Correspondence

Utility-Driven Adaptive Preprocessing for Screen Content Video Compression

Shiqi Wang, *Member, IEEE*, Xinfeng Zhang, Xianming Liu, Jian Zhang, *Member, IEEE*, Siwei Ma, *Member, IEEE*, and Wen Gao, *Fellow, IEEE*

***Abstract*—In this work, we propose a utility-driven preprocessing technique for high-efficiency screen content video (SCV) compression based on the temporal masking effect, which was found to be a fundamental attribute that plays an important role in human visual perception of video quality, but has not been fully exploited in the context of SCV coding. Specifically, we investigate the temporal masking effect from the perspective of perceived utility, which allows us to preserve the quality of the high utility content and substitute the low utility regions with the corresponding smooth version. To distinguish the regional utilities, a specifically designed block type identification algorithm for screen content is employed to measure the local properties. Subsequently, the Gaussian filter is applied to smooth out the high-frequency components in the detected low utility regions to save consumption bits. Validations based on subjective testings show that the proposed approach is capable of achieving significant bitrate savings with little sacrifice on the final utility compared with the conventional SCV coding scheme.**

***Index Terms*—Block type identification, screen content video (SCV), temporal masking, utility information.**

I. Introduction

Recent years have witnessed dramatically increased interest and demand for the mobile computer devices, such as laptops, tablets, PDAs, smartphones. Due to the constraints imposed by the local computing capacity and data resources on such devices, many remote computing and virtualization scenarios have emerged. The purpose of these applications is to access the remote data and control the computational resources via the networks [1]–[4]. In these scenarios, remote computing can be achieved by the users' interactions with the local interface. The content of such interface is usually generated by the computer, and can be regarded to be a kind of time-varying screen content video (SCV) composed of both computer generated textual/graphical and natural images. In such an environment, there is considerable concern regarding how the captured SCVs can be efficiently delivered.

Manuscript received November 3, 2015; revised April 7, 2016 and June 30, 2016; accepted August 28, 2016. Date of publication November 4, 2016; date of current version February 14, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61322106, Grant 61632001, Grant 61571017, Grant 61300110, and Grant 61672193, and in part by the National Basic Research Program of China (973 Program) under Grant 2015CB351800. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shahram Shirani. *(Corresponding author: X. Liu.)*

S. Wang and X. Zhang are with the Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore 637553 (e-mail: wangshiqi@ntu.edu.sg; xfzhang@ntu.edu.sg).

X. Liu is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: xmliu.hit@gmail.com).

J. Zhang, S. Ma, and W. Gao are with the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: jian.zhang@pku.edu.cn; swma@pku.edu.cn; wgao@pku.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2625276

Compared with natural videos, SCVs exhibit distinguished properties in both spatial and temporal domains. In spatial domain, the noise free screen content frames are usually featured with repeated patterns, thin lines, limited colors and large smooth areas [5]. In temporal domain, various types of motion, such as long range, irregular and global motions, are usually involved in a typical SCV. In addition to the conventional hybrid video coding scheme [6] and advanced super-resolution based image/video compression approaches with superior coding performance [7]–[9], these distinguished properties have motivated numerous specifically developed SCV coding techniques, many of which have been adopted into the screen content coding (SCC) extension to the High Efficiency Video Coding (HEVC) standard [10]. For example, the residuals of the textual content are usually sparse and sometimes contain sharp directional edges, which may not follow the vertical or horizontal Discrete Cosine Transform (DCT) directions. In view of this, strategies of skipping the transform process were proposed [11]–[13]. Moreover, it is observed that the screen content typically contains a limited number of distinct colors, inspired by which color table/palette method has been widely investigated [5], [14]. The intra motion compensation approach was also proposed to remove the redundancy from the repeated patterns occurred in one frame [15]. Recently, to further reduce the redundancy of the prediction residuals in different color components, the adaptive color-space transform technique was subsequently adopted [16]. In temporal domain, inspired by the observation that motion in screen content is usually based on full-pel resolution, adaptive motion vector resolution (AMVR) scheme was proposed in [17] to adapt the resolutions of motion vectors between full- and sub-pel. Moreover, hash based block matching techniques for text/graphics have been developed for intra and inter block search [18].

In addition to these advanced coding techniques, perceptually relevant properties of screen content image (SCI) have also been extensively studied in the literature. In [19]–[21], various screen quality assessment algorithms were developed based on the SCI quality assessment database in [21]. Yet, these methods only focus on perceptual quality for an individual image, and the substantial difference between SCI and SCV lies in the visual sensitivity exploration in the temporal domain. For natural video coding, the temporal masking effects have been widely exploited to improve the coding efficiency [22]–[24]. However, most of them are designed and validated on natural videos, which may not always share the same properties of SCVs.

As widely hypothesized in computational vision science, the major task of the human visual system (HVS) when viewing an image is to act as an optimal information extractor [25]. This motivates us to study the temporal masking properties of SCVs from the perspective of perceived utility, as high utility corresponds to high information content that needs to be extracted through the interactive screen-remoting mechanism. However, with numerous work proposed to evaluate the utility of natural image [26]–[28], much less work has been dedicated to SCVs. Based on the philosophy of the image utility assessment,

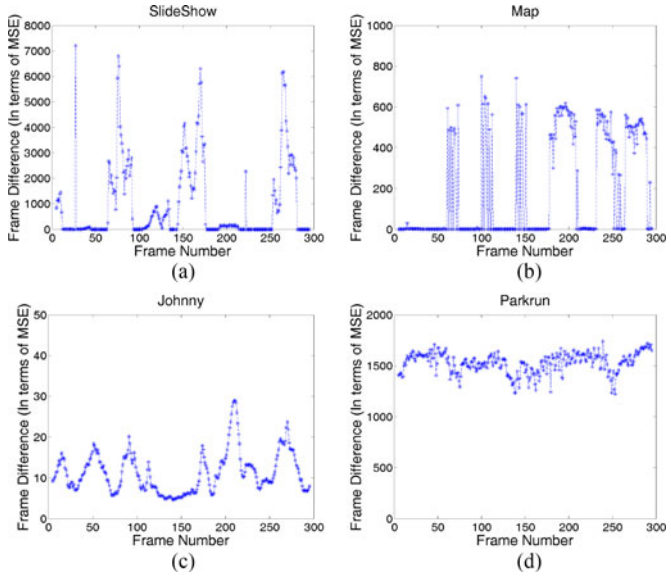

Fig. 1. Variations of the temporal frame difference for SCVs and natural video sequences.

