reduce the data to be transmitted, the temporal representation is shown in Fig. 1(b). When a frame is part of a SOI, it is transmitted fully (frame rate is equal to the original rate) while in the cases where a frame is not part of a SOI, the frame rate is decreased or frames are transmitted only upon requests of a user. In order to improve the user's perceptual satisfaction, we applied the spatial representation as shown in Fig. 1(c). The ROIs in the frame are zoomed in and rendered on the small displays instead of the entire frame.

The remainder of this section will be structured as follows: Sec. 3.1 introduces the sampling based dynamic attention model to obtain the ROIs and SOIs. Sec. 3.2 introduces the attention based representation from the obtained ROIs and SOIs.

2.1 A sampling based dynamic video attention model

Video data are spatio-temporal data. User attention comes from both the *spatial* and *temporal* directions. Past work on the extraction of visual attention uses the saliency map representation to denote the visual attention in an image [5,6,8]. Those image based attention models fail to model the *temporal* correlations of the attention and thus are not

suitable to maintain the dynamics of the attention in video sequences.

Based on the *Sequential Importance Sampling* (SIS) algorithm [9], we introduce a sampling based dynamic video attention model to obtain and maintain the dynamic attention in both *spatial* and *temporal* directions.

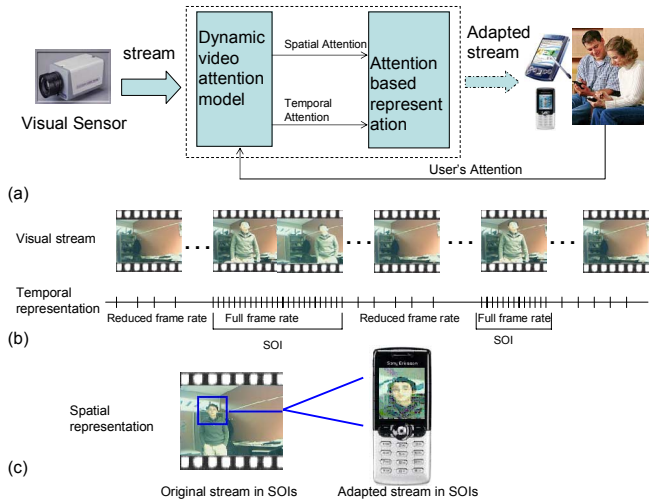


Fig. 1 Real-time attention based video content representation for tiny devices. (a) Our framework, (b) The temporal representation (c) The spatial representation.

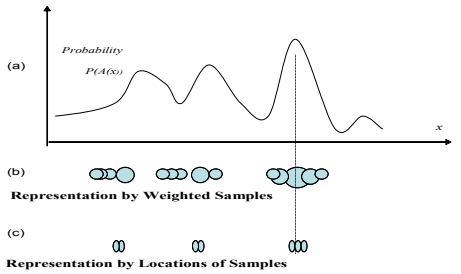


Fig. 2 The representation of the distribution.(a) Continuous representation, (b) A set of weighted samples (dot indicates the location of the sample and the size of the dot indicates the weight), (c) Only the locations of a set of samples.

2.1.1. Sampling based approach

We use samples to represent the attention [1,2]. There are two ways of representing the attention by using samples, either by using weighted samples (Fig. 2(b)), or, by using only the location of the samples (Fig.2 (c).

Representation by weighted samples: The attention is maintained by the weighted samples as shown in Fig.2(b). It means the distribution of the attention is captured by both weights and the location of the samples (Fig.3(c,g,k)).

Representation by samples: The attention is only maintained by the location of the samples (Fig.2 (c)). This method does not require the weights to represent the distribution of the attention. This representation is used for obtaining the ROIs (Fig.3(d,h,l)). However, it needs to be generated by re-sampling the weighted samples.

Different from the image based methods, the sampling method provides a flexible representation of the attention with minimal assumptions. The number of samples employed can be adjusted to achieve a balance between the accuracy of the approximation and the computation load. Moreover, it is easy to incorporate this representation within a dynamical system which can model the temporal continuity of attention (shown in Fig. 3), if we consider each sample as a particle and each particle having its own dynamics.

2.1.2. Our dynamic attention framework

Fig. 3 shows an overview of our sampling based dynamic attention model. In this section, we discuss the different concepts in our framework.

Spatial attention Attention samples (AS) are used to maintain the spatial distribution of the attention. The attention samples are defined as;

$$AS(t)[i],\quad i=[0,N_A(t)] \tag{1}$$

where $AS(t)$ is the set of attention samples at a given time t. $N_A(t)$ is the number of attention samples at time t and $AS(t)[i]$ is the ith attention samples at time t.

