Exploiting spatio-temporal characteristics of human vision for mobile video applications

Rashad Jillani, and Hari Kalva
Department of Computer Science and Engineering
Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, USA

ABSTRACT

Video applications on handheld devices such as smart phones pose a significant challenge to achieve high quality user experience. Recent advances in processor and wireless networking technology are producing a new class of multimedia applications (e.g. video streaming) for mobile handheld devices. These devices are light weight and have modest sizes, and therefore very limited resources – lower processing power, smaller display resolution, lesser memory, and limited battery life as compared to desktop and laptop systems. Multimedia applications on the other hand have extensive processing requirements which make the mobile devices extremely resource hungry. In addition, the device specific properties (e.g. display screen) significantly influence the human perception of multimedia quality. In this paper we propose a saliency based framework that exploits the structure in content creation as well as the human vision system to find the salient points in the incoming bitstream and adapt it according to the target device, thus improving the quality of new adapted area around salient points. Our experimental results indicate that the adaptation process that is cognizant of video content and user preferences can produce better perceptual quality video for mobile devices. Furthermore, we demonstrated how such a framework can affect user experience on a handheld device.

Keywords: Human Vision, H.264, Scalable Video Coding, Mobile Video, Saliency, Perceptual Quality

1. INTRODUCTION

H.264/AVC has been designed to provide a technical solution appropriate for broadcast, storage device, and conversation service over wireless networks, VOD or multimedia streaming services. Recent developments in video encoding standards such as the H.264 have resulted in highly efficient compression [1]. Experimental results show that H.264's coding performances overcome MPEG-4 at low bit rates [2]. The current video compression algorithms also support multiple reference frames for motion compensation. As the number of reference frames is increased, the complexity increases proportionally.

Resource constrained devices typically manage the complexity by using a subset of possible coding modes thereby sacrificing video quality. This quality and complexity relationship is evident in most video codecs used today. Most H.264 encoder implementations on mobile devices today do not implement the standard profiles fully due to high complexity. For example, Intra 4x4 prediction is usually not implemented due to complexity. In [3] complexity is reduced by using machine learning algorithm to predict Intra MB coding. In addition, several rate control techniques have been proposed to achieve balance among efficiency and complexity; e.g., mode selection and post-quantizer control. Most content available today is not developed for mobile devices. When such content is delivered to mobile devices scaling down to the resolution of the device is the most common solution. When down scaling is employed mobile devices use techniques such as letterboxing to maintain the aspect ratio. Using the valuable display space on mobile devices for letterboxing is not an effective use of resources. Systems adopting content to mobile devices should exploit the available display fully. The classic problem of quality vs. bitrate tradeoffs should be balanced. Proper rate control can significantly improve the performance by reducing time-out effects, packet loss thus enhance the video quality and guarantee quality of service (QOS) [4]. The rate control algorithms that appear in literature primarily focus on achieving target bitrate and do not consider the content of video and human perception. Similar work has been done in model based coding where content modeling drives bit allocation [7].

There exist technologies for video streaming in networks with fluctuating bandwidth that define the regions of interest in a video. MPEG-4 selective enhancement [5] is used in the enhancement layer of MPEG-4 FGS (Fine Grained Scalability) in order to stream better quality of video within selected image regions. However, MPEG-4 selective enhancement does not provide quality improvement for the base layer. FGS-MR video encoding [9] uses MR (Multi-

Applications of Digital Image Processing XXXI, edited by Andrew G. Tescher,
Proc. of SPIE Vol. 7073, 70730N, (2008) · 0277-786X/08/$18 · doi: 10.1117/12.796746

Resolution) frames based on MR masks to improve the rate distortion performance. However, this process introduces additional complexity in the encoding process since canny edge detection necessary for FGS-MR is a resource consuming process.