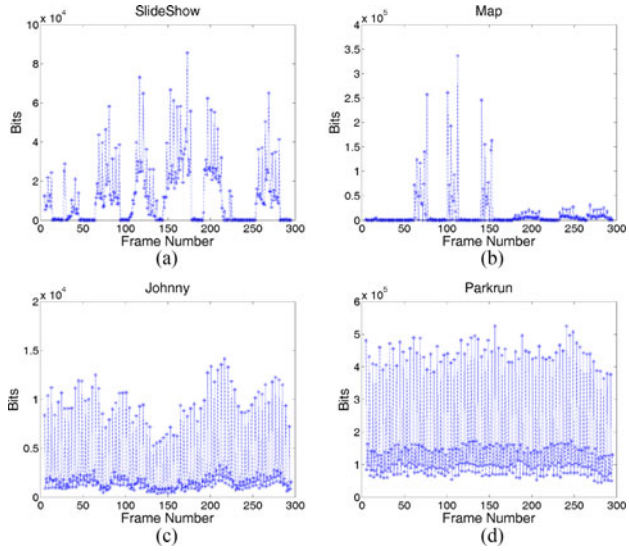

Fig. 2. Variations of the coding bits for SCVs and natural video sequences.

appropriately injected distortions can be tolerated in lower utility regions as long as the underlying task is reliably performed, as the HVS will regard the usefulness of the image as a surrogate for a reference [27]. Regarding SCV, the utility can be characterized by the temporal stability in screen content refreshing. For instance, the fading in/out between two slides or the zooming in/out operations may produce low utility regions, as the major task of these frames is to guarantee a smooth transition during content refreshing. As such, a smooth version of such content can provide a utility equivalent SCV.

In this paper, we propose an utility-driven adaptive preprocessing approach for SCV compression. The approach adaptively identifies and processes the low utility content using Gaussian low-pass filtering. As such, higher coding efficiency is ensured by effectively reducing the coding bitrate of the low utility regions. The proposed scheme applies on the captured SCV such that the SCV generation process is not affected. Experimental results show that the proposed scheme can significantly save the bitrate in transmitting the SCVs.

II. Characteristics of Screen Content Video

In this section, we analyze the characteristics of the SCV, especially focusing on the video content fluctuations along the temporal direction. Specifically, four video sequences (Parkrun, Johnny, Map and Slideshow) are employed for investigation. The resolution of them is 1280×720 and 300 frames are used for testing. Among them, Map and Slideshow are typical SCVs, Johnny is the natural video with static background, and Parkrun is the natural video with global motion.

The mean square error (MSE) between two adjacent frames is firstly computed to demonstrate the content variations. This is a simple but effective measure that can well reflect the frame difference caused by any motion or scene change. The variations of frame difference are demonstrated in Fig. 1, from which we can observe that there are two poles for SCVs. At one extreme, adjacent screen frames remain quite stable, leading to approximate zero difference. At the other extreme, the irregular and global motions caused by zooming in/out, flipping over and dragging etc. may produce extraordinarily large difference. However, as the content of natural video evolves over time, the frame differences of natural videos are usually larger than zero and exhibit smooth variations over a period of time.

Furthermore, we compress these sequences using HEVC codecs to investigate the content variations in terms of the frame level coding bits. In particular, the natural videos are compressed with HEVC main profile (HM16.2) and the SCVs are compressed with the SCC extension of HEVC (HM16.2-SCM3.1), respectively. The configuration is Low Delay and the quantization parameter (QP) is set to be 24 for all the sequences. The coding bit as a function of frame index serves as another indicator to reflect the residual energy and motion activities, and their variations are illustrated in Fig. 2. Since the hierarchical bit allocation structure (rate-GoP) is employed in HEVC [29], [30], each frame will be coded with different QPs and Lagrangian multipliers according to the corresponding layer index, leading to periodic coding bits variations within every four frames. We can observe that there exist significant coding bits variations for SCVs, which originate from the SCV content fluctuations with irregular motions. However, such large variations of coding bits may bring challenges in the deployment of SCV compression, as many SCV applications are constrained by limited buffer space and bandwidth. These observations demonstrate the potentials and necessities of reducing the coding bits in those frames with irregular motions, and suggest us to exploit the perceptual redundancy to improve the SCV coding efficiency.

III. Screen Content Video Preprocessing

A. Motivation

The motivation behind the SCV preprocessing is to generate the utility equivalent frames that consume less bits than the directly recorded ones. In human-computer interaction, the time variant interface is usually subject to various operations, such as zooming in/out, fading in/out, windows moving, and page scrolling to accomplish the goal of screen sharing. These operations usually create large content variations, and some of them cannot be efficiently coded as there does not exist an appropriate reference for prediction. However, the central role of such unstable content is ensuring smooth transitions rather than providing useful information, as the high utility content usually remains steady in several consecutive frames for the purpose of information extraction. As a result, the capability of blurring such low utility content for compression allows one to save bitrate at the current moment of low utility content, and reserve bandwidth and buffer capacities for future frames that desire more bitrates to maintain the SCV quality.