The elements of $AS(t)[i]$ are defined as;

$$AS(t)[i].x,\ AS(t)[i].y,\ AS(t)[i].w \tag{2}$$

where x is the position in the x direction, y is the position in the y direction of an image plane, and, w is the weight of the attention sample $AS(t)[i]$.

The amount of the attention in one particular position is measured by the weights of the attention samples $AS(t)[i].w$. Depending on the application, the attention can be measured by different features [1,2]. For instance, motion is an important activity that needs to be monitored in the surveillance domain. The motion attention can be obtained by updating the weights as follows:

$$AS(t)[i].w = \left| I_t(AS(t)[i].x, AS(t)[i].y) - I_{t-1}(AS(t)[i].x, AS(t)[i].y) \right| \tag{3}$$

where the weight is the absolute difference of corresponding pixel intensity values of two neighboring frames $I_t(x,y)$and $I_{t-1}(x,y)$. The procedure of updating is shown in Fig. 3 (For instance, from (*f*) to (*g*)).

Besides the attention samples, we also use sensor samples (SS) denoted as : $SS(t)[i]$, $i=\{0, N_S\}$. The sensor samples are uniformly spread over the image and are used to probe the image [1] whether there is a shift in attention (see the white dots in Fig.3(b,f and j)). The definition is the same as shown in the equation (1) and (2).

Temporal attention In a video sequence, there is a varying amount of attention over time. To measure the amount of the attention, we introduce the concept of attention saturation denoted as $ASat(t)$.

In this paper, $ASat(t)$ is measured by the average of the weights of the entire set of sensor samples for a given time window slice. Thus, $ASat(t)$ is defined as:

$$ASat(t) = f_N\left(\frac{1}{n}\sum_{q=[t-n,t]}\left(\frac{1}{N_S}\sum_{i=1}^{N_S} SS(t)[i].w\right)\right) \tag{4}$$

where f_N is the mapping function which is used to normalize the value into the range [0,1], and, N_S is the number of sensor samples. n denotes the temporal neighborhood. The purpose of averaging n number of recent sensor sample sets is to suppress noise and maintain temporal continuity.

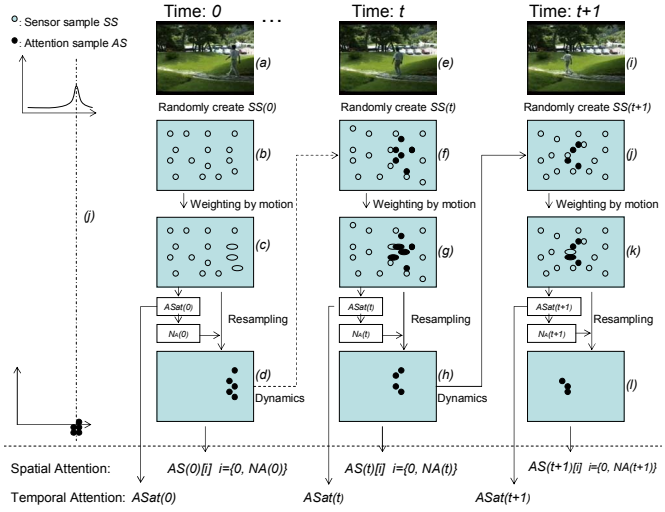


Fig. 3 Our sampling based dynamic video attention model. (*a,e,i*) Sample frames. (*b,f,j*) Randomly created *SSs* and dynamical-updated *ASs* (only in *f&j*) from previous time slice. (*c,g.k*) weighted *SSs* and *ASs* (only in *g&k*). (*d,h,j*) N_A number of the *ASs*. (*j*) *ASs* in (*d*) represent the attention distribution.

Fig. 3 illustrates our dynamic attention model finds the attention in a pedestrian monitoring scenario. For each time instance, the model outputs the temporal attention ($ASat(t)$) and spatial attention ($AS(t)$). The procedure in detail goes as follows:

Step 1. The *SSs* are randomly created to sense the entire scene in (*b*), (*f*) and (*j*). If the ASs exist in the previous time slice, they are dynamically updated (Step 4) to the current time slice like the dark dots shown in (*f*)and (*j*). Note that time 0 indicates the situation when there are no previous *ASs, i.e.* start of the system or a sequence of interest.

