Adaptive Video Presentation for Small Display While Maximize Visual Information

Yandong Guo, Xiaodong Gu, Zhibo Chen, Quqing Chen, and Charles Wang

Thomson Corporate Research, Beijing
eemars@gmail.com, xiao-dong.gu@thomson.net

Abstract. In this paper we focus our attention on solving the contradiction that it is more and more popular to watch videos through mobile devices and there is an explosive growth of mobile devices with multimedia applications but the display sizes of mobile devices are limited and heterogeneous. We present an intact and generic framework to adapt video presentation (AVP). A novel method for choosing the optimal cropped region is introduced to minimize the information loss over adapting video presentation. In order to ameliorate the output stream, we make use of a group of filters for tracking, smoothing and virtual camera controlling. Experiments indicate that our approach is able to achieve satisfactory results and has obvious superiority especially when the display size is pretty small.

Keywords: Adaptive video presentation, mobile device, optimal cropped region, maximize visual information, Kalman filter, and virtual camera control.

1 Introduction

It becomes more and more popular to watch videos through mobile devices and there is an explosive growth of mobile devices with multimedia applications. Unfortunately, there are two obstacles to browse videos on mobile devices: the limited bandwidth and the small display sizes of mobile devices. Thanks to the development of network, hardware and software, the bandwidth factor is expected to be less constraint while the limitation on display size remains unchangeable in the foreseeable future.

If sub-sampling each frame according to the resolution of the output device while preserving the intact video contexts, the excessive reduction ratio will lead to an ugly experience. It will be a good solution if cropping the most important part, called the region of interest (ROI), from the original video, discarding partial surroundings and then resizing the cropped region to the display size of the output device. How to get the optimal cropped region (OCR) is the key technique in the solution.

There are several approaches used for browsing videos on mobile devices by cropping region of interest [4-6]. In [4], the authors proposed a semi-automatic solution for this problem. The method proposed in [5] is focused on the technique of virtual camera control. And the approach in [6] is for special scenario of static panoramic capturing. However, the perceptual result is affected by the display

G. Qiu et al. (Eds.): VISUAL 2007, LNCS 4781, pp. 322–332, 2007.

resolution but none of these solutions has considered providing optimal cropped region (OCR) according to the display sizes, which means maximize the viewer received information.

In this paper, we presented an intact self-adaptive solution which can be used to all kinds of videos and can be adapted to various display sizes. Our main contribution is to propose a novel algorithm which can get OCR adaptively according to the display size while minimizing the information loss. Furthermore, we improved the visual camera control by adding zooming in/out operations when necessary.

The paper is organized as follows. In the following section, all the components of the system are explained briefly. The OCR choosing is described in Section 3 while the tracking and filtering is shown in Section 4. In Section 5, the virtual camera control is presented and the last section concludes this paper.

2 System Architecture

A complete framework of our approach is shown in Fig.1. First of all, we analyze the input sequence to extract the attention objects (AO), e.g., human faces, balls, texts or other saliency objects. These attention objects are divided into two groups according to the models by which we analyze the sequences: saliency attention objects by saliency models (bottom-up process) and semantic attention objects by semantic models (top-down process).

We adopt the Itti's model [1] as the saliency model to produce the saliency map and use the model mentioned in [10] as the semantic model to detect the saliency attention objects. Since "The purpose of the saliency map is to represent the conspicuity or the 'saliency'-at every location in the visual field by a scalar quantity and to guide the selection of attended locations, based on the spatial distribution of saliency" [1], we can get the conclusion that human attention is an effective and efficient mechanism for information prioritizing and filtering. Moreover, supposing that an object with larger magnitude will carry more information, we can calculate the information carried by the *i*-th saliency object as:

$$Infor_i^{saliency} = \sum_{x,y \in R} I_{x,y} \,. \tag{1}$$

Where $I_{x,y}$ denotes the scalar quantity of the pixel (x, y) in the saliency map, R denotes the region of the *i*-th object.

By employing the top-down model [2], we obtain the position, the region and the quantity of information of the semantic attention objects. In the following part of paper, we take football as an example of semantic attention objects, and use the trajectory-based ball detection and tracking algorithms mentioned in [3] to locate the ball. The information carried by semantic attention objects can be calculated as:

$$Infor_i^{semantic} = W_i \times area_i^{semantic} \,. \tag{2}$$

Where $area_i$ denotes the magnitude of the semantic attention degree and W_i is used to unify the attention models by giving a weight to each kind of semantic attention objects.

