Lu SP, Zhang SH. Saliency-based fidelity adaptation preprocessing for video coding. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 26(1): 195–202 Jan. 2011. DOI 10.1007/s11390-011-1122-y

# Saliency-Based Fidelity Adaptation Preprocessing for Video Coding

Shao-Ping Lu (卢少平), *Student Member*, *CCF*, *ACM*, and Song-Hai Zhang (张松海), *Member*, *CCF*, *ACM*, *IEEE*

*Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*

*Beijing Engineering Research Center for Intelligent Processing of Visual Media and Content Security, Beijing 100084, China*

E-mail: lushaoping@cg.cs.tsinghua.edu.cn; sh-zhang04@mails.tsinghua.edu.cn

Received May 4, 2009; revised October 31, 2010.

**Abstract** In this paper, we present a video coding scheme which applies the technique of visual saliency computation to adjust image fidelity before compression. To extract visually salient features, we construct a spatio-temporal saliency map by analyzing the video using a combined bottom-up and top-down visual saliency model. We then use an extended bilateral filter, in which the local intensity and spatial scales are adjusted according to visual saliency, to adaptively alter the image fidelity. Our implementation is based on the H.264 video encoder JM12.0. Besides evaluating our scheme with the H.264 reference software, we also compare it to a more traditional foreground-background segmentation-based method and a foveation-based approach which employs Gaussian blurring. Our results show that the proposed algorithm can improve the compression ratio significantly while effectively preserving perceptual visual quality.

**Keywords** visual saliency, bilateral filter, fidelity adjustment, region-of-interest, encoder

## 1 Introduction

Appropriate bit allocation to provide optimal visual quality is a critical task in video coding. The human visual system works in such a way that people tend to concentrate on certain important details of a scene, while paying less attention elsewhere. Taking into account the importance of different parts of a scene is, we believe, a significant clue to the bit allocation problem for video coding.

Recently, several techniques have been developed using the human visual system's characteristics to guide region classification for bit-rate control. These take into account texture and spatial motion of regions[1-2], or specifically find face regions[3]. In [4-5], more bits are allocated to areas perceived to be foreground while the background is allowed to have a lower visual quality. Algorithms of the above kind process the image at the macroblock-level. Such region-based quantization adjustment may face several problems. Firstly, although significant progress has been made in image segmentation, it is still a challenging problem. Secondly, such regions do not correspond directly to macroblocks. A further issue is that quantization distortion is even more apparent to viewers than motion flickering or image blur[6]. Large quantization distortion between regions may be unacceptable to human visual perception.

Extraction and description of priorities of regions using a saliency model remain an open issue. Itti *et al.*[7] made one of the first attempts to develop saliency-based visual attention models. More recently, several video compression algorithms have been developed based on regions of interest (ROI). In [8], the author employed an eye-tracking device and coarse object segmentation technique to automatically select ROI before encoding. To produce blurring effects appropriate to the priority of each region, a 6-level Gaussian pyramid was constructed and applied to the image. It is hard to achieve perceptually seamless results if regions have clearly different blurring effects[8-9].

The conception of bilateral filtering was first introduced in [10] as a means of sharpening edges while blurring small discontinuities. It has been widely used in various fields such as photograph enhancement[11-12], video processing[13-14], optical flow and motion estimation[15], dynamic range image compression[16] and video compression[17]. The bilateral filter has proven to be generally applicable due to its simple formulation and fast non-iterative implementation.

In this paper, we present a video coding scheme based on the bilateral filter and saliency computation to selectively remove details from the image, allowing it to be encoded in fewer bits. The bilateral filter's control

Short Paper

This work was supported partially by the National High-Tech Research and Development 863 Program of China under Grant No. 2009AA01Z330, the National Natural Science Foundation of China under Grant Nos. 61033012 and 60970100.

parameters are locally adjusted by a perception-based image saliency map, to provide greater smoothing in less important areas of the image, while retaining details in more important areas. Use of a bilateral filter also ensures that sharp boundaries between regions are retained even after smoothing. Further steps are taken to ensure temporal coherence of smoothing between successive frames.

The reminder of this paper is organized as follows. In Section 2 our new saliency-aware adaptive fidelity adjustment mechanism is described. Section 3 presents experimental results based on JM12.0 of H.264/AVC to show the effectiveness of the proposed framework. We finish the paper with conclusions in Section 4.

