

Looking into Video Frames on Small Displays

Xin Fan^{*1}, Xing Xie², He-Qin Zhou¹, Wei-Ying Ma²
Department of Automation, University of Science and Technology of China¹
Hefei, 230027, P.R.China

van@mail.ustc.edu.cn, hqzhou@ustc.edu.cn

Microsoft Research Asia²
5F, Sigma Building, No.49, Zhichun Road. Beijing, 100080, P.R.China
{xingx, wyma}@microsoft.com

ABSTRACT

With the growing popularity of personal digital assistants and smart phones, people have become enthusiastic to watch videos through these mobile devices. However, a crucial challenge is to provide a better user experience for browsing videos on the limited and heterogeneous screen sizes. In this paper, we present a novel approach which allows users to overcome the display constraints by zooming into video frames while browsing. An automatic approach for detecting the focus regions is introduced to minimize the amount of user interaction. In order to improve the quality of output stream, virtual camera control is employed in the system. Preliminary evaluation shows that this approach is an effective way for video browsing on small displays.

Categories and Subject Descriptors

I.4.9 [Image Processing and Computer Vision]: Applications

General Terms

Algorithms, Human Factors

Keywords

Video adaptation, form factor, mobile device, virtual camera control, adaptive content delivery

1. INTRODUCTION

With the growing popularity of personal digital assistants and smart phones, people have become enthusiastic to watch videos through these mobile devices. Though a few commercial video players such as Windows Media Player and PocketTV have been developed to enable users to browse videos from the small-form-factor devices, the limited bandwidth and small window sizes remain to be two critical obstacles.

Although there have been many approaches for adapting videos,

^{*} This work was conducted while the first author was a visiting student at Microsoft Research Asia.

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

MM'03, November 2-8, 2003, Berkeley, California, USA.

Copyright 2003 ACM 1-58113-000-0/00/0000...\$5.00.

most of them only focused on compressing and caching. In this paper, we will focus on the display constraints, since excessive resolution or quality reduction will cause degradation of the user perception. Currently the main resources of mobile videos are traditional well-edited clips, but we notice that real-time programs or spontaneous video clips such as home videos, surveillance videos will become more and more popular [4]. For this kind of video, it is more possible and also in a great demand to optimize the contents for different display conditions.

In our system, the users are allowed to zoom into a set of focus regions, for instance, a region containing a human face or a text paragraph which can hardly be recognized after down-sampling. The focus regions will be displayed using a higher resolution, therefore, be better perceived. A user interface similar to a typical image viewer is provided to navigate among different regions in video frames.

However, this approach is not very convenient since the users have to frequently interact with the system. In order to minimize the amount of user interaction, an automatic approach for detecting the focus regions is introduced in Section 2. It is based on our previous work on image adaptation [1][2][5]. In [1][2][5], instead of treating an image as a whole, we manipulate each region-of-interest in the image separately, which allows delivery of the most important region to the client when the screen size is small.

Previous results on image adaptation can be extended to video adaptation if we simply consider each video frame as an image. A similar idea appears in [8], but their process is fully manual and how to compose the cropped images into an acceptable video stream is not presented. In order to improve the quality of output stream, we employ virtual camera control in the system, which will be presented in Section 3.

2. VISUAL ATTENTION MODELING

In this section, we define a visual attention model to reveal the regions most probably attracting the user's attention in a video frame. Automatic generation of the attention model is presented afterwards.

2.1 Visual Attention Model

As shown in Definition 1, in our model, a set of information carriers – *attention objects* (AOs) are defined, which usually represent semantic objects, such as a human face, a flower or a text annotation.

Definition 1: $\{AO_i\} = \{(ROI_i, AV_i, MPS_i)\}, \quad 1 \leq i \leq N \quad (1)$

Where ROI is referred in our model as a spatial region corresponding to an AO . ROI can be in arbitrary shapes and allows to overlap. Referring to human cognitive systems, the AO s carry different amounts of information, i.e., are of different importance. The attention value (AV) indicates the weight of each AO in contribution to the information contained in the image. Since the delivery of information is significantly dependent on the area of presentation, minimal perceptible size (MPS) is introduced as a threshold to avoid excessively sub-sampling during the reduction of display size.

2.2 Automatic Attention-based Modeling

Four types of attention objects are taken into account in our model: motion objects, face objects, saliency objects and text objects.

Different from static pictures, the moving part in a scene is usually noticeable and it has been considered in many video transcoding systems. In our implementation, video sequences are stored as MPEG format and the motion information is measured by a similar method to [6] where foreground motion can be effectively discriminated from camera motion or background motion. The AV of motion object is estimated by its size, spatial/temporal coherence and motion intensity. It is supposed that an object with larger magnitude, greater motion intensity or more consistent motion will be more important:

$$AV_{motion} = Area_{motion} \times W_{motion}^{intensity} \times W_{motion}^{coherence} \quad (2)$$

Early stages of attention processing are deployed by ensemble of low-level features such as contrast, orientation, and intensity etc. Due to the heavy computations of the traditional saliency model, we adopt a contrast-based model [7] to produce the saliency map and determine the attention-getting areas. An example image and its saliency map are shown in Figure 1.