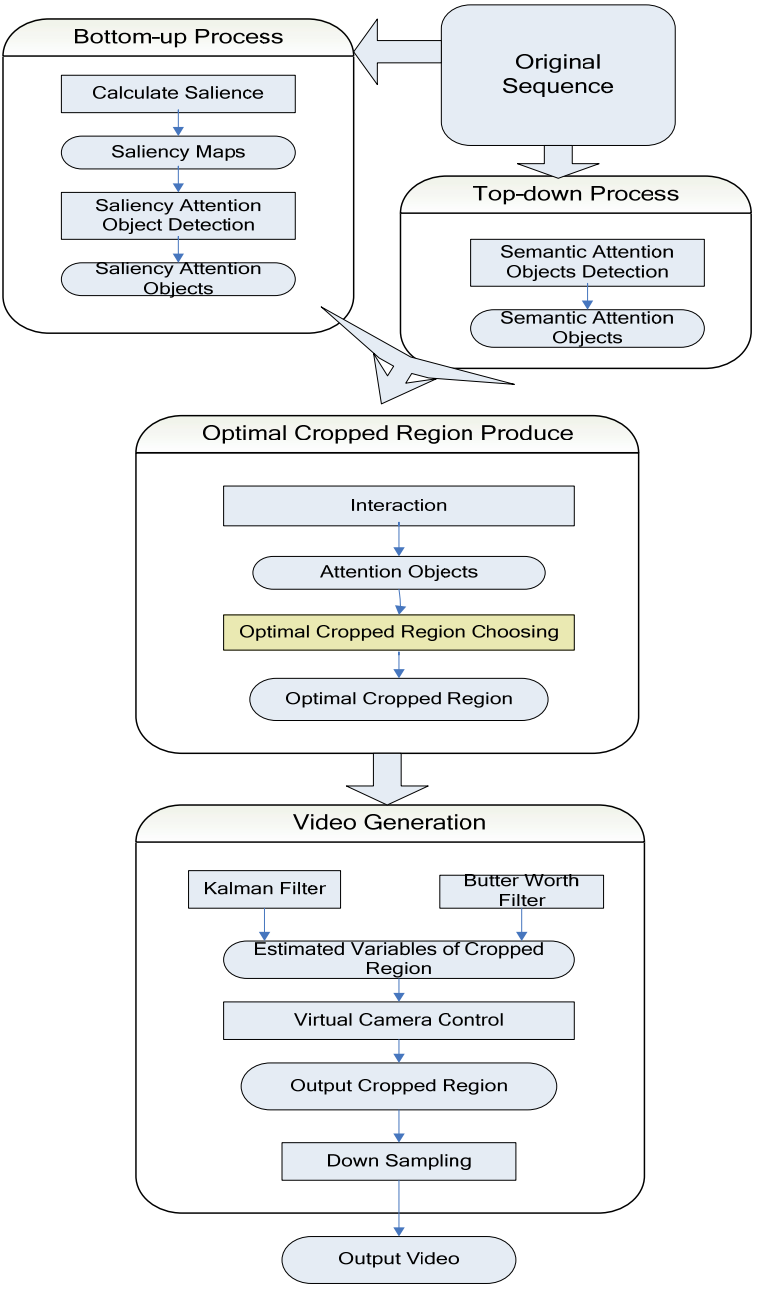
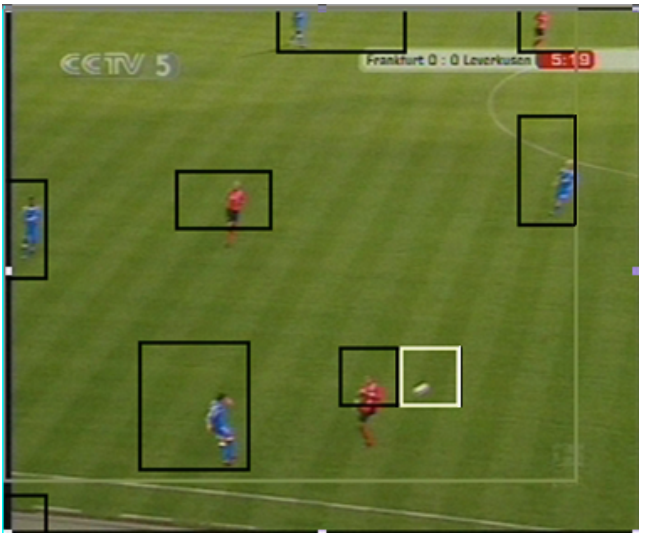