In this paper we take into account the perceptual quality of the image and the areas in that image that receive most user attention. We try to improve the end user experience by coding only the regions of interest at the same given bitrate but at higher quality. It provides us the opportunity to concentrate more on the area in the image where more likely user attention will be focused. By paying more attention to this area we enhance the perceptual quality of that area while neglecting and sometimes sacrificing the content in the areas where most likely user attention will not be focused. Our technique reduces the bitrate while improving the quality of the selected regions. Related work in this area includes video codecs that have been proposed to address these problems inherent to any foveated video compression technique (e.g., encode high-priority regions first, then lower-priority regions, in a continuously-variable-bitrate encoding scheme [8]).

The rest of the paper is organized as follows. Section 2 discusses the background information and the details of the mobile devices and services used today. The proposed approach to mobile video adaptation using region of interest (ROI) and computational models of human attention are discussed in Section 3. Experiments and results are presented in Section 4 and Conclusions in Section 5.

2. BACKGROUND AND MOTIVATION

With the availability of powerful mobile devices, video-based entertainment content is no longer limited to the traditional living room environment. With the introduction of smaller, more portable devices in consumer market and the availability of a variety of A/V (Audio/Video) content, consumers now have multitude of choices for viewing their favorite entertainment whenever they want. These new devices enable them to be more mobile with their entertainment as they are no longer bound to the living rooms.

This new trend in mobility can be experienced in many ways, whether through additional multimedia options throughout the house or via miniaturized entertainment devices for experiences in the outdoor environment. In both cases, this newfound flexibility offers many new business opportunities. However, there are a number of technical challenges inherent in each link of the entertainment chain that must be overcome in order to successfully adopt these devices. These include, but not limited to, the creation of content and preparing it for distribution; transferring or copying content, and playing or rendering the content on various mobile devices.

Everyday, consumer marketplace is introduced with an ever-increasing number of mobile devices and in order to meet the needs of consumers and ultimately capture the market, electronics manufacturers are striving to develop a broad range of new concepts. We present the explanation of each type of mobile device and its characteristics [10].

2.1 Devices

Following is a brief description of mobile devices' types classified based on features and functionality:

2.1.1 Mobile phone/handsets

Traditional handsets now come with a variety of new features that are powered by a number of operating systems. There is a large and varied spectrum of devices available in fewer than two primary classifications: (i) Feature Phones which integrate numerous entertainment options while supporting voice communication. And (ii) Smart Phones which provide broadband applications and other data processing functions.

a) Feature Phone

These are typically "candy bar" or "flip" styled handhelds with a standard numeric keypad. They feature menus that can support such media streaming applications as music, games, text messaging, and digital video/imaging via a built-in camera.

b) Smart Phone

These phones employ the mobile operating systems such as Symbian, Palm, WindowsMobile, RIM, MacOS. Examples of these phones are Blackberry, Treo, Motorola Q, SonyEricsson and iPhone. All of these have some type of QWERTY keyboard in one form or another.

c) Java and/or Brew Phone

Java can be used to support simple games and basic media players for multiple brands of handsets. Since the processing power of each handset varies, Java applications need to be customized for each device. In contrast, Brew is a cross platform environment that is more efficient than Java and utilizes the full processing power for a limited number of handsets. The devices for basic 3D games, video recording, user-generated content and video conferencing are on the horizon.

2.1.2 Portable media players

Comprised of ubiquitous MP3 devices, this category is now dominated by the Apple iPod. Although countless other media players are now available, none has established anywhere near the level of market penetration or public awareness as the iPod.

These players offer two types of connectivity options.

a) Tethered

These portable devices have required USB, iLink (IEEE 1394) or some other type of cabled connection to a host Mac or PC in order to transfer content. Recently a number of devices have been developed that can also achieve connectivity using wireless technology. Examples of these devices include: iPod, iPhone, PlayStation Portable (PSP) and Zune.

b) Un-tethered

Other handheld media devices are now coming to market that no longer require a host computer. These devices can communicate via Wi-Fi directly with the retail store to obtain media content. Additionally, the use of the latest SOC ("System On Chip") technology allows digital and analog DVR capability while reducing costs. Examples of these devices include SanSa Connect and Archos Series 5.