Figure 1. An example of saliency detection.

The AV s of saliency objects are calculated as:

$$AV_{saliency} = \sum_{(i,j \in R)} B_{i,j} \cdot W_{saliency}^{i,j} \quad (3)$$

where $B_{i,j}$ denotes the value of pixel (i,j) in the saliency map. Since people often pay more attention to the region near the center of an image, a normalized Gaussian template centered at the image is used to assign the position weight $W_{saliency}^{i,j}$.

Some heuristic rules are employed to calculate the MPS of above two types of AO s. For example, bigger regions can be scaled down more aggressively than smaller ones. Furthermore, the MPS can be automatically adjusted according to the user browsing history.

Face objects and text objects are defined and generated in the same way as our prior work [1]. In order to combine different

types of AO s into a unified attention model, the AV of each AO is normalized to $(0,1)$ and the final AV is computed as:

$$AV_i = w_k \cdot AV_i^k / \sum_i AV_i^k \quad (4)$$

where AV_i^k represents the AV of AO_i detected in model k and w_k is the weight of model k , e.g. face model, text model or motion model, which manifests the contribution during the attention-guiding function. In our system, motion objects are considered most attention-getting and semantic objects play a more important role than saliency objects. w_k can also be automatically adjusted according to user browsing behavior.

3. VIRTUAL CAMERA CONTROL

Based on the visual attention model, we can easily find out where the focus regions are in a video frame according to different constraints. Before going to the discussion of generating focus regions, we first introduce how to combine them into a smooth video stream.

A straightforward approach is to directly present the focus regions to the user. However, this naïve approach will cause jitters in the video sequences since the frames will be discontinuous after cropping. We notice that in some tracking systems [9], virtual camera control is introduced to smooth the tracking process. Therefore, we assume that there is also a virtual video camera steered to pan and zoom in the video frames. Two types of focuses are introduced here: Camera Focus (CF) and Target Focus (TF). Camera Focus stands for the focus displayed to the users and Target Focus is the destination focus either manually assigned or automatically determined. Corresponding display ratios of the two types of focus regions are defined as Camera Ratio (CR) and Target Ratio (TR), respectively. The Euclidean distance between CF and TF is denoted as Δd and the difference of CR and TR is defined as Δr . The direct focus shifting is substituted for a smooth following process from the current focus region to the target focus region, with a set of pan and zoom operations. In our system, the virtual camera is in one of following three statuses:

- *Fixation status:* If both Δd and Δr are very small, the virtual camera will be fixed in order to avoid unpleasant dithers in the video stream.
- *Following status:* When either Δd or Δr is larger than a predefined threshold, we will let the virtual camera smoothly follow the new target focus. In moving object tracking, IIR filtering is often used to smooth temporally the spatial derivatives. In our algorithm, a recursive implementation of the second order filter is adopted:

$$C(k) = \alpha_1 C(k-1) + \alpha_2 (T(k) + T(k-1)) \quad (5)$$

where $\alpha_1 + \alpha_2 / 2 = 1$, $\alpha_1, \alpha_2 > 0$, $C(k)$ is either the position of CF or CR at time k and $T(k)$ is the corresponding position of TF or TR.

- *Shifting status:* When either Δd or Δr is less than a threshold, we will directly steer the virtual camera to the new target region in order to eliminate the lag as a result of the IIR filter:

$$C(k) = T(k) \quad (6)$$

In addition, the virtual camera will also come to shifting status when a shot boundary occurs, which is mainly because that we consider the contents are not related between two different shots.

4. SYSTEM FRAMEWORK

A friendly user interface is critical to facilitate the video browsing on mobile devices since the input device is very limited. Three browsing modes are proposed in our system: manual, semi-automatic and full-automatic. They provide three levels of automation and users can choose the one he most prefers. Virtual camera control is employed in all three modes to ameliorate the resulting video stream. The framework of our system is illustrated in Figure 2.

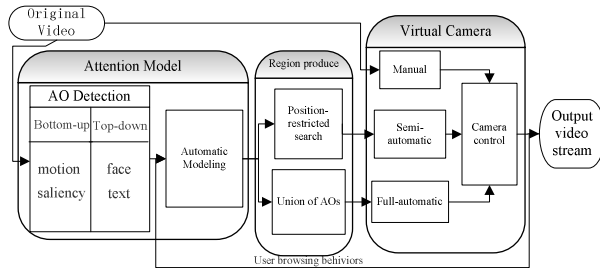


Figure 2. The framework of our system.