Fig. 1. The System Architecture

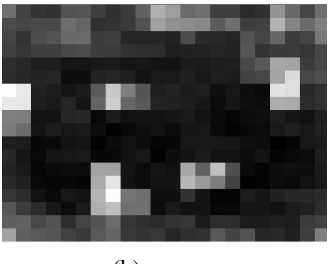
The saliency model by Itti determines the saliency objects by the nonlinear integration of low-level visual cues, mimicking processing in primate occipital and

posterior parietal cortex. It makes no assumption on video content and thus is universally applicable. By making use of the saliency model our approach is able to deal with various videos. Moreover, we amend the results of the saliency model by adopting the semantic model, because the semantic model can provide more accurate location of attention objects. If a semantic object and a saliency object cover similar region (judged by the threshold) the region got by semantic model will be chosen for the attention object.

The saliency attention objects and the semantic objects are integrated to get a uniform attention model. An example frame is shown in Fig. 2, the saliency attention objects are marked by the black rectangles and the semantic attention objects are marked by the white rectangle.



(a)



(b)

(c)

Fig. 2. (a) is the frame in which the attention objects were marked; (b) is the saliency map of the frame calculated by Itti's model and (c) is the map to show the saliency attention objects

As shown in Fig. 2, we usually get more than one AO. If we draw the cropped region including all AOs, we have to zoom out the cropped region at very large ratio

when the display resolution is very low compared with the original resolution. However, when the display resolution is not very low compared with the original resolution, larger cropped region including more AOs may be the better choice. For the reason above, the cropped region should be adapted to different display sizes to preserve as much information as possible by trade-off between cropping and reduction.

The process of choosing OCR is a noisy process. We assume that the noise is Gaussian so we use the Kalman Filter to estimate the central coordinates of OCR and use IIR to smooth the areas of OCR temporally. Kalman Filter and IIR reduce most of the noise inherent in the OCR choosing, and produce a new region sequence (EOCR). If the EOCR is used to down-sample directly, the quality of the output video is often jittery. The movement and zoom in/out of EOCR is less smooth than that of a physical camera, which has inertia due to its mass. Therefore, we use an additional filtering step called virtual camera control (VCC) to produce a smoother and more pleasing video output.

3 Optimal Cropped Region (OCR) Choosing

When there are several AOs dispersed in one frame, if we draw the cropped region including all the AOs, we may get a large cropped region. Down sampling such a large cropped region into display size will lead to an excessive reduction ratio and the display sequence will be blurred badly when the screen used for display is pretty small. Because there is loss of details during the process of image resizing which cannot be recovered afterwards and when the resize ratio becomes larger the information loss increases rapidly [8], outputting all the AOs may not always be the best choice and we have to abandon some of the AOs to get the largest information output.

Which ones of the AOs should be involved in the cropped region and which ones should not? This is the problem we will solve in this section. In another words, that is to say, how to keep balance between information loss from down sampling and that from abandoning AOs?

Definition 1. The optimal cropped region (OCR) is the region which can generate display sequence with the minimum information loss.

In the following part, we will discuss how to measure the information loss along the process of cropping and down sampling.

When we choose some of the AOs to form a cropped region, the sum of information actually got by viewers is,

$$InforSum^{cropped\_region} = \sum_{i \in CR} infor_i^{Saliency} + \sum_{j \in CR} infor_i^{Sematic}. \quad (3)$$

Where *CR* denotes the cropped region. Then, we will measure the information loss during the down sampling which is a fine to coarse image representation. In [9], Mario Ferraro considered the fine-to-coarse transformation of an image as an isolated irreversible thermodynamical system whose channels are dynamical subsystems. The information loss over the transformation is measured by P, the average of the density of entropy production across the image,

$$\sigma = \left(\frac{\nabla f(x,y,t)}{f(x,y,t)} \right)^2 . \qquad (4)$$

$$\mathrm{P} = \iint_{\Omega} f(x,y,t)\sigma(x,y,t)dxdy .$$

The density σ measures the local loss of information and t is a non-negative parameter that defines the scales of resolution at which the image is observed; small values of t correspond to fine scales (cropped region before down-sampling), while large values correspond to coarse scales (cropped region after down-sampling). (x, y) is used to denote the coordinates of a pixel in a discrete lattice.