2.1.3 Notebooks/laptops/ultra-portables

These devices are full function PCs or Macs that support a variety of Internet media services and websites. Several of the high-speed services that are listed in the next section may be bundled with these devices or enabled by plugging in a small USB connector. Most of these devices also incorporate some form of Wi-Fi and/or wide area network (WAN) capability.

"Ultra Mobile PCs" (UMPCs) are another new miniaturized device that runs on the latest Windows operating system. Many of these devices also support short range Bluetooth wireless capability in addition to built-in Wi-Fi or Wide Area Network (WAN).

2.1.4 Portable digital receivers

Like television, radio is also transitioning to the digital age and new handheld receivers are now appearing that support audio content delivery. Some examples include satellite radio devices from Sirius and XM; as well as "inband on-channel" (IBOC) digital terrestrial receivers from HD Radio. Many of these devices also support mobile video and/or GPS functionality.

2.2 Content Distribution to Mobile Devices

A/V contents can be distributed via a number of platforms:

2.2.1 Online video stores

Video content can be purchased on the Internet through various online retail stores that offer movies, broadcast TV content and music videos to consumers. Some of this content is specifically aimed at portable devices, while other content requires some adaptation for portable use. Content selection is growing rapidly, although not as widely available as DVDs. Examples of online video stores include Apple's iTunes Store, Wal-Mart's Video Download Store, Amazon's Unbox, and AOL-video.

2.2.2 Video streaming and/ or downloading through mobile carrier network

Most cell phone operators offer their own video services for a subscription fee that is added to the consumer's monthly bill. Most of these services stream video directly to the phone due to limitations in storage capacity, although some offer

download capability. Mobile content encompasses broadcast TV shows, news, weather and sports. Examples of mobile carrier network services include MobiTV, Verizon Wireless' V CAST, Sprint Movies and AT&T's Cellular Video.

2.2.3 Mobile TV

Mobile TV, like standard TV, offers one or more channels that deliver content on a pre-determined schedule. In contrast, video streaming enables the selection of content on-demand. There are two types of Mobile TV; the first is coupled to a cell phone service and requires a special type of cell phone with Mobile TV capability. Examples of these services include Verizon Wireless' V Cast Mobile TV, and Sprint's Pivot (in cooperation with Comcast). Crown Castle and Qualcomm's MediaFlo also wholesale and license content to several carriers.

The second type of service is tied to Satellite radio operators that have added TV content to their programming and require separate compatible devices. Both Sirius and XM satellite radio operators have announced mobile satellite TV services.

2.2.4 Retail download kiosks

Downloading from retail kiosks is also being developed for portable media players and devices. This is already offered for portable game devices like Sony's PlayStation Portable (PSP) and Nintendo's Gameboy DS. Other companies working on the same concept for A/V content include Microsoft (Zune downloads from the TableTop PC) and PortoMedia.

2.2.5 Time and place shifting enablers

Recently new types of mobile devices have been introduced that enable users, while traveling, to download or stream video content stored at home over the Internet.

Examples of such enablers include Sling Media's Slingbox, Sony's LocationFree TV, Orb Networks' Orb, TiVo ToGo and various PC-based Digital Video Recorder applications. Additionally, with the advent of affordable high efficiency video encoders, a new generation of mobile devices are also available that can be taken anywhere on the go.

2.3 Content Adaptation

Following are the basic methods with which video can be adapted for playback on mobile devices.