Generally speaking, the low utility content can be a whole frame, or only regions within the SCV. To identify such content, the spatial and temporal local properties of each frame should firstly be accessed to distinguish the block type. Subsequently, we perform an object level region detection to locate the low utility content. Finally, a Gaussian smooth filter is applied to generate the utility equivalent SCV.

B. Block Type Identification

The local statistical properties of the SCV are captured by means of block type identification. Specifically, each frame is divided into non-overlapping 16×16 blocks, which are further categorized into one of three types: skip, text, and natural image [1], [31], [46]. We firstly introduce the detection algorithms for skip and text, and then the remaining blocks are classified as natural image blocks.

1) Skip Block: SCVs often exhibit high temporal redundancy in terms of motion compensation. Therefore, detecting the skip block is one essential procedure in block type identification [1], [31], [32], [46]. One extreme case is that there is a high probability of fixed blocks with no scene change. In addition, scrolling up/down and moving a window typically produce content updates, which exhibits large areas of global motion [33]. As such, skip blocks are further classified into fixed blocks and global motion blocks. Fixed block is easy to examine by comparing the current block with the co-located block in the adjacent frames, and if there is no difference, the block is identified as fixed. To detect the blocks that are subjected to global motion, we employ the global motion detection technique that makes use of feature comparison to locate the global motion areas [34]. In this manner, the long range motions that frequently occur in SCV can be efficiently estimated. It is also worth mentioning that the future frames will also be used for detecting the skip block to infer its temporal properties, which may introduce a slight delay in this process.

2) Text Block: Motivated by the observation that text blocks are usually characterized by sharp edges, we firstly employ a binary feature derived from the number of high gradient pixels in an $M \times N$ block for high gradient block type classification [32], [35], [46]

$$\zeta = u\left(\sum_{k=1}^{M-1}\sum_{l=1}^{N-1}\gamma_{k,l} - T_{\mathrm{HF}}\right) \tag{1}$$

where T_{HF} represents a certain threshold of high gradient pixels and function u is defined as a step function as follows:

$$u(x) = \begin{cases} 1, & x > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The parameter $\gamma_{k,l}$ is a binary value indicating whether the (i,j)-th pixel is a high gradient pixel

$$\gamma_{k,l} = u(|I_{k,l} - I_{k-1,l}| > I_{\mathrm{th}} \vee |I_{k,l} - I_{k,l-1}| > I_{\mathrm{th}}) \tag{3}$$

where I_{th} is the pre-defined threshold and $I_{k,l}$ represents the luma samples at the location (k,l).

However, in a typical SCV frame, the high gradient block can be either a text block, or an edge block from natural image regions. To further differentiate the pictorial and textual content, the limited color is employed as another distinguished feature of text block [31], [32], [46]. Specifically, the text block can be represented by a few number of colors and their corresponding positions, which are also known as base color and index map. Such unique feature has been widely employed in screen content compression and processing [5], [35], [36]. One typical example is illustrated in Fig. 3, in which only limited number of sample values exists in the text block. As such, the text block can be represented using their index map, as shown in Fig. 3(c). To effectively identify

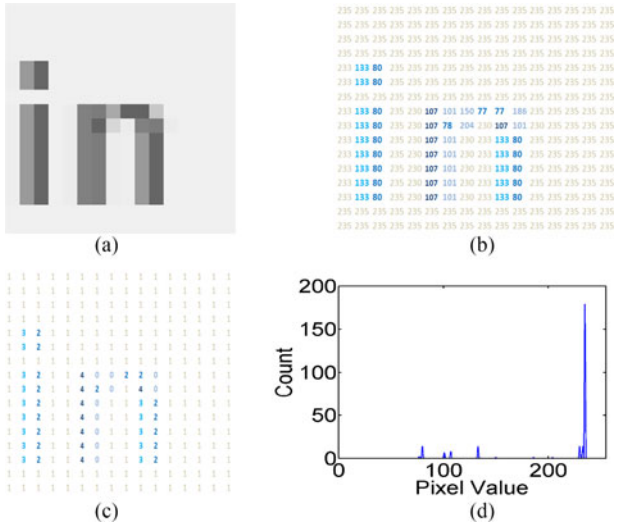

Fig. 3. Demonstration of the limited color representation. (a) Amplified luminance block. (b) Luminance pixel values of (a). (c) The map of major color index. (d) The histogram of pixel value.

these base colors, we build the pixel value histogram where a majority of pixels are converged to a small number of colors, as illustrated in Fig. 3(d). These colors are treated as base colors and the number of base colors is limited to four. Equal size windows around the base colors are used to range the sample values and those that cannot be represented by base colors are identified as escape colors. When the number of escape colors is smaller than a certain threshold, suggesting that limit colors suffice to represent the block, the current block is categorized into text block. Otherwise, it is treated as a natural image block.

C. Low Utility Region Detection

Given the block types, the low utility regions are adaptively detected for each frame. First of all, skip blocks should not be included in the low utility region for further processing, since such blocks may easily get referenced by the subsequent frames or inherited from the previous frames. In view of this, including skip blocks in the low utility region may break the temporal consistency and decrease the coding efficiency. Regarding the remaining text and natural image blocks, we locate the low utility regions using a probabilistic strategy.