In the case of image down-sampling, when the cropped region is scaled down from the original scale t_0 to the scale t_1, we use the operator T which takes the original image at one scale to another scale to give the transformation,

$$T_{downsample} : I(\square, t_0) \to I(\square, t_1) . \qquad (5)$$

In order to measure the information loss during the image down-scaling process, there must be a reference image to the original one. Therefore, we up-sample the down-scaled display image by the uniform algorithm to the resolution of cropped region at scale t_0 and choose it as the reference image:

$$T_{upsample} : I(\square, t_1) \to I'(\square, t_0) . \qquad (6)$$

Where $T_{downsample}$ and $T_{upsample}$ indicate the transformation of down-sampling and up-sampling, respectively.

We calculate the information loss ratio *InforLossRatio* over the transformation of $T_{downsample}$ and $T_{upsample}$ by (4).

With assumption that there is no information loss during the up-sampling process, we can get the remaining information of display image by,

$$InfoSum^{display} = (1 - InforLossRatio) \times \sum_{i \in CR} infor_i^{Saliency} + \sum_{i \in CR} infor_i^{Sematic} . \qquad (7)$$

Because the semantic attention objects do not lose its semantic meaning during the transformation, we do not multiply the information attenuation factor to them when they are larger than the minimal perceptible size which can be predefined according to the class of the objects.

We calculate the $InfoSum^{display}$ of display images got by cropped regions including different AOs and at last, we set the cropped region with the largest $InfoSum^{display}$ as the OCR and calculate the central coordinates (x_{OCR}, y_{OCR}) and area $Size_{OCR}$ of the optimal cropped region for the next use.

4 Tracking and Smoothing

After the OCR determination, we get a sequence of their central coordinates and a sequence of their size. The center of OCR tracking is generally a noisy process. We

assume that the noise is Gaussian Noise, and estimate the coordinates and velocities of the centers by Kalman Filter [7]. We describe the model of the discrete-time system by the pair of equation below:

$$\text{System Equation: } s(k)=\Phi s(k-1)+\Gamma w(k-1)\ . \tag{8}$$

$$\text{Measurement Equation: } z(k) = Hs(k) + n(k)\ . \tag{9}$$

Where $s(k) = [x(k), y(k), v_x(k), v_y(k)]^T$ is the state vector, $x(k)$ and $y(k)$ are the horizontal and vertical coordinates of the center of OCR at time k, respectively. $v_x(k)$ is the velocity in the horizontal direction and $v_y(k)$ is the velocity in the vertical direction. $w(k\text{-}1)$ is the Gaussian noise caused by choosing, representing the center acceleration in the horizontal and vertical directions.

And $z(k) = [x_{OCR}(k), y_{OCR}(k)]^T$ is the measurement vector at time k . $n(k)$ denotes the noise caused by measure and supposed to be Gaussian white noise.

The state transition matrix and coupling matrix

$$\Phi = \begin{pmatrix} 1 & 0 & T & 0 \\ 0 & 1 & 0 & T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \Gamma = \begin{pmatrix} \frac{T^2}{2} & 0 \\ 0 & \frac{T^2}{2} \\ T & 0 \\ 0 & T \end{pmatrix}$$

And the measurement matrix, $H = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$

The Kalman Filter estimates the coordinates and the velocities of the centers of OCRs from the results of OCR choosing. After the recursive procedure, we can get the estimated vector $\hat{s}(k) = [\hat{x}_{OCR}(k), \hat{y}_{OCR}(k), \hat{v}_x(k), \hat{v}_y(k)]$ with the minimum mean-squared error of estimation.

The area of OCR $Size_{OCR}$ may fluctuate intensively. We observed that the zooming of a physical camera is a more smooth process, so we use IIR filter to smooth the area of OCR temporally and the output of IIR is indicated by $\hat{Size}_{OCR}(k)$ [10].

5 Virtual Camera Control

If we use the tracking and smoothing results to move the cropped region directly, the quality of the output video is often jittery. The resulting motion is less smooth than that of a physical camera, which has inertia due to its mass. Virtual camera control (VCC) mentioned in [3] is used to solve the problem. The basic elements of VCC by

[3] are that the change of coordinates less than the threshold will be discarded and the coordinates are set to be constant, whereas the monotone continuous change of coordinates larger than the thresholod will be tracked and the coordinates are set to be changed smoothly. The VCC they mentioned is only used to control the centroid motion because the size of cropped region is constant in their application. However, the size of cropped regions can be changed in our algorithm and we ameliorate the VCC by adding zooming in/out function as one state of the state machine.