Step 2. The weights of the *SSs* and the *ASs* (in (*c*), (*g*) and (*k*)) are adjusted by the measurements from the motion features (by using Eq. (*3*)). The weights are indicated by the size of the points in (*c*), (*g*) and (*k*).

Step 3. The temporal attention ($ASat(t)$) is calculated from the weight-adjusted *SSs*. Consequently, $AS(t)$ is created by resampling the *SSs* (shown from (*c*) to (*d*)) or both the *SSs* and *ASs* (shown from (*g*) to (*h*) and from (*k*) to (*l*), respectively). The number of the created $AS(t)$s is controlled by the obtained $ASat(t)$. For more detail, we refer to [1,2]. Obviously, those $AS(t)$s created represent the distribution of the spatial attention as shown in (*d*), (*h*) and (*l*).

Step 4. Each *AS* follows its own dynamics (for instance, constant velocity) and is dynamically updated to the next time slice (dark dots shown from (*h*) to (*j*)).

Now we introduce the extraction of the region of interest and sequence of interest from those samples.

2.1.3. Region/Sequence of interest extraction

Region of interest (ROI): A region of interest represents the area in the image at time t that has a high attention of the user. The set of ROIs at time t is denoted as $R(t)$. There are $N_R(t)$ number of ROIs at a given time instance t. Thus, $R(t) \subset AS(t)$.

There are several ways to generate those ROIs. One straight forward way is to select those attention samples that have a high enough weight, i.e. thresholding the $AS(t).w$. Clustering method then can be used to find out different (spatial) clusters among the selected attention samples.

Alternatively, one can evaluate the attention samples by another measurement of the user's attention. For instance, in the video surveillance domain, the attention samples can be evaluated using a face detector at the position of the sample. Then each of the attention samples that are positively evaluated are combined, to generate a region of interest (Fig.6).

Sequence of Interest (SOI): The sequence of interest represents frames where there is a high attention of the user. It is defined as

$$S(t) = \{s_i\} \quad i = [0, N_S(t)], s_i = [t^S{}_i, t^E{}_t] \tag{5}$$

where $S(t)$ is the total set of SOIs so far, and $N_S(t)$ is the number of the SOIs so far. s_i is an individual SOI. A SOI includes all the ROIs ($R(t)$) from the beginning of the sequence (*i.e.* time slice $t^S{}_i$) to the end of the sequence (*i.e.* time slice $t^E{}_i$).

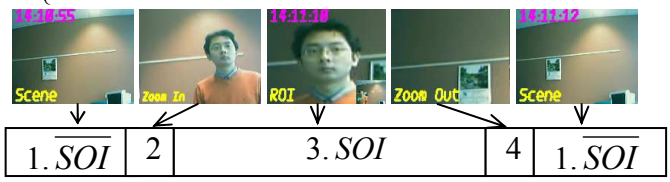
The method to obtain the SOIs from the $ASat(t)$ is straightforward. As shown in Equation (6), the frame whose $ASat(t)$ is bigger than some threshold T is considered to belong to the sequence of interest.

$$F(t) = \begin{cases} SOI & ASat(t) > T \\ \overline{SOI} & otherwise \end{cases} \tag{6}$$

where $F(t)$ indicates whether frame at time t belongs to a SOI (SOI) or not ($\overline{SOI}$).

If a frame at time t belongs to SOI i, we use the following equation to find whether it is the starting time $t^S{}_i$, ending time $t^E{}_i$ or other time of the SOI $t^M{}_i$.

$$t = \begin{cases} t_i^S & F(t-1) == \overline{SOI} \\ t_i^E & F(t+1) == \overline{SOI} \\ t_i^M & otherwise \end{cases} \tag{7}$$



Mode: 1: Scene (reduce frame rate); 2. Zoom in; 3: Focus ; 4. Zoom out

Fig. 4 Attention based representation.

2.2. Attention based representation

After we obtain the SOIs and ROIs, the next step is how to represent those attended content to the user in a perceptually pleasant manner. For simplicity, we did only consider one ROI for a given moment in time.

By combining the temporal and spatial representation as shown in Fig.1(b,c), the final attention based representation is illustrated in Fig.4. In order to smooth the transitions between the scene ($\overline{SOI}$) and the focus mode (SOI), we make use of a pseudo zoom in /out between the two modes.

In order to reduce the ambiguity in the focus mode, the overall scene is also scaled down and shown in the bottom-right region of the frame.

3. EXPERIMENTS

We applied our representation to a video surveillance scenario. We captured video streams (320x240 sized frames) in a lab environment and obtained attention based representation (reduced to 160x120 sized frames) by using the proposed method in Sec.2. Our method performs in real-time. Some example frames are shown in Fig. 6.