## 2 Method

### 2.1 Architecture Overview

The framework of our proposed scheme, shown in Fig.1, comprises three major modules: (i) a visual saliency model computation, (ii) a local detail scalable bilateral filter computation and (iii) a video coder.

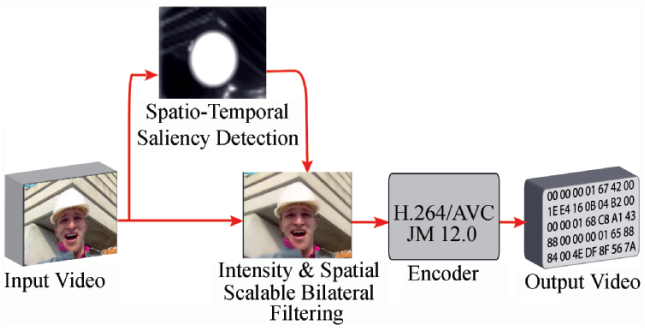


Fig.1. Framework of the proposed scheme.

Our aim is to adaptively preserve the visually important detail while smoothing other regions so that the encoding bit-rate can be effectively reduced. To do so, we use the following steps. Firstly, a spatio-temporal saliency model is computed by combining bottom-up and top-down analysis. Secondly, the bilateral filter is used to smooth the detail in the image in a way that is locally adjusted according to the saliency map. We do so in a way that avoids visual artifacts. The resulting degree of smoothing affects bit allocation in the encoder: the processed image is sent to an H.264/AVC encoder.

### 2.2 Spatio-Temporal Saliency Model

Researchers attempt to model the human visual system by creating models of saliency based on psychology and neurobiology. The human attention model in [18] takes both the top-down (task-related) and the bottom-up (scene-related) into account. Since the initial work on saliency detection in [7], a variety of visual saliency algorithms have been developed over the past decade[8,19-21].

When considering top-down saliency, particular attention has been paid to human faces due to their particular importance; face detection is becoming a mature technology, and is increasingly being incorporated into saliency models[19,22-24].

As Fig.2 shows, our model for computing saliency comprises both bottom-up and top-down attention components, as well as a spatial saliency smoothing process. Our approach follows [7-8, 19]. A similar bottom-up and top-down framework is also used in [22]. For the bottom-up saliency model, firstly, multi-scale feature maps are extracted corresponding to color, intensity, orientation, contrast, flicker and motion features. An *activation map* is constructed employing a graph-based feature model proposed in [19]. This has good performance and allows parallel implementation; it uses a Markov chain to provide the equilibrium distribution for the activation map. For the top-down saliency model, the widely used face detection algorithm in [25] is employed. A Gaussian convolution is used to blend the detected face area with the surrounding background, to avoid a sharp transition in saliency. Other top-down models could also be used, but faces are one of the most important subjects in video, which is why we have a specific saliency model for them. The top-down and bottom-up models are combined in a way which uses the top-down saliency if it is high, and the bottom-up saliency indicates an area of sufficient interest, and uses the bottom-up saliency otherwise. The computed saliency is smoothed to provide spatial coherence when the saliency is used for filter control. Temporal coherence is also enforced by blending saliency from frame to frame.

In detail, we combine the bottom-up saliency map with the top-down face saliency map using the following

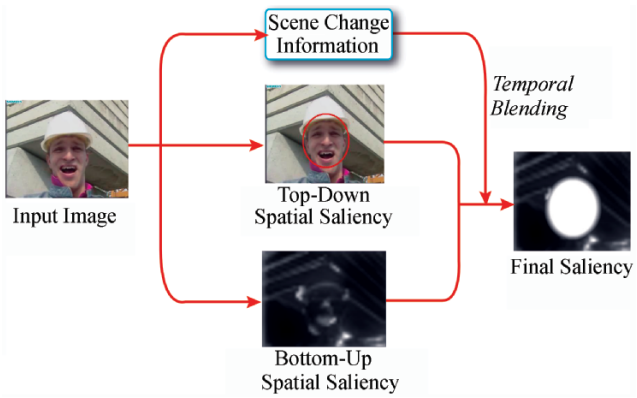


Fig.2. Saliency computation, combining bottom-up and top-down models.