Let the binary value m ($m \in \{1,0\}$) denote whether the current block is included in the low utility region or not. Given the (i,j)-th block in the t-th frame, the posterior distribution of $m_{i,j,t}$ can be predicted with the Bayes' theorem

$$p(m_{i,j,t}|f_{i,j,t}) = \frac{p(f_{i,j,t}|m_{i,j,t})p(m_{i,j,t})}{p(f_{i,j,t})} \tag{4}$$

where $f_{i,j,t} \in \{1,0\}$ indicates the block type, and 1 denotes text and 0 denotes natural image, respectively. Therefore, the maximum a posteriori (MAP) estimation of $m_{i,j,t}$ is given by

$$\begin{aligned} \hat{m}_{i,j,t} &= \operatorname{argmax} p(m_{i,j,t}|f_{i,j,t}) \\ &= \operatorname{argmax} p(f_{i,j,t}|m_{i,j,t})p(m_{i,j,t}). \end{aligned} \tag{5}$$

The likelihood function can be rewritten as

$$p(f_{i,j,t}|m_{i,j,t}) = p(f_{i,j,t}|1)^{m_{i,j,t}} p(f_{i,j,t}|0)^{1-m_{i,j,t}}. \tag{6}$$

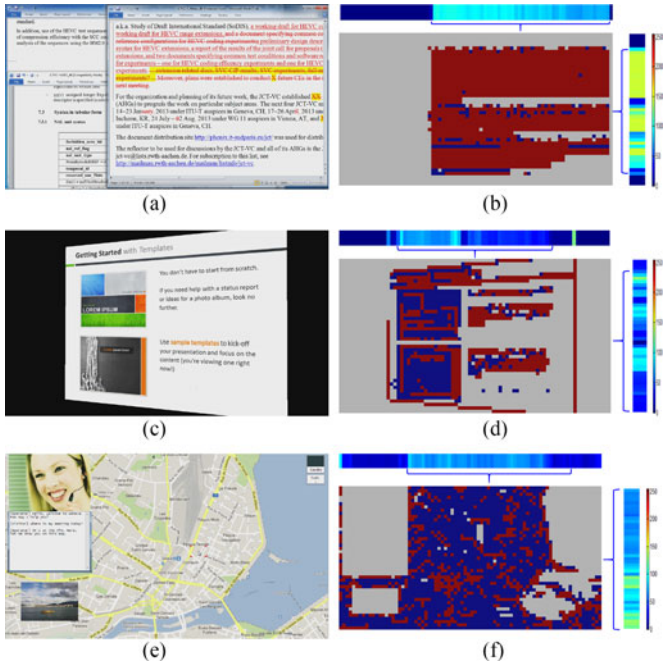

Fig. 4. Illustration of the low utility region detection process in SC_Wordediting, SC_Slideshow and SC_Map. Gray: skip block; red: text block; blue: natural image block. The heat map is employed to visualize the local vertical and horizontal activities.

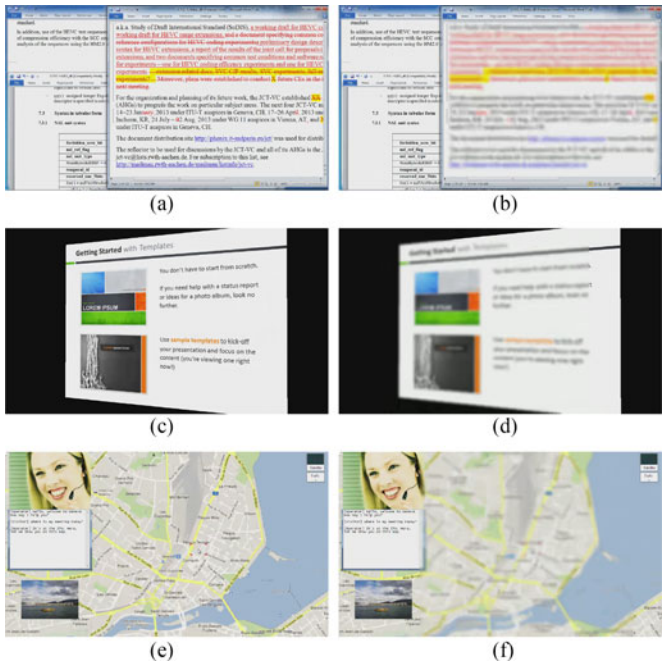

Fig. 5. Comparisons between the anchor and proposed SCV coding schemes of SC_Wordediting (288-th), SC_Slideshow (150-th), and SC_Map (143-th) frames with QP = 37. (a), (c), (e): anchor; (b), (d), (f): proposed.

Inspired by the pairwise interaction markov random field, we model the prior distribution $p(m_{i,j,t})$ based on the previous t_0 frames in the following form:

$$p(m_{i,j,t}) \propto \exp\left(\sum_{s=0}^{t_0} \beta_{t,t-s}\epsilon_{m_{i,j,t}=m_{i,j,t-s}}\right) \tag{7}$$

where $\beta_{t,t-s}$ follows 1-D Gaussian distribution with standard deviation (std) that is empirically chosen as $\sigma = 1.5$

$$\beta_{t,t-s} = \frac{\alpha}{\sigma\cdot\sqrt{2\pi}}\exp\left(-\frac{s^2}{2\sigma^2}\right). \tag{8}$$

Here the parameter α is used to balance the prior and likelihood. The term $\epsilon_{m_{i,j,t}=m_{i,j,t-s}}$ is set to one if $m_{i,j,t} = m_{i,j,t-s}$ and zero otherwise.

With (6) and (7), the log-posterior distribution of $\ln p(m_{i,j,t}|f_{i,j,t})$ is computed as follows:

$$\begin{aligned}\ln p(m_{i,j,t}|f_{i,j,t}) = {} & \ln p(f_{i,j,t}|0) + \phi_{i,j,t}\cdot m_{i,j,t} \\ & + \sum_{s=0}^{t_0}\beta_{t,t-s}\epsilon_{m_{i,j,t}=m_{i,j,t-s}}\end{aligned} \tag{9}$$

where $\phi_{i,j,t}$ denotes the log-likelihood ratio

$$\phi_{i,j,t} = \ln[p(f_{i,j,t}|1)/p(f_{i,j,t}|0)]. \tag{10}$$

In practice, considering that we are specifically interested in the case when $m_{i,j,t} = 1$, the likelihood is empirically defined as follows:

$$p(1|1) = \frac{e}{1+e} \quad \text{and} \quad p(0|1) = \frac{1}{1+e} \tag{11}$$

where e denotes the base of the natural logarithm.