SOIs are estimated by utilizing Eq. 6. The estimated SOIs *v.s.* a ground truth (which is identified by manually labeling the frames) is illustrated in Fig.5. When we reduce the frame rate of the $\overline{SOI}$ sequences (as shown in Fig. 1(b)), the data to be transmitted can be extremely reduced without losing the interest of the users.

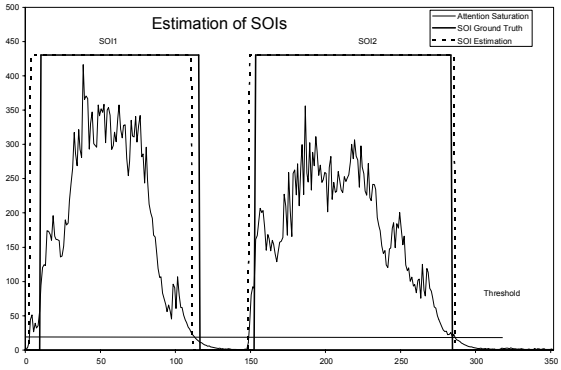
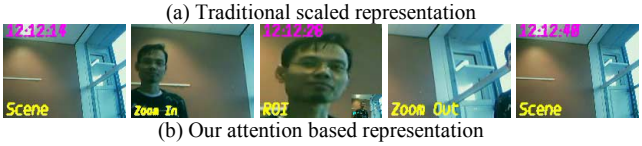


Fig. 5 Estimation of SOIs by thresholding the attention saturation.

Group (1)

(a) Traditional scaled representation



(b) Our attention based representation

Group (2)

(c) Traditional scaled representation

(d) Our attention based representation

Fig. 6 Sample frames of the test video clips for the user study.

Next, we carried out a subjective evaluation. We compared our representation (Fig. 6.(b,d)) with the traditional scaled one(down scaling (to 160x120) only in Fig.6(a,c)). Each one of the 17 participants was asked to view two sets of the video streams and to complete a number of questionnaires to evaluate our method.

The initial results are encouraging: 71% of the users judges that the zoom in/out action is done logically, and, 100% of the users judges that it improves the surveillance tasks, *i.e.* it improves the recognition of the person. 65% of the users answered that our attention based representation could help to improve the visual satisfaction while compared to the traditional scaled one.

4. CONCLUSION

In this paper, we proposed an attention based representation of video content on tiny devices. Experimental results show that our method can reduce the amount of the visual data to be transmitted and improves the perceptual satisfaction in a video surveillance domain given the constrained display size of the mobile device. More complicated representations will be investigated in the future to deal with the multiple attention regions in one frame.

5. REFERENCES

[1] J. Wang, and M. S. Kankanhalli, "Experiential Sampling for Multimedia Analysis," *In Proc. of ACM Multimedia 2003*, Berkeley, November 2003.

[2] J. Wang, M. S. Kankanhalli, W-Q. Yan, and R. Jain, "Experiential Sampling for Video Surveillance," In *Proc. 1st ACM Int. Workshop on Video Surveillance*, Berkeley, November 2003.

[3] P. Brusilovsky, "Adaptive Hypermedia," *User Modeling and User-Adapted Interaction* 11: 87-110, Kluwer Academic Publishers, Netherlands 2001.

[4] ISO/IEC JTC1/SC29/WG11/N4242 (2001) ISO/IEC 15938-5 FDIS Information Technology – Multimedia Content Description Interface – Part 5 Multimedia Description Schemes, Sydney, Australia, July 2001.

[5] L. Chen, X. Xie, X. Fan W. Ma, H. Zhang, and H. Zhou., "A visual attention model for adapting images on small displays," Microsoft Research Asia, 2002.

[6] Y-F. Ma and H-J. Zhang, "Contrast-based Image Attention Analysis by Using Fuzzy Growing," *In Proc. of ACM Multimedia 2003*, Berkeley, November 2003.

[7] X. Fan, X. Xie, H. Zhou and W. Ma, "Looking into Video Frames on Small Displays," *In Proc. of ACM Multimedia 2003*, Berkeley, November 2003.

[8] L. Itti, and C. Koch, "Computational Modeling of Visual Attention," *Nature Reviews Neuroscience*, Vol. 2, No. 3, pp. 194-203, Mar 2001.

[9] A. Doucet, S. J.Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statist. Comp.*, 10:197-208, 2000.