4.1 User Interface

Three types of user controls are introduced: direction, zoom and mode switch, as shown in Figure 3 (a). The first two are defined similar to a typical image viewer and the mode switch control is used to change video browsing modes which will be introduced in Section 4.2.

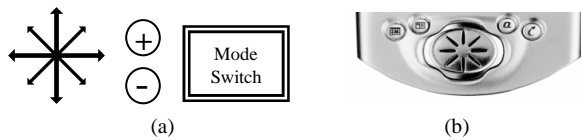


Figure 3. User interface of our system.

They can be easily implemented on current mobile devices. For example, in Figure 3 (b), we can map the buttons on a Pocket PC into these user controls. More advanced interface can employ a set of sensors instead of direct inputs [3].

4.2 Browsing Modes

As to our experience, it's difficult and annoying to locate and zoom into the attractive regions in a fully manual way. Based on the visual attention model, three browsing modes and corresponding user interface are designed to deal with different user scenarios.

4.2.1 Manual browsing mode

This is the basic browsing mode. Both direction and zoom controls are enabled in this mode. Users can scroll to a desirable position and zoom into the detail by hand. A picture-in-picture view is employed to help locating the focus regions. When the user clicks on the thumbnail image, a camera focus shifting will start to change the current focus to the destination focus. As mentioned before, this approach causes large amounts of user interaction which degrade user perception.

4.2.2 Full-automatic browsing mode

For some types of videos, there may be only a small number of AOs in the image, e.g., a surveillance video. In this case, we propose an automatic browsing approach where the MPS is ignored and the target region is defined as the union of all attention objects detected. Both direction and zoom controls are disabled in this mode. The resulting video stream uses more screen space to display the attention-getting regions while cropping out the other parts. However, when video frames contain many separate focuses, this approach will have less difference with the original down-sampling scheme.

4.2.3 Semi-automatic browsing mode

Different from the full-automatic browsing mode, the most attractive region produced by our previous algorithm [1] will be first presented to the user when entering this mode. It is implemented as searching the optimal region which contains as many perceptible attention objects as possible under the display constraint. Here "perceptible" stands for that the display area of an attention object is larger than its MPS.

Users can use the direction controls to scroll to the other parts of the video frame. The duration or times of press on direction buttons denote the relative distance between the current focus and the desirable region. An approximate position of the target focus is calculated and a "restricting rectangle" is defined which covers all possible target focuses. A position-restricted searching is then performed to find the new focus and the virtual camera will be automatically shifted to it. This scheme facilitates the locating of focus regions and only slight adjustment is needed afterwards.

In the position-restricted searching algorithm, the AOs included or partial included in the restricting rectangle should be displayed in the output stream. If no such AO exists, the restricting rectangle itself is looked as an AO. A branch and bound searching [1] with these AOs as its root is executed to find the optimal solution, as shown in Figure 4.

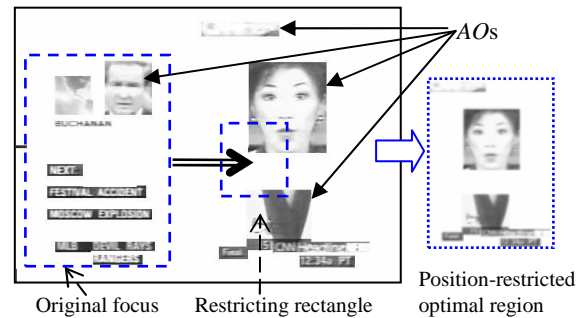


Figure 4. An example of position-restricted searching.

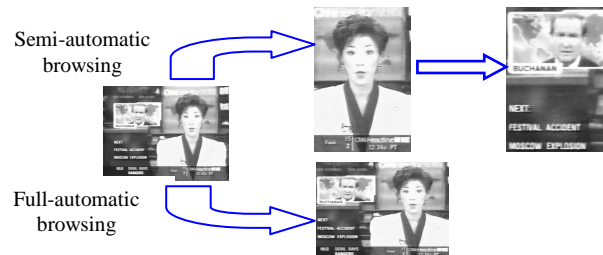


Figure 5. Comparison of full-automatic browsing mode and semi-automatic browsing mode.

A comparison of the full-automatic browsing mode and the semi-automatic browsing mode is illustrated in Figure 5. From the figure, we can see that in this case, the semi-automatic browsing mode is better since there are multiple focuses.