approach. For a point $x$ in the image, the final saliency value $S(x)$ is defined as:

$$S(x) = \begin{cases} S_B(x), & \text{if } S_T(x) = 0, \\ \max(S_T(x), S_B(x)), & \text{if } S_T(x) \neq 0,\ S_B^r \geqslant T, \\ S_B(x), & \text{if } S_T(x) \neq 0,\ S_B^r < T, \end{cases} \tag{1}$$

where $T$ is a threshold value and we set it as $T = 10n$ and $n$ is the number of pixels in the face region $r$ detected by top-down face detection. Here $S_T(x)$ and $S_B(x)$ are the top-down and bottom-up saliency respectively, $S_B^r$ is the sum of bottom-up saliency in a face region $r$.

**Algorithm 1.** Pseudocode for Spatio-Temporal Saliency

```
S'_0(x) = S_0(x); //initial frame saliency
while (! finished all frames) do
  scene change detection;
  if (scene changed) then
    update γ(t) based on scene change;
    for (each pixel x in the image) do
      S'_t(x) = (1 − e^{−γ(t)}) · G(S_t(x)) + e^{−γ(t)} · S'_{t−1}(x);
    end for
  else
    for (each x in the image) do
      S'_t(x) = S_t(x);
    end for
  end if
end while
```

There are 3 cases to be considered. If the point $x$ lies outside any facial region determined by the facial saliency map ($S_T(x) = 0$), the saliency value $S(x)$ is simply equal to the bottom-up saliency $S_B(x)$. If $x$ belongs to a face region $r$, the sum of the bottom-up saliency in $r$ region $S_B^r(x)$ is calculated. If this is greater than or equal to a threshold value $T$, the saliency in $x$ is set to the larger of the bottom-up saliency and the facial saliency. We use a threshold $T$ because the face detection algorithm sometimes regards some small regions as face regions by mistake and we find such a threshold can effectively avoid most of them. Otherwise, the total bottom-up value in the region $r$ is less than $T$, and the value $S(x)$ should be set to $S_B(x)$. If a facial saliency region exists but it is considered insignificant by the bottom-up map, we use the bottom-up value as the final saliency and ignore the facial information.

Although the bottom-up model above takes flicker and motion features into account, we still need to consider temporal coherence of the computed saliency. To avoid performing a global pre-computation for the entire video sequence, we dynamically adjust saliency, which ensures visually satisfactory results. In general, the motion is small for each pixel in a video clip. Therefore, we smooth the saliency over time, using pseudocode from Algorithm 1 to determine the final saliency for each frame. We assume that the video is divided into a number of distinct scenes, and that we want to provide temporal coherence within each scene, but not between them. Here, $S'_t(x)$ is the adjusted saliency value for pixel $x$ in frame $t$, while $S_t(x)$ is the saliency value computed by (1). $G(\cdot)$ is the Gaussian convolution, and here we set the Gaussian kernel as $\sqrt{5}$. We use $e^{-\gamma(t)}$ to control blending of an exponential decay of the saliency from previous frames with the saliency of the current frame[26]. The decay factor $\gamma(t)$ is set to

$$\gamma(t) = N + t - t_0 \tag{2}$$

where $N$ is the average number of frames in previous scenes (this is our best estimate for the length of the current scene), and $t - t_0$ is the number of frames since the change of the last scene. $N$ is used to limit the influence of previous frames. For the first frame, or the first frame after a scene changes, we set the saliency to $S_t(x)$. Various approaches can detect whether a scene has changed[27-28]. Here we employ the color histograms-based shot-change detection method[28] for its low computational cost.

### 2.3 Extended Bilateral Filter

A bilateral filter replaces pixels in the image by a weighted mean of their neighbors. The weight of each neighbor pixel is related to both its 2D spatial distance in the image plane and distance in intensity in each color channel. As a result, a bilateral filter can effectively sharpen edges while removing small differences in intensity values.