We use the $x_o(k), y_o(k), Size_o(k)$ to denote the central position and the size of region used to output at time *k*, respectively. When the inequalities

$$| Size_o(k-1) - \hat{Size}_{OCR}(k) | < \sigma_s$$

is satisfied for a certain constant σ_s, the sizes remain unchangeable.

$$Size_o(k) = Size_o(k-1) . \tag{10}$$

Otherwise, we start the zoom operation of virtual camera,

$$Size_o(k) = \alpha_1 Size_o(k-1) + \alpha_2 \hat{Size}_{OCR}(k) . \tag{11}$$

$$\alpha_1 + \alpha_2 = 1, \ \alpha_1, \alpha_2 > 0 \tag{12}$$

6 Experiment Results

We choose several video sequences, including sports match videos, news videos, home videos and surveillance videos. The normalized viewer received information

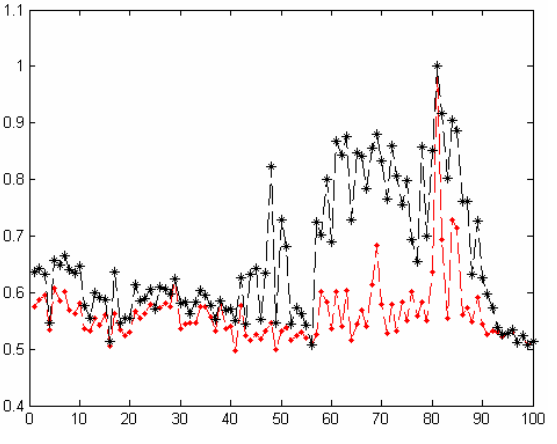


Fig. 3. The curve of the normalized viewer received information different cropped region
* With OCR choosing
•The cropped region with all attention objects

curve of a sequence about soccer match is shown in Fig. 3. The horizontal axis in the figure is the temporal axis while the vertical axis is used to indicate the normalized viewer received information. The figure indicates that the AVP with OCR choosing mechanism can output more visual information than AVP with the cropped region including all attention objects.

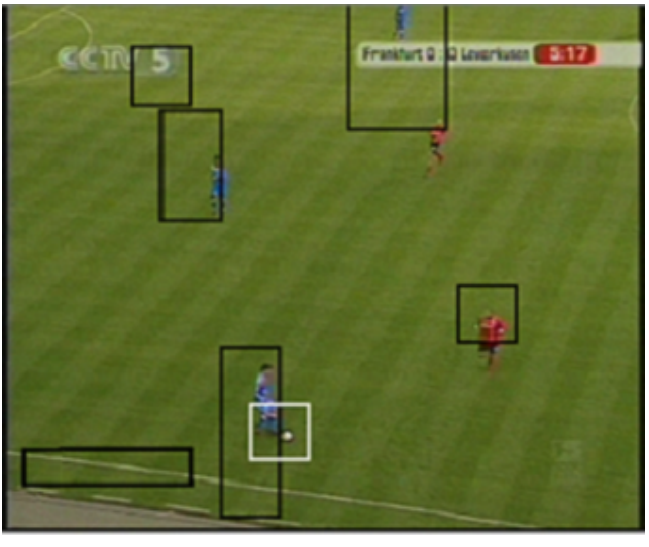


Fig. 4. The original frame with the attention objects marked. The saliency attention objects are marked in black rectangle while the semantic attention object is marked in white rectangle.

Since the evaluation of video presentation quantity is associated with the viewers' subjective feeling, we invited nine volunteers and showed the result sequences created by our approach, the sequences created by the framework without OCR choosing, and the sequences by down sampling the original sequences. We asked them to choose the sequence with the most satisfactory browsing experience. The feedback is encouraging and all sequences created by our approach got the highest evaluation.

There are the frames of the result sequences created by different approaches in different display sizes which are similar with the display sizes of real mobile devices as an example in figure 5. And their original frame with attention objects marked is shown in Fig. 4. The experiment results show that the larger the down-scaling ratio is, the larger the information loss ratio is, so the smaller display size usually leads to smaller cropped region. But there is no uniform function can describe the relationship between the down-scaling ratio and the information loss ratio precisely. We can see that the cropped regions in our approach are adjusted according to the display sizes of mobile devices to gain an optimal cropped region.