- **Scaling**: the process of converting a video signal from one resolution to another. If the source material has a 1920x1080 resolution, the source material must be scaled or rendered for display on a portable device with a lower resolution (640x480).
- **Color space conversion**: the process of translating from one color space to another. Different color spaces are better for different applications; and some types of equipment has limiting factors that dictate the size and type of color space that can be used.
- **Transcoding**: the process of converting one digital codec (compression/decompression) of a movie into another. The original format is first decoded and then re-encoded into the desired format. By way of example, a video source using MPEG2 will need to be transcoded into a format such as Windows Media Video (WMV) or AVC. Transcoding may or may not be transparent (lossless).
- **Transrating**: the process of transferring video from one bit rate to another. This is often required because some video formats and portable devices can only support specific data rates. This process can also impact the size of the data file (the lower the bit rate, the smaller the file size).
- **Transcrypting**: the process of converting from one encryption scheme (often referred to as DRM) to another. Portable devices may often be limited to specific DRM formats, For example, Apple iPods use Fairplay DRM, whereas Microsoft Zune devices use Windows Media DRM. In order for music content to play properly on these devices, they must be transcrypted into the appropriate format.

While viewing the video contents on mobile devices, a number of factors affect the quality of user experience. These include data file size, bitrate, compression schemes, and overall resolution of the display. Service providers take into consideration these factors and try to maximize the quality of video for mobile devices e.g. YouTube (www.youtube.com) is optimized to play videos at low bit rate and a low resolution for 320X240. Table 1 describes resolution of some of the mobile devices available in the market.

Table 1. Mobile devices' resolutions

Mobile Device	Resolution	Type
Apple iPhone	480x320	Smart phone
Apple iPod	320x240	Portable media player
Microsoft Zune	320x240	Portable media player
Samsung Instinct	240x432	Smart phone
Samsung BlackJack II (i617)	320x240	Smart phone
Samsung SCH-a890	176x220	Feature phone
Sanyo MM-5600	240x320	Smart phone
Black Berry Curve 8320	320x240	Smart phone
HP iPAQ PDA	480x640	PDA

Content providers have to balance several technical issues to deliver an optimal consumer experience while delivering A/V content to mobile devices. While better quality is their target, that target is limited by several other factors such as bitrate, A/V codec selection, screen resolution and more.

Generally, better quality means higher bitrate and larger A/V file for delivery to consumer. Similarly, larger screen resolution means larger file size. Of course, size of A/V contents affects the download time on consumer device. Although it is possible to make the downloading A/V content a background task, the performance of streaming A/V content depends upon the available bandwidth of the user's device connection. Therefore, the emphasis is put upon the reduction of the size of A/V content from the providers resulting in the limited bitrate and resolution at the cost of better streaming quality for the user. The selection of video compression scheme can also impact the bitrates and quality metrics, as certain compression schemes such as H.264 can achieve higher quality at lower bitrates, thus maximizing bandwidth.

The current solution for content adaptation is down-scaling for most of the content providers. A/V contents can be prepared either offline (i.e. preprocessed before delivery) or on the fly (i.e. processed in real time based on the characteristics of the target device). As a result, the quality for the target device may not be suitable for viewing the contents if the original contents are prepared with the HD or SD services in mind. This paper discusses different approaches to scale and adapt the video contents for the miniaturized devices limited by the bandwidth and screen resolution. The goal is to improve the end user experience while possibly deviating from what the original content creator had produced.

3. METHODOLOGY

In this section, we will analyze different architectures that are proposed to address the problem of delivering A/V contents to mobile devices.

3.1 H.264/SVC ROI Framework

The Scalable Video Coding is the extension of the H.264/AVC video compression standard. H.264/AVC was developed jointly by ITU-T and ISO/IEC JTC 1. These two groups created the Joint Video Team (JVT) to develop the H.264/AVC standard. H.264/AVC has been very strongly embraced by the industry and is now being developed in virtually all new existing applications of digital video technology[14]-[19]. The purpose of SVC [11] is to extend the capabilities of the H.264/AVC design to address the needs of applications to make video coding more flexible for use in highly heterogeneous and time-varying environments. Instead of multiple encodings of each video source so as to provide the optimized bit stream to each client, scalable coding provides a unique bit stream whose syntax enables a flexible and low complexity extraction of the information so as to match the requirements of different devices and networks. Such an application scenario is shown in figure 1.