Let $x$ be the location of some pixel in the image, and $\hat{x}$ be the location of one of its neighbors in the neighborhood $\Omega$: those pixels no further than $\sigma_s$ from $x$. Let $C_x$ be the input color of pixel $x$. Then the bilateral filter computes its output color $C'_x$ as:

$$C'_x(\sigma_s, \sigma_r) = \frac{1}{W_\Omega} \sum_{\hat{x} \in \Omega} \omega_s(x, \hat{x}, \sigma_s) \omega_r(x, \hat{x}, \sigma_r) C_{\hat{x}}, \tag{3}$$

where $W_\Omega$ normalizes the sum of the weights

$$W_\Omega = \sum_{\hat{x} \in \Omega} \omega_s(x, \hat{x}, \sigma_s) \omega_r(x, \hat{x}, \sigma_r). \tag{4}$$

In the above, $\omega_s(x, \hat{x}, \sigma_s)$ is the weight for spatial distance between pixels and is given by

$$\omega_s(x, \hat{x}, \sigma_s) = \exp\Big( - \frac{(x - \hat{x})^2}{2\sigma_s^2} \Big). \tag{5}$$

$\sigma_s$ denotes the spatial scale factor: blurring will be greater with larger $\sigma_s$. If $\sigma_s$ is too large, significant visual features across boundaries may be lost by bilateral filtering. Similarly $\omega_r(x, \hat{x}, \sigma_r)$ weights averaging of neighbors according to their difference in intensity; $\omega_r(x, \hat{x}, \sigma_r)$ is defined as:

$$\omega_r(x, \hat{x}, \sigma_r) = \exp\Big(-\frac{(C_x - C_{\hat{x}})^2}{2\sigma_r^2}\Big) \qquad (6)$$

where $\sigma_r$ is an intensity scale factor. When $\sigma_r$ is small, the filter will preserve almost all the intensity differences. As its value increases, the filter will tend to a standard Gaussian blur filter. Bilateral filtering is performed in Lab color space for best results (all channels are treated in the same way): in this space, Euclidean distances correlate closely to perceived color discrepancy[10]. Our key idea is to use the saliency map to locally adjust $\sigma_s$ and $\sigma_r$.

One example of bilateral filtering performed by the above method is shown in Fig.3. The original image, Fig.3(a), was transformed into CIE-Lab color space and then processed by the bilateral filter. The filtered result is shown in Fig.3(b) and close-ups are shown in Figs. 3(c) and 3(d).

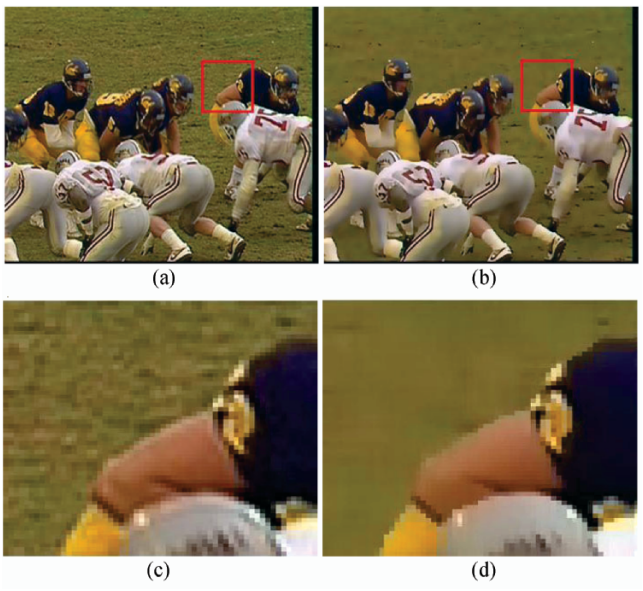


Fig.3. Results of bilateral filtering using fixed values ($\sigma_r = 4.25$, $\sigma_s = 3$) over the whole image. The bilateral filter smoothes small details while preserving sharp boundaries. (a) Original image. (b) Bilateral filtered image. (c) Zoom of (a). (d) Zoom of (b).

Figs. 4(a) and 4(b) show the luminance channel values for the close-up regions. Note that the filter smoothes the grass background while preserving the visual contrast boundaries and retaining detail elsewhere.

Considerable high frequency information is removed by the bilateral filter from each color channel in relatively uniform low-interest areas without significant visual quality loss, allowing the bit-rate to be reduced. The bit-rate achieved with various settings for the intensity scale $\sigma_r$ and the spatial scale $\sigma_s$ is shown in Fig.5 for the football sequence. Larger values for $\sigma_r$ and $\sigma_s$ allow the coder to allocates fewer bits.