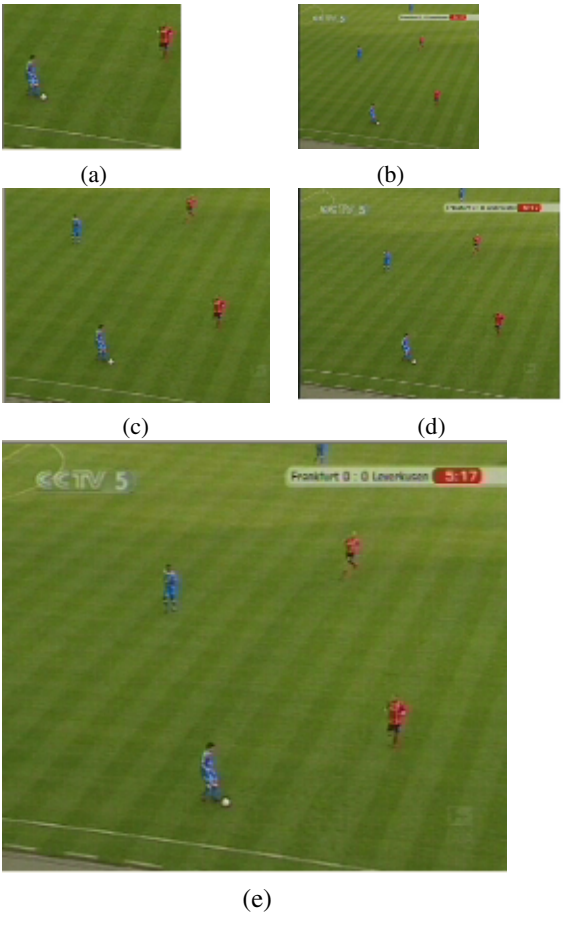


Fig. 5. (a), (c) are from the sequences created by our approach with the display size 88x72 and 132x108, respectively. (b), (d) are from the sequences created by the approach get cropped region including all the AOs, (without OCR choosing) with the display size 88x72 and 132x108, respectively. When the display size is 264x216, the approach with OCR choosing and the one without OCR choose gain the same cropped region, which is shown in (e).

7 Conclusion

In this paper we presented an intact framework for adapting video presentation. The framework integrates attention objects detection, optimal cropped region choosing, tracking and smoothing process and virtual camera control to output pleasant sequences. The whole process is automatic, robust and has obvious advantages when the display size is pretty small or when there are multi-attention-objects. This adaptive video presentation approach not only can be used for watch videos through mobile devices but also has potential use for transferring video bit streams through the network with limited bandwidth by transferring only saliency parts of the video. We

will do experiments to demonstrate this potential use of the approach. Moreover, in our future work, we plan to accelerate the optimal cropped region choosing process and employ more rules from cinematography to ameliorate the mechanism of virtual camera control.

References

1. Itti, L., Koch, C., Niebur, E.: A Model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on PAMI, 1254–1259 (1998)
2. Yu, X., Leong, H.W., Xu, C., Tian, Q.: Trajectory-Based Ball Detection and Tracking in Broadcast Soccer Video. IEEE Transactions on multimedia, 1164–1178 (2006)
3. Tovinkere, V., Qian, R.J.: Detecting semantic events in soccer games: Towards a complete solution. In: ICME, pp. 1040–1043 (2001)
4. Fan, X., Xie, X., Zhou, H.Q., Ma, W.Y.: Looking into video frames on small displays. In: MM 2003, November 2-8, pp. 247–250 (2003)
5. Aygun, R.S., Zhang, A.: Integrating virtual camera controls into digital video. In: Multimedia and Expo, ICME 2004, pp. 1503–1506 (2004)
6. Sun, X., Kimber, D., Manjunath, B.S.: Region of interest extraction and virtual camera control based on panoramic video capturing. IEEE Transactions on PAMI, 981–989 (2005)
7. Kay, S.M.: Fundamentals of statistical signal processing. Prentice-Hall, Englewood Cliffs (1993)
8. Cover, T.M., Thomas, J.A.: Elements of information theory. Rscience Publication (1991)
9. Ferraro, M., Boccigone, G.: Information properties in fine-to-coarse image transformations. In: Image Processing, ICIP 1998, pp. 757–761 (1998)
10. Mitra, S.K.: Digital signal processing. McGraw-Hill, New York (1993)
11. Ma, Y.F., Zhang, H.J.: Contrast-based image attention analysis by using fuzzy growing. In: MM 2003, Berkeley, California, USA, pp. 374–380 (2003)