This implies that the text blocks most likely locate in the low utility region, which is in line with the design philosophy of the proposed scheme. The low utility regions in the smooth transitions are mainly composed of unstable text blocks and uniformly flat areas, from which the useful information is difficult to extract. As the uniformly flat areas do not need such preprocessing, only the areas with abundant text blocks are taken into account. Moreover, it is reasonable to assign natural image blocks with much lower probability, which further improves the robustness of the algorithm in the scenario of natural video playing. As a result, the probability is transformed into a weighting factor assigned for each block type, which is given by

$$\rho_{i,j} = \begin{cases} 0, & \text{skip block} \\ p(1|f_{i,j,t}), & \text{otherwise} \end{cases}. \tag{12}$$

To finally locate the low utility region, two features representing the horizontal and vertical image block activities, which are obtained by accumulating $\rho_{i,j}$ along the vertical and horizontal directions, are employed as follows:

$$A_{\mathrm{Hor}}(i) = \sum_{j=1}^{H}\rho_{i,j} \quad \text{and} \quad A_{\mathrm{Ver}}(j) = \sum_{i=1}^{W}\rho_{i,j} \tag{13}$$

where H and W indicate the height and width of the frame and are measured in terms of a 16×16 block. As illustrated in Fig. 4, typical

TABLE I
RATING CRITERIONS FOR SUBJECTIVE EVALUATION

Description	Score
The first one is much worse than the second one	−3
The first one is worse than the second one	−2
The first one is slightly worse than the second one	−1
The first one has the same utility as the second one	0
The first one is slightly better than the second one	1
The first one is better than the second one	2
The first one is much better than the second one	3

TABLE II
PERFORMANCE OF THE PROPOSED ADAPTIVE PREPROCESSING SCHEME

Gaussian std	QP	SC_Map				SC_Slideshow				SC_Wordediting				Kimono			
		R_{anc}	R_{pro}	ΔR	S_{avg}	R_{anc}	R_{pro}	ΔR	S_{avg}	R_{anc}	R_{pro}	ΔR	S_{avg}	R_{anc}	R_{pro}	ΔR	S_{avg}
3.5	—	———	———	———	0.214	———	———	———	0.429	———	———	———	1.000	———	———	———	0.000
	22	1832.68	1427.61	22.10%	0.143	599.01	377.58	36.97%	0.286	1176.47	985.33	16.25%	1.071	7236.17	7236.17	0.00%	-0.071
	27	1182.27	981.89	16.95%	0.286	345.98	234.11	32.33%	0.500	845.60	717.19	15.19%	0.786	2873.64	2873.64	0.00%	0.500
	32	721.62	621.85	13.83%	0.214	203.93	149.23	26.83%	0.357	577.48	514.31	10.94%	0.857	1282.30	1282.30	0.00%	0.000
	37	427.71	370.88	13.29%	0.429	126.92	98.44	22.44%	0.429	400.61	371.63	7.24 %	0.786	595.42	595.42	0.00%	-0.286
5.5	—	———	———	———	0.286	———	———	———	0.214	———	———	———	1.071	———	———	———	-0.143
	22	1832.68	1381.56	24.62%	0.357	599.01	357.86	40.26%	0.286	1176.47	932.31	20.75%	1.143	7236.17	7236.17	0.00%	0.357
	27	1182.27	955.01	19.22%	0.286	345.98	222.51	35.69%	0.286	845.60	680.88	19.48%	1.214	2873.64	2873.64	0.00%	0.143
	32	721.62	607.67	15.79%	0.214	203.93	142.07	30.33%	0.357	577.48	489.13	15.30%	1.000	1282.30	1282.30	0.00%	0.214
	37	427.71	364.61	14.75%	0.286	126.92	94.10	25.86%	0.500	400.61	356.88	10.92%	1.071	595.42	595.42	0.00%	0.000

SCV frames are locally analyzed and the distributions of the horizontal and vertical activities are depicted. In each direction, the blocks that are concentrated in the high value ranges indicate rich text blocks, corresponding to a high possibility of being inside a low utility region. Therefore, a threshold that is dependent on the average activity value along each direction is applied to filter the low utility region. With such a procedure, the bound of the natural video region is adaptively detected by capturing the properties of SCVs at the object level.

D. Low Utility Region Processing

Blur is essentially a natural effect which was discovered to be highly relevant to the motion in the perception of HVS [37]. In the literature, various types of blur have been purposely added to the video sequences to enhance the visual experience [38], [39]. However, little has been done in the context of SCV compression. In this work, the circular-symmetric Gaussian filter is applied to process the identified low utility region and smooth out the high frequency information. The reasons of adopting the Gaussian kernel function are manifold. First, as the main task of low utility content is to ensure the smooth transitions in the irregular motions, it is natural to apply the circular-symmetric low-pass filter to further process it. Second, the Gaussian filter is well designed to achieve the redundancy reduction, and has been widely adopted explicitly or implicitly as a HVS channel in preprocessing the signal for similarity comparison [40], [41]. Third, the Gaussian filter is friendly for implementation, enabling its applications in real scenarios [42]. The decoded frames in Fig. 4 with and without preprocessing are illustrated in Fig. 5, from which we can observe that the unstable regions with abundant text will get blurry, such that transitions in SCV playing may become smoother and consume less coding bits simultaneously.