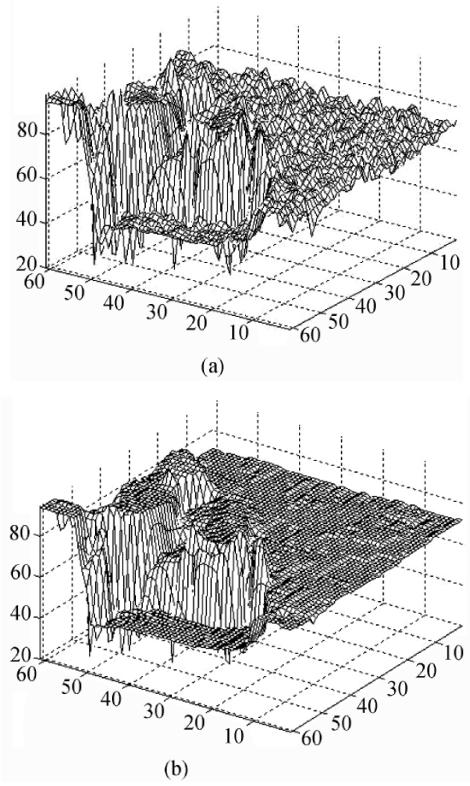


Fig.4. Details of the bilateral filtering results. See that the bilateral filter smoothes small details while preserving sharp boundaries. (a) Luminance of Fig.3(c). (b) Luminance of Fig.3(d).

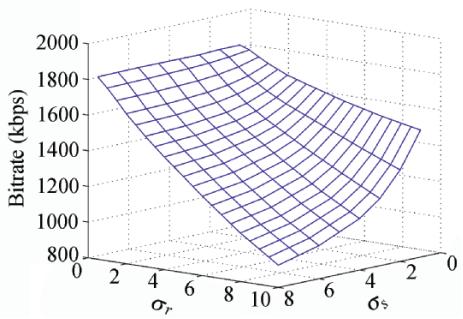


Fig.5. Rate distribution achieved after bilateral filtering for the football clip while varying $\sigma_s$ and $\sigma_r$. $QP$ is constant at 28.

To construct the map between visual saliency and the scale used for the bilateral filter, we define an intermediate function $I(\kappa, x)$ as:

$$I(\kappa, x) = \frac{\kappa}{2}\left(1 - \tan\left(3\left(\frac{S'(x)}{128} - 1\right)\right)\right). \quad (7)$$

$S'(x)$ is the spatio-temporal saliency value computed as above, which is normalized to 256 across the image. $\kappa$ is a user-defined threshold variable. This function maps the saliency between 0 and $\kappa$. We then simply map the intermediate function to the intensity and spatial scale factor in the bilateral filter. The local intensity scale factor $\sigma_r(x)$ at pixel location $x$ is set to:

$$\sigma_r(x) = I(10, x) \quad (8)$$

while the spatial scale factor $\sigma_s(x)$ is set to:

$$\sigma_s(x) = \lfloor I(8, x) \rfloor. \quad (9)$$

Our experiments show that setting $\sigma_r > 10$ or $\sigma_s > 8$ produces results which are no longer visually acceptable because too much is smoothed.

## 3 Experimental Results

We implemented our video coding scheme using the H.264/AVC reference software JM12.0[29]. Rate-distortion optimization (RDO) mode was enabled, Hadamard transforms and variable block sizes were selected. The search range was set to 32. To evaluate our scheme, the following experiments were done: (i) output from the proposed scheme was compared to that produced by the H.264 reference software using both objective and subjective quality metrics; (ii) we compared results from our scheme with those produced by related algorithms in the literature, such as [8-9].

We first consider subjective image quality. In this experiment, the quantization parameters ($QP$) of our scheme and JM12.0 were fixed at 28. The $QP$ is a quantization step size and larger value of it can produce higher compression but also more distortion. We used the standard football, mother, news and carphone video sequences, all of which were in YUV 4:2:0 CIF format with a 30 Hz frame rate. Detailed visual results are shown in Fig.6. The most important regions (such as the face and the player's bodies) and global textures are preserved well, while the uniform areas (such as the view from the window, and the grass) are smoothed. Note that fidelity adjustment results have smooth boundaries, which effectively avoids any block artifacts or other seams between regions.