The basic concept of SVC is to enable the creation of compressed bit stream that is comprised of partial bit streams and it enables the transmission and decoding of partial streams. SVC structures the data of a compressed video bit stream into *layers*. The *base layer* is an ordinary H.264/AVC bit stream, while one or more *enhancement layers* provide improved quality for those decoders that are capable of using it. The SVC bit stream always consists of a lower-resolution/quality version of the video signal (base layer) and the full-resolution/quality video (enhancement layer). With a little bit of increase in decoder complexity relative to a single layer H.264/AVC, SVC provides network friendly scalability at a bit stream level. Furthermore, SVC makes it possible to rewrite the fidelity-scalable SVC bit streams to single-layer H.264/AVC bit stream lossless. The target applications of the SVC extension of H.264/AVC vary from video conferencing as well as for mobile to high-definition broadcast and professional editing applications.

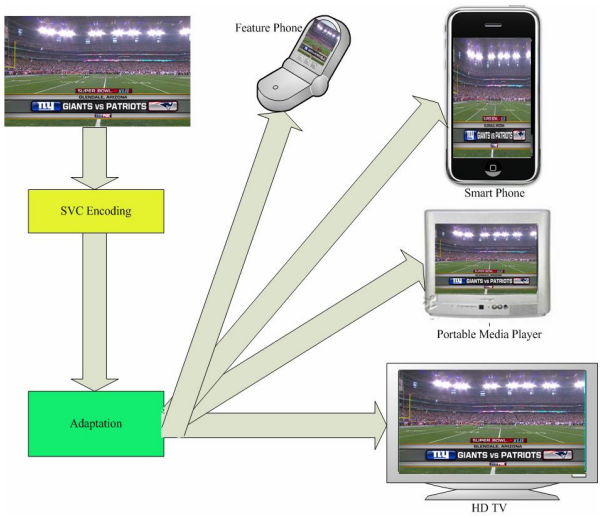

Fig. 1. H.264/SVC encoded delivery and streaming for different target devices after adapting the incoming SVC stream.

Three fundamental types of scalability were enabled in H.264/SVC. The first is *temporal scalability*, in which the enhancement layer provides an increase of the frame rate of the base layer. To a large extent this was already supported in the original H.264/AVC standard—but new supplemental information has been designed to make such uses more powerful. The next is *spatial scalability* (or resolution scalability), in which the enhancement layer offers increased picture resolution for receivers with greater display capabilities. Finally, there is *quality scalability*, sometimes also referred to as SNR or fidelity scalability, in which the enhancement layer provides an increase in video quality without changing the picture resolution.

H.264 SVC extension enables provisions for ROI based adaptation using Flexible Macroblock Ordering (FMO). This technique is formally identified as interactive ROI (IROI) scalability [12] [13]. A ROI typically is a region within the video pane containing visual information that is more interesting than the other parts of the video pane. IROI scalability means that the encoder generates a scalable bit-stream without knowledge about the possible ROIs (coding process). Then, the ROI can be selected on-the-fly by the user (selection process). Because there are no predefined fixed ROIs, the user can select his own ROIs at any moment. This information about the selected ROIs is sent to an extraction engine in order to customize the scalable bitstream and to generate an adapted bitstream (extraction process). Finally, the adapted bitstream can be decoded such that the ROI has a higher visual quality than the surrounding background (decoding process). In the case of multiple ROIs, they can be equally important or they might have different levels of importance. IROI scalability can be used to extract a sub-stream that defines ROI. However, this method does not provide sequence level cropping and if required, sequence level cropping must be performed offline by putting some restrictions in temporal prediction scheme of H.264.