IV. EXPERIMENTAL RESULTS

In this section, the proposed adaptive preprocessing scheme is implemented and validated in terms of the utility based subjective testing. Specifically, four sequences (SC_Map, SC_Slideshow, SC_Wordediting, and Kimono) are used, which cover common screen-sharing scenarios, such as web browsing, office working (slide and word), and natural video playing. It is worth mentioning that the natural video sequence Kimono is treated as a special SCV to examine the robustness of the scheme in the scenario of full screen natural video playing. As such, it is converted into YUV4:4:4 format as well. The test videos are preprocessed and compressed with the SCC extension of HEVC in platform HM16.2-SCM3.1. The common test conditions in the development of SCC extension [43] are used in the test. Low delay B coding structure is employed to simulate the real-time communication environment, and four quantization parameter (QP) values ranging from 22 to 37 are used.

To assess the utility of the preprocessed SCVs, subjective studies were further conducted, in which 14 non-expert subjects (10 males, 4 females) were invited. Specifically, in each trial, a subject is shown a pair of video sequences compressed at the same QP value, and is asked to provide the score based on the guidelines in Table I. Each pair is played in a random order and the obtained score is further processed, such that the value larger than zero indicates that the preprocessed video with the proposed scheme is inferior to the compared one in terms of utility. In each pair, the videos are played on after the other. The subjects were asked to offer their opinions based on the utility resemblance. In other words, the conveyed information is used as the criterion to evaluate how much information loss has been incurred by the proposed preprocessing scheme. Specific instructions and examples were given before the test.

The performance of the proposed adaptive preprocessing scheme is shown in Table II, where R_{anc} and R_{pro} indicate the bitrate generated by the original and prepropcessed sequences. The bitrate variation ΔR is given by

$$\Delta R = \frac{R_{\mathrm{anc}} - R_{\mathrm{pro}}}{R_{\mathrm{anc}}}. \tag{14}$$

To evaluate the utility variation, the average subjective score is calculated, which is denoted to be S_{avg}. The performances with two different Gaussian windows of stds 3.5 and 5.5 are demonstrated. The first row in each Gaussian window corresponds to the subjective tests in the uncompressed case, and the rest rows show the results obtained by using different QPs to compress the SCVs. It is observed that for sequences SC_Map and SC_Slideshow, the utility degradations are ignorable while up to 24% and 40% ΔR can be achieved. For sequence SC_WordEditing, the results show that the preprocessing may slightly decrease the utility of the SCV, given the fact that the high frequency screen operations may produce some flickering artifacts after preprocessing. For the natural video sequence Kimono, as the proposed scheme makes the sequence untouched, the bitrates are identical and the utility variations can be regarded as the random noise in subjective testing. Furthermore, we compute the score variance at each QP point, and the average value of all QP points for each sequence is demonstrated in Table III, which suggests that subjects have a relatively good agreement on judging the utility of SCVs. This further confirms the robustness of the proposed scheme. Moreover, the frame level comparisons of the coding bits are shown in Fig. 6, which demonstrates that the proposed scheme can achieve significant bitrate savings for the frames with high content variations, such that the bandwidth and buffer capacities for future frames that desire more bitrates can be reserved.

TABLE III
SCORE VARIANCE FOR EACH SEQUENCE

Sequences	SC_Map	SC_Slideshow	SC_Wordediting	Kimono
Variance	0.4527	0.4247	0.7484	0.3714

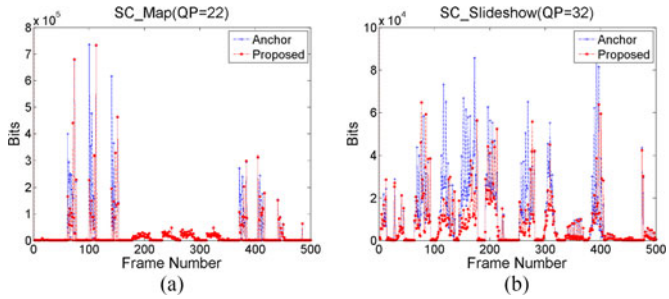

Fig. 6. Frame level comparisons of coding bits. Anchor: original SCV compression; proposed: preprocessed SCV compression (std = 3.5).

TABLE IV
PERFORMANCE COMPARISONS OF THE PROPOSED AND DIVISIVE NORMALIZATION-BASED PERCEPTUAL CODING SCHEMES

Sequences	QP	Std	R_{anc}	R_{pro}	ΔR	S_{avg}
SC_Map	27	3.5	1151.2	981.9	14.7%	0.429
	37	5.5	413.9	364.6	11.9%	0.214
SC_Slideshow	27	3.5	335.3	234.1	30.2%	0.357
	37	5.5	120.8	94.1	22.1%	0.643
SC_Wordediting	27	3.5	827.6	717.2	13.3%	0.786
	37	5.5	399.7	356.9	10.7%	1.000

To further show the advantage of our approach, the proposed scheme is compared with the conventional perceptual video coding algorithm based on divisive normalization [44], [45], which was incorporated into the HEVC SCC extension platform HM16.2-SCM3.1. Subjective testings with the same protocols as in Table II are conducted, and the results are shown in Table IV. We can observe that compared with the perceptual coding algorithm, our approach can still obtain significant rate reduction with little sacrifice on S_{avg}. This is mainly due to the fact that the proposed preprocessing scheme considers the distinct temporal properties of SCVs, such that the coding performance can be further improved from the perspective of temporal masking.

V. CONCLUSION

We propose a utility-driven preprocessing technique to save the coding bits of low utility content in screen content video compression. The novelty of this paper lies in identifying and processing the low utility regions to generate a utility equivalent SCV. Unlike the previous perceptual natural video compression schemes that maintain a constant video quality frame by frame, the proposed approach allows to serve users with the time variant interface of large variations in terms of frame-level quality, such that the perceptual redundancies can be further removed by taking the temporal masking into account. Subjective results demonstrate superior performance as compared to HEVC screen content extension by offering significant rate reduction, while keeping the similar level of utility information. Moreover, the proposed scheme also provides useful evidence that the screen content coding performance can be further improved by taking advantages of the meaningful perceptual cues, and opens up new space in regulating the bitrate for practical SCV rate control in real application scenarios.