A bit-rate gain comparison for JM12.0 and our method is shown in Fig.7. The video sequences were encoded by the default JM coder, and our scheme with constant $QP$ respectively. The graph shows that the number of bits needed for our coding scheme is almost 40% fewer for the football and mother video sequences,


Fig.6. Comparison of results with $QP = 28$. Note that in our result the face and the player's bodies are well preserved while uniform areas are smoothed. (a) JM12.0 result (292 kbps). (b) Our result (243 kbps). (c) JM12.0 result (1843 kbps). (d) our result (1120 kbps).

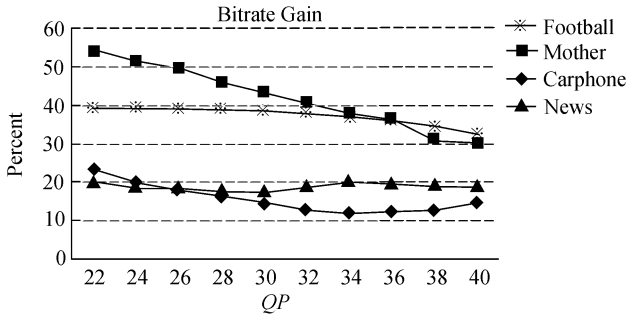


Fig.7. Our method results in a bit-rate gain of at least 20% and up to 50% over the JM12.0 method. This means significantly fewer bits are needed to code images with our method for any quantization parameter.

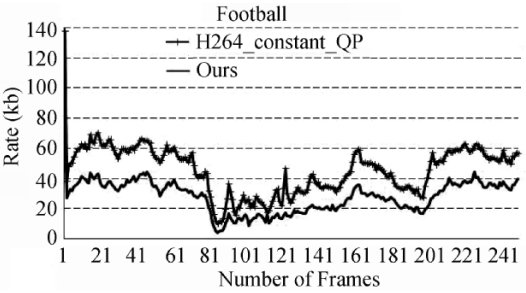


Fig.8. Bit-rate performance comparison between JM12.0 and our algorithm with $QP$ fixed at 28.

and 17% fewer for the carphone and news sequences. As $QP$ increases, the bit gain decreases, and ultimately

the bit-rate of the two methods gradually converges. A frame-by-frame bit-rate comparison for the football clip is shown in Fig.8. Note also that these results imply that our algorithm avoids fluctuation in bit allocation between frames.

Fig.9 shows the image quality degradation of our scheme under different $QP$s for the mother video sequence. Since most high frequency information has been processed by our adaptive bilateral filter, we find that the degradation is not very obvious except for the face regions when $QP$ gradually becomes large.

Fig.10 compares our algorithm with two other approaches: Cavallaro's foreground-background segmentation based method[9], and Itti's foveation filter model with multiple Gaussian pyramids[8]. Fig.10(b) is taken from [9]. We believe the visual results in the regions around the two persons in Fig.10(c) are more acceptable with our scheme. Uniform areas such as the floor and the wall are smoothed with less visual distortion while most of the structural information in the image is preserved. Our result also avoids seaming effects between foreground and background, and other visual artifacts. Fig.10(e) is taken from Itti's homepage. Fig.10(f) shows that our algorithm can preserve information in unimportant regions more effectively. Changes in blurring are more gradual and less noticeable with our method.

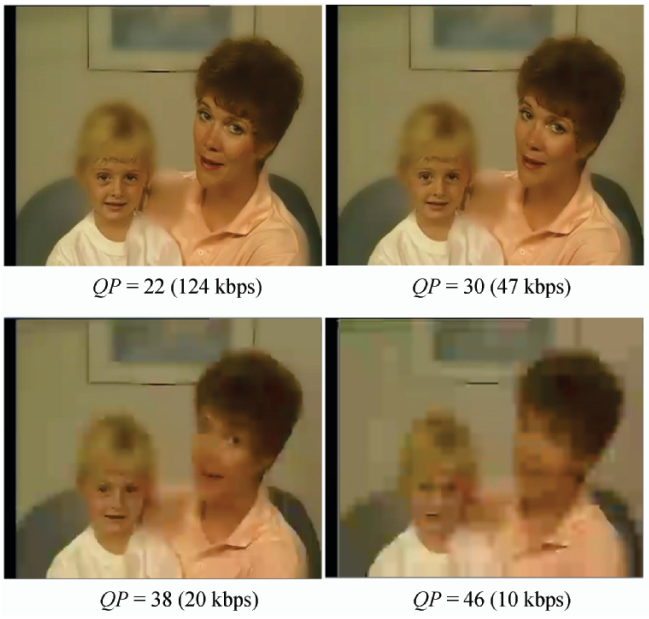


Fig.9. Visual results of our scheme show the trade-off between bit allocation and perceptual information. As $QP$ goes up, the number of bits goes down, and eventually perceptual information is reduced.