There are two types of approaches that have been used for content adaptation based on IROI: H.264/AVC FMO and the XML-driven content adaptation framework [21] [22] [23]. A similar IROI scalability technique has been proposed by Lambert et al [9]. In H.264, the FMO tool has been utilized to code an ROI into Network Abstraction Layer (NAL) Units (NALUs), and therefore this technique also suffers from the limitations of IROI scalability of the SVC extension of H.264/AVC. In addition ROI identification in these methodologies is based on user preferences or profiles that define the coordinates of ROI explicitly. Therefore, this framework could easily be classified as user driven. However, the identification and selection of possible ROIs, if automated, can significantly improve the utility of IROI features.

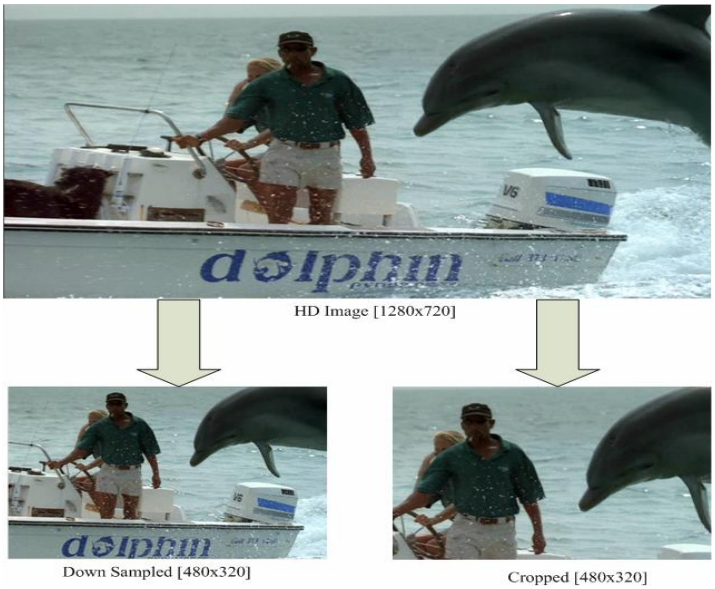

Figure 2 shows the comparison of the perceptual quality of down-sampling and IROI based cropping. Obviously, cropping produce better perceptual quality because unlike down-sampling it preserves the fine details of the selected area of the image. But the down-sampled HD images targeted for mobile devices do not produce good perceptual quality on the short display screens and the detail of the whole frame is obscured. On the other hand, cropping may produce good results but selection of ROI is the major problem in this case. Moreover, the overhead of using FMO is significant as well as the sizes of XML descriptions for ROIs. These shortcomings make this framework more complex for mobile devices. However, this framework may be better suited for surveillance systems.

Fig. 2. Comparison of down sampling and cropping from HD format to 480x320.

3.2 Saliency Based Framework

Simply scaling images reduces the size of important features. If the image contains a single important feature, one can crop the image and scale it to fit. Images with multiple important features present a more challenging case for adapting it for mobile video applications. In these cases, you might waste valuable image area with unimportant regions between important features. For many images, the key content is a small set of several objects. To effectively display the images on small displays, you must display the objects at a sufficient size so that they are easy to recognize. Other objects in the image, as well as the precise relationships between objects, are less important. Our automatic saliency based ROI detection method helps generate these small images. Our goal is to effectively display small images by preserving the recognizability of important image features while concentrating to important regions in given images. The basic premise of this approach is that if we can identify the important objects in a given image, we can display them at a higher resolution/quality in the target image so that they are more recognizable. Such cropping necessarily comes at the expense of other less important parts of the image. In contrast to traditional resizing and down-sampling techniques, a saliency detection method identifies the important objects by displaying them in their actual size, making them easier to recognize on small display screens.

Recent research has shown that a bit stream of video content is not only a combination of 1s and 0s, it also contains semantic information. On the other hand, rigid psychological experiments have confirmed that the Human Visual System (HVS) has a property of non-uniform spatial resolution, with respect to the attentional focus in sight, which is called "foveation" [20]. Based on this property

- People will be unaware of quality degradation in periphery regions of a video frame.
- People will expect better quality in regions that receive attentional focus than that of the other regions.

Attention refers to the ability of one human to focus and concentrate upon some visual or auditory 'object'. According to human psycho-visual nature, people often pay more attention to some salient stimuli in videos. Itti et al [9] proposed a neurobiological model of visual attention which predicts regions that should be given high priority during compression, that is, regions of high *cognitive importance* for scene understanding and thus improved the compression ratio. A snapshot of the saliency map at each frame directly determines the priority to be given to every spatial location in the frame. It also concludes that there are few regions in an image that attract the attention of human eye based on various parameters. We based our ROI detection technique on this conclusion and use it to have a general idea about the salient regions in an image. We use approximation of this model to calculate the salient points in an image. This model is also

called bottom-up saliency-based visual attention model. A detail discussion about this model is out of the scope of this paper and has been presented elsewhere [9].

In a number of video-based application scenarios, the ROI can be chosen right away at encoder side and this information can be used to encode the bitstream targeted for mobile devices to be streamed and ultimately played efficiently. Due to the limitation in the availability of wireless bandwidth and small display area of mobile devices, the task of keeping the size of bitstream small and complexity of encoding/decoding manageable is more difficult than it is for other network based video applications. Saliency based model finds salient locations in an image based on the intensity, colors and orientations of objects and the image is encoded based on these salient regions with minimal intervention from user based on a set criteria. Based on different saliency points, a high resolution image containing different regions of the source image can be cropped to the lower resolution target image for mobile devices.

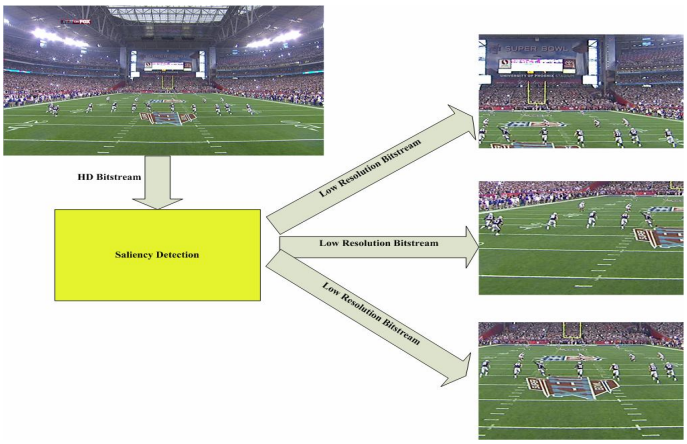

Fig. 3. Process of finding salient regions and encoding the incoming bitstream based of points of saliency.

Figure 3 shows an example of the implementation of saliency based approach. Source image is from a 720p bitstream while saliency detection mechanism finds points of saliency in the incoming image and encodes the bitstream based on this information in combination with the target resolution of the device and the user experiences a better perceptual quality of the selected region based on saliency. It must be noted that the saliency points can be distinguished from each other based on their intensity. Therefore, if there are multiple salient points, we can assign preference order with them and along with their location in the source image; and this information can be used to encode different parts of the source image as shown in figure 3.

Figure 4 shows an exhaustive example of this process where an incoming HD bitstream of high bit-rate is going through the saliency detection process and then adapted according to the specified varying requirements based on the targeted networks and devices with different resolutions and bit-rates. It must be noted that the output images is neither the result of resizing input images nor just cropping at random regions. Saliency detection process is capable of producing the output images that are taken from the same input images but produces different displays for the same output bit-rate and resolution. At a given time, the focus of saliency points which is based on the intensities may occur at multiple regions in an image. Saliency detection process determines the appropriate region according to the given preferences and prepares the output image after adapting it to the given conditions of output bitrate and resolution. In this process, it makes ensure to include the single or multiple saliency points in the output image and adapts the orientation accordingly.