ACKNOWLEDGMENT

The authors would like to thank the Associate Editor and anonymous reviewers for their valuable comments that significantly helped us in improving the quality of the paper.

REFERENCES

[1] Y. Lu, S. Li, and H. Shen, "Virtualized screen: A third element for cloud-mobile convergence," *IEEE Multimedia Mag.*, vol. 18, no. 2, pp. 4–11, Feb. 2011.
[2] R. A. Baratto, L. N. Kim, and J. Nieh, "Thinc: A virtual display architecture for thin-client computing," *ACM SIGOPS Operating Syst. Rev.*, vol. 39, no. 5, pp. 277–290, 2005.
[3] "Virtual network computing (VNC)". [Online]. Available: https://www.realvnc.com/, accessed Nov. 2016.
[4] Microsoft, "Remote desktop protocol (RDP)". [Online]. Available: https://msdn.microsoft.com/en-us/library/aa383015(v=vs.85).aspx, accessed Nov. 2016.
[5] C. Lan, G. Shi, and F. Wu, "Compress compound images in H.264/MPGE-4 AVC by exploiting spatial correlation," *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 946–957, Apr. 2010.
[6] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.
[7] C. A. Segall, A. K. Katsaggelos, R. Molina, and J. Mateos, "Bayesian resolution enhancement of compressed video," *IEEE Trans. Image Process.*, vol. 13, no. 7, pp. 898–911, Jul. 2004.
[8] H. Shen, L. Zhang, B. Huang, and P. Li, "A MAP approach for joint motion estimation, segmentation, and super resolution," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 479–490, Feb. 2007.
[9] Q. Yuan, L. Zhang, H. Shen, and P. Li, "Adaptive multiple-frame image super-resolution based on U-curve," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3157–3170, Dec. 2010.
[10] J. Xu, R. Joshi, and R. Cohen, "Overview of the emerging HEVC screen content coding extension," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 50–62, Jan. 2016.
[11] A. Gabriellini, M. Naccari, M. Mrak, and D. Flynn, "Spatial transform skip in the emerging high efficiency video coding standard," in *Proc. Int. Conf. Image Process.*, 2012, pp. 185–188.
[12] A. Gabriellini, M. Naccari, M. Mrak, D. Flynn, and G. Van Wallendael, "Adaptive transform skipping for improved coding of motion compensated residuals," *Signal Process., Image Commun.*, vol. 28, no. 3, pp. 197–208, 2013.
[13] M. Mrak and J.-Z. Xu, "Improving screen content coding in HEVC by transform skipping," in *Proc. 20th Eur. Signal Process. Conf.*, 2012, pp. 1209–1213.
[14] W. Zhu, W. Ding, J. Xu, Y. Shi, and B. Yin, "Screen content coding based on HEVC framework," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1316–1326, Aug. 2014.
[15] D. Kwon and M. Budagavi, "RCE3: Results of test 3.3 on intra motion compensation," in *Proc. Joint Collaborative Team Video Coding-N0205, 14th Meet.*, 2013. vol. 25, pp. 1–8.
[16] L. Zhang *et al.*, "Adaptive color-space transform for HEVC screen content coding," in *Proc. Data Compress. Conf.*, 2015, pp. 233–242.
[17] B. Li, J. Xu, G. Sullivan, Y. Zhou, and B. Lin, "Adaptive motion vector resolution for screen content," in *Proc. 19th Meeting Joint Collaborative Team Video Coding-S0085*, 2014, pp. 1–14.
[18] W. Zhu, W. Ding, J. Xu, Y. Shi, and B. Yin, "Hash-based block matching for screen content coding," *IEEE Trans. Multimedia*, vol. 17, no. 7, pp. 935–944, Jul. 2015.
[19] S. Wang, K. Gu, K. Zeng, Z. Wang, and W. Lin, "Perceptual screen content image quality assessment and compression," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 1434–1438.
[20] K. Gu *et al.*, "Saliency-guided quality assessment of screen content images," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1098–1110, Jun. 2016.
[21] H. Yang, Y. Fang, and W. Lin, "Perceptual quality assessment of screen content images," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4408–4421, Nov. 2015.