4.3 Utilization of User Behavior

It has been long discovered that user behavior contains valuable human wisdom and mining this knowledge could greatly help tuning human-computer interaction. As mentioned in Section 2.2, some parameters in the visual attention model, such as *MPS* and the weights of different models can be automatically adapted to the user browsing behavior. We adjust these parameters when a new target focus is reached and the virtual camera is in a fixation status. The weight w_k corresponding to the *AOs* of model k within the display area will be increased according to a centered normalized Gaussian template.

$$w_k' = \max_j \{\mu_{k,j}\} \cdot w_k \quad (7)$$

where w_k' denotes the new weight for model k and $\mu_{k,j}$ is the position weight of the j th *AO* of model k in the Gaussian template. The *MPS* of *AOs* of model k will also be increased according to the zoom ratio of current displayed objects:

$$MPS_{k,j}' = \lambda_1 \cdot MPS_{k,j} + \lambda_2 \cdot Ratio_{k,j} \quad (8)$$

where $\lambda_1 + \lambda_2 = 1$. $MPS_{k,j}'$ denotes the new *MPS* of the j th *AO* of model k and $Ratio_{k,j}$ is the zoom ratio of the *AO*.

5. IMPLEMENTATION

Currently we have implemented a prototype video browser on desktop PC based on Microsoft DirectShow. All the attention detection modules including face detection and saliency detection can be performed in real-time, thus, users will not feel any delay during the browsing. However, due to the limited capabilities, it will be difficult to directly port the detection modules to mobile devices. We plan to setup a proxy which accepts the client's requests and process the video streams for them according to their display sizes.

Since there is no agreed-upon ground truth or benchmark available and proper evaluation requires subjective results from a significant number of users, we only invite six volunteers to try our prototype and provide their judgments on its performance. We chose nine clips from a video library as our testing data, including home videos, surveillance videos and news videos. Two subjective assessment questions were given to them:

1. *Does the output result under full-automatic browsing mode improve the browsing experiences? (Improved, No difference, Worse)*
2. *Do you think the idea of using a virtual camera is acceptable for small displays? (Acceptable, Not acceptable)*

Preliminary feedbacks are very encouraging. Averagely, it was considered that majority (62.9%) of output results under full-automatic browsing mode had improved the browsing experience. In most cases (81.5%), the participants had the opinion that zooming and panning in the original video was an acceptable

solution for the limited display sizes. We also found that better feedback was received for surveillance videos and home videos, which aligns with our initial motivations. More user study experiments will be carried out in the future.

6. CONCLUSIONS

In this paper, we proposed a novel solution for video browsing on mobile devices. Most existing work on video transcoding focused on reducing bandwidth cost and the limited display size is of deficient consideration. In our approach, three browsing modes and corresponding user interfaces are proposed to preserve the attended information with high quality. A visual attention model is utilized to detect the most informative regions and virtual camera control is employed to improve the output video stream. Our approach can be easily applied to real-time video transcoding proxies, though porting to client devices still need a lot of efforts. In our future work, we plan to employ more rules from cinematography to improve the results of virtual camera control.

7. ACKNOWLEDGEMENTS

We would like to express our special appreciation to Tao Mei, Xiaodong Gu, Liquan Chen and Yusuo Hu for their insightful suggestions. We also thank all the voluntary participants in our user study experiments.

8. REFERENCES

- [1] L.Q. Chen, X. Xie, X. Fan, W.Y. Ma, H.J. Zhang, and H.Q. Zhou, A visual attention model for adapting images on small displays, *ACM Multimedia Systems Journal*, to appear.
- [2] X. Fan, X. Xie, W.Y. Ma, H.J. Zhang, and H.Q. Zhou, Visual attention based image browsing on mobile devices, *Proc. of ICME 2003, Vol. I*, p53-56, Baltimore, USA, Jul. 2003.
- [3] K. Hinckley, J. Pierce, M. Sinclair, and E. Horvitz, Sensing techniques for mobile interaction, *ACM UIST 2000, San Diego, USA*, p91-100, Nov. 2000.
- [4] R. Jain, Mobile multimedia, *IEEE Multimedia*, Vol.8, No. 3, p1-1, Jul. 2001.
- [5] H. Liu, X. Xie, W.Y. Ma, and H.J. Zhang, Automatic browsing of large pictures on mobile devices, *ACM Multimedia 2003, Berkeley, CA, USA*, to appear.
- [6] Y.F. Ma and H.J. Zhang, A model of motion attention for video skimming, *IEEE ICIP 2002, New York, USA*, p129-132, Sep. 2002.
- [7] Y.F. Ma and H.J. Zhang, Contrast-based image attention analysis by using fuzzy growing, *ACM Multimedia 2003, Berkeley, CA, USA*, to appear.
- [8] K.B. Shimoga, Region of interest based video Image transcoding for heterogeneous client displays, *Packet Video 2002, Pittsburgh, USA*, Apr. 2002.
- [9] X. Sun, J. Foote, D. Kimber, and B.S. Manjunath, Panoramic video capturing and compressed domain virtual camera control, *ACM Multimedia 2001, Ottawa, Canada*, p229-238, Sep. 2001.