Fig.10. Our method provides a more visually pleasing compressed image as compared to Cavallaro and Itti's methods because it avoids seaming effects (near the boundaries of the two persons in (c)) and better preserves structural information in the unimportant areas (the floor and the wall in (c), the grass and sky areas in (f)). (a) Input sequence (30 413 kbps). (b) Cavallaro's result (∼150 kbps). (c) Our result (∼150 kbps). (d) Input sequence (110 592 kbps). (e) Itti's result (∼1000 kbps). (f) Our result (∼1000 kbps).

## 4 Conclusions

This paper presents a novel fidelity adjustment scheme for region-of-interest-based video coding. Fidelity is adaptively adjusted by a locally adjustable bilateral filter which is set according to results of visual saliency analysis. Less perceptually important areas in the images are smoothed more strongly than perceptually important areas. Our experiments show that such variable smoothing allows the video coder to allocate fewer bits to unimportant regions, enabling greater coding gain. The algorithm nevertheless effectively preserves visually important information while gradually degrading the unimportant regions.

Instead of partitioning the whole picture into macroblocks, this algorithm applies a pixel-based process. Furthermore, it does not need any segmentation, which is often difficult to do reliably in practice. Since the bilateral filter changes gradually, transitions between areas of different fidelity are smooth and region boundaries remain sharp.

Our new fidelity adjustment scheme provides a novel approach to the trade-off between bit allocation and preservation of perceptual information for video coding. Our scheme is independent of the video codec used. We show significant reduction in bit rates necessary compared to H.264/AVC JM12.0 with constant *QP*. Our scheme can be incorporated in the traditional bit-rate control framework to obtain improved performance. Additionally, the strict mapping between fidelity and bits in our scheme makes it possible to adjust the fidelity to give a particular target bit-rate. Fidelity control can also be adaptively adjusted according to the bandwidth available in network applications.

Further work is still needed to improve the perceptual quality in conjunction with bit allocation. Color space transformation between YUV and CIE-Lab is not lossless, and may cause color distortion in some video frames. Our proposed fidelity adjustment algorithm requires saliency detection, and saliency map computation is an ongoing topic of research in computer vision. Our future work will focus on achieving more robust spatio-temporal fidelity to improve the perceptual quality of the results.

**Acknowledgements** We would like to thank anonymous reviewers and editors for their valuable suggestions to improve the paper. Also, thanks go to Ralph R. Martin for his helpful discussions.

## References

[1] Tao B, Dickinson B W, Peterson H A. Adaptive model-driven bit allocation for MPEG video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 2000, 10(1): 147-157.

[2] Tang C W, Chen C H, Yu Y H, Tsai C J. Visual sensitivity guided bit allocation for video coding. *IEEE Transactions on Multimedia*, 2006, 8(1): 11-18.

[3] Chen M J, Chi M C, Hsu C T, Chen J W. ROI video coding based on H.263+ with robust skin color detection technique. *IEEE Transactions on Consumer Electron*, 2003, 49(3): 724-730.

[4] Chai D, Ngan K N. Foreground/background video coding scheme. In *Proc. IEEE Int. Symp. Circuits Syst*, Hong Kong, China, Jun. 9-12, 1997, pp.1448-1451.

[5] Lee S, Pattichis M S, Bovik A C. Foveated video compression with optimal rate control. *IEEE Transactions on Image Process*, 2001, 10(7): 977-992.

[6] Wang D, Speranza F, Vincent A, Martin T, Blanchfield P. Towards optimal rate control: A study of the impact of spatial resolution, frame rate and quantization on subjective video quality and bit rate. In *Proc. SPIE 2003*, Lugano, Switzerland, Jul. 8-11, 2003, pp.198-209.

[7] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Anal. and Machine Intell.*, 1998, 20(11): 1254-1259.

[8] Itti L. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, 2004, 13(10): 1304-1318.

[9] Cavallaro A, Steiger O, Ebrahimi T. Semantic video analysis for adaptive content delivery and automatic description. *IEEE Transactions on Circuits and Systems for Video Technology*, 2005, 15(10): 1200-1209.

[10] Tomasi C, Manduchi R. Bilateral filtering for gray and color images. In *Proc. ICCV*, Bombay, India, Jan. 4-7, 1998, pp.839-846.

[11] Eisemann E, Durand F. Flash photography enhancement via intrinsic relighting. *ACM Transactions on Graphics*, 2004, 23(3): 673-678.

[12] Huang H, Zang Y, Rosin P L, Qi C. Edge-aware level set diffusion and bilateral filtering reconstruction for image magnification. *Journal of Computer Science and Technology*, 2009, 4(24): 734-744.

[13] Bennett E P, McMillan L. Video enhancement using per-pixel virtual exposures. *ACM Transactions on Graphics*, 2005, 24(3): 845-852.

[14] Winnemöller H, Olsen S C, Gooch B. Real-time video abstraction. *ACM Transactions on Graphics*, 2006, 25(3): 1221-1226.

[15] Xiao J J, Cheng H, Sawhney H, Rao C, Isnardi M. Bilateral filtering-based optical flow estimation with occlusion detection. In *Proc. ECCV*, Graz, Austria, May 7-13, 2006, pp.211-224.

[16] Paris S, Durand F. A fast approximation of the bilateral filter using a signal processing approach. In *Proc. ECCV*, Graz, Austria, May 7-13, 2006, pp.568-580.

[17] Pham T Q, Van Vliet L J. Separable bilateral filtering for fast video preprocessing. In *Proc. IEEE ICME*, Amsterdam, Netherlands, Jul. 6-9, 2005, pp.454-457.

[18] William J. The Principles of Psychology. Cambridge, MA: Harvard University Press, 1981.

[19] Cerf M, Harel J, Einhäuser W, Koch C. Predicting human gaze using low-level saliency combined with face detection. In *Proc. NIPS*, Vancouver, Canada, Dec. 3-7, 2007, pp.241-248.

[20] Sebe N, Lew M S. Comparing salient point detectors. *Pattern Recognition Letters*, 2003, 24(1): 89-96.

[21] Robert J P, Iyer A, Itti L, Koch C. Components of bottom-up gaze allocation in natural scenes. *Journal of Vision*, 2005, 5(8): 692-692.

[22] Tsapatsoulis N, Pattichis C, Rapantzikos K. Biologically inspired region of interest selection for low bit-rate video coding.

In *Proc. ICIP*, San Antonio, USA, Sept. 16-19, 2007, pp.305-308.

[23] Chen W F, Liu C H, Lander K, Fu X L. Comparison of human face matching behavior and computational image similarity measure. *Science in China Series F: Information Sciences*, 2009, 52(2): 316-321.

[24] Lee K W. Guiding attention by cooperative cues. *Journal of Computer Science and Technology*, 2008, 5(23): 874-884.

[25] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*, Hawaii, USA, Dec. 11-13, 2001, pp.511-518.

[26] Paris S. Edge-preserving smoothing and mean-shift segmentation of video streams. In *Proc. ECCV*, Marseille, France, Oct. 12-18, 2008, pp.460-473.

[27] Zhu S H, Liu Y C. Two-dimensional entropy model for video shot partitioning. *Science in China Series F: Information Sciences*, 2009, 52(2): 183-194.

[28] Gargi U, Kasturi R, Strayer S H. Performance characterization of video-shot-change detection methods. *IEEE Transactions on Circuits and Systems for Video Technology*, 2000, 10(1): 1-13.

[29] H.264/AVC reference software [online]. http://iphome.hhi.de/suehring/html.

**Shao-Ping Lu** is a Ph.D. candidate at Department of Computer Science and Technology in Tsinghua University. His research interests include image and video process. He is a student member of China Computer Federation and ACM.

**Song-Hai Zhang** obtained his Ph.D. degree in 2007 from Tsinghua University. He is currently a lecturer of computer science at Tsinghua University, China. His research interests include image and video processing, geometric computing. He is a member of China Computer Federation, ACM and IEEE.