[22] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "SSIM-motivated rate-distortion optimization for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 4, pp. 516–529, Apr. 2012.
[23] Z. Chen and C. Guillemot, "Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable-distortion model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 6, pp. 806–819, Jun. 2010.
[24] L. Xu, S. Li, K. N. Ngan, and L. Ma, "Consistent visual quality control in video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 6, pp. 975–989, Jun. 2013.
[25] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May. 2011.
[26] M. Budagavi, "Video compression using blur compensation," in *Proc. IEEE Int. Conf. Image Process.*, 2005, vol. 2, pp. II–882–II-885.
[27] D. M. Rouse and S. S. Hemami, "Natural image utility assessment using image contours," in *Proc. IEEE Int. Conf. Image Process.*, Nov. 2009, pp. 2217–2220.
[28] D. M. Rouse, Y. Wang, F. Zhang, and S. S. Hemami, "A novel technique to acquire perceived utility scores from textual descriptions of distorted natural images," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 2505–2508.
[29] H. Li, B. Li, and J. Xu, "Rate-distortion optimized reference picture management for high efficiency video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1844–1857, Dec. 2012.
[30] S. Wang, S. Ma, S. Wang, D. Zhao, and W. Gao, "Rate-GOP based rate control for high efficiency video coding," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 6, pp. 1101–1111, Dec. 2013.
[31] S. Wang, J. Fu, Y. Lu, S. Li, and W. Gao, "Content-aware layered compound video compression," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2012, pp. 145–148.
[32] H. Shen, Y. Lu, F. Wu, and S. Li, "A high-performanance remote computing platform," in *Proc. Int. Conf. Pervasive Comput. Commun.*, 2009, pp. 1–6.
[33] D. Miao, J. Fu, Y. Lu, S. Li, and C. W. Chen, "High frame rate screen video coding for screen sharing applications," in *Proc. IEEE Int. Symp. Circuits Syst.*, Jun. 2014, pp. 2157–2160.
[34] B. O. Christiansen and K. E. Schauser, "Fast motion detection for thin client compression," in *Proc. Data Compression Conf.*, 2002, pp. 332–341.
[35] S. Wang, K. Gu, S. Ma, and W. Gao, "Joint chroma downsampling and upsampling for screen content image," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 9, pp. 1595–1609, Sep. 2016
[36] Z. Pan, H. Shen, Y. Lu, S. Li, and N. Yu, "A low-complexity screen compression scheme for interactive screen sharing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 6, pp. 949–960, Jun. 2013.
[37] T. L. Harrington and M. K. Harrington, "Perception of motion using blur pattern information in the moderate and high-velocity domains of vision," *Acta Psychologica*, vol. 48, no. 1, pp. 227–237, 1981.
[38] M. Potmesil and I. Chakravarty, "Modeling motion blur in computer-generated images," *ACM SIGGRAPH Comput. Graph.*, vol. 17, no. 3, pp. 389–399, 1983.
[39] G. J. Brostow and I. Essa, "Image-based motion blur for stop motion animation," in *Proc. 28th Annu. Conf. Comput. Graph. Interactive Technol.*, 2001, pp. 561–566.
[40] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[41] G. Zhai, A. Kaup, J. Wang, and X. Yang, "Retina model inspired image quality assessment," in *Proc. Visual Commun. Image Process.*, 2013, pp. 1–6.
[42] S. Khorbotly and F. Hassan, "Recursive implementation of gaussian filters with switching and reset hardware," in *Proc. Int. Midwest Symp. Circuits Syst.*, 2013, pp. 1399–1402.
[43] H. Yu, R. Cohen, K. Rapaka, and J. Xu, "Common test conditions for screen content coding," in *Proc. Joint Collaborative Team on Video Coding-S1015,* 2014, pp. 1–6.
[44] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "Perceptual video coding based on SSIM-inspired divisive normalization," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1418–1429, Apr. 2013.
[45] A. Rehman and Z. Wang, "SSIM-inspired perceptual video coding for HEVC," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 497–502.
[46] Z. Pan, H. Shen, Y. Lu, S. Li, N. Yu, "A low-complexity screen compression scheme for interactive screen sharing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 6, pp. 949–960, Jan. 2013.

Shiqi Wang (M'15) received the B.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2008, and the Ph.D. degree in computer application technology from Peking University, Beijing, China, in 2014.

He was previously a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. He is currently a Research Fellow with the Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore. He has proposed more than 30 technical proposals to ISO/MPEG, ITU-T, and AVS standards.

Xinfeng Zhang received the B.S. degree in computer science from the Hebei University of Technology, Tianjin, China, in 2007, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2014.

He is currently a Research Fellow with Nanyang Technological University, Singapore. His research interests include image and video processing and compression.

Xianming Liu received the B.S., M.S., and Ph.D degrees in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 2006, 2008, and 2012, respectively.

He is an Associate Professor with the Department of Computer Science, HIT. In 2011, he spent half a year with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, as a Visiting Student, where he then worked as a Postdoctoral Fellow from December 2012 to December 2013. Since 2014, he has been working as a Project Researcher with National Institute of Informatics, Tokyo, Japan. He has authored or coauthored more than 40 International Conference and Journal Publications, including top IEEE journals such as the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, and the the IEEE TRANSACTIONS ON MULTIMEDIA; and top conferences, such as Computer Vision and Pattern Recognition, International Joint Conference on Artificial Intelligence, and Data Compression Conference.

Dr. Liu was the recipient of the IEEE ICME 2016 Best Student Paper Award.

Jian Zhang (S'12–M'13) received the B.Sc. degree in mathematics from the Harbin Institute of Technology (HIT), Harbin, China, in 2007, and the M.Eng. and Ph.D. degrees in computer science and technology from HIT in 2009 and 2014, respectively.

He is currently working as a Postdoctoral Fellow with the Institute of Digital Media, School of Electronics Engineering and Computer Science, Peking University, Beijing, China. His research interests include image/video compression and restoration, compressive sensing, sparse representation, and deep learning.

Dr. Zhang was the recipient of the Best Paper Award and Best Student Paper Award at the IEEE International Conference on Visual Communication and Image Processing in 2011 and 2015, respectively.

Siwei Ma (S'03–M'12) received the B.S. degree from Shandong Normal University, Jinan, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005.

From 2005 to 2007, he was a Postdoctorate with the University of Southern California, Los Angeles, CA, USA. He is currently a Professor with the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing, China. He has authored or coauthored more than 100 technical articles in refereed journals and proceedings in the areas of image and video coding, video processing, video streaming, and transmission.

Wen Gao (S'87–M'88–SM'05–F'09) received the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991.

From 1991 to 1995, he was a Professor of computer science with the Harbin Institute of Technology, Harbin, China, and a Professor with the Institute of Computing Technology of the Chinese Academy of Sciences, Beijing, China. He is currently a Professor of computer science with Peking University, Beijing, China. He has authored or coauthored extensively, including five books and more than 600 technical articles in refereed journals and conference proceedings in the areas of image processing, video coding and communication, pattern recognition, multimedia information retrieval, multimodal interface, and bioinformatics.

Prof. Gao served or serves on the Editorial Board for several journals, such as the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT, the EURASIP *Journal of Image Communications*, the *Journal of Visual Communication and Image Representation*. He chaired a number of prestigious international conferences on multimedia and video signal processing, such as IEEE ICME and ACM Multimedia, and also served on the advisory and technical committees of numerous professional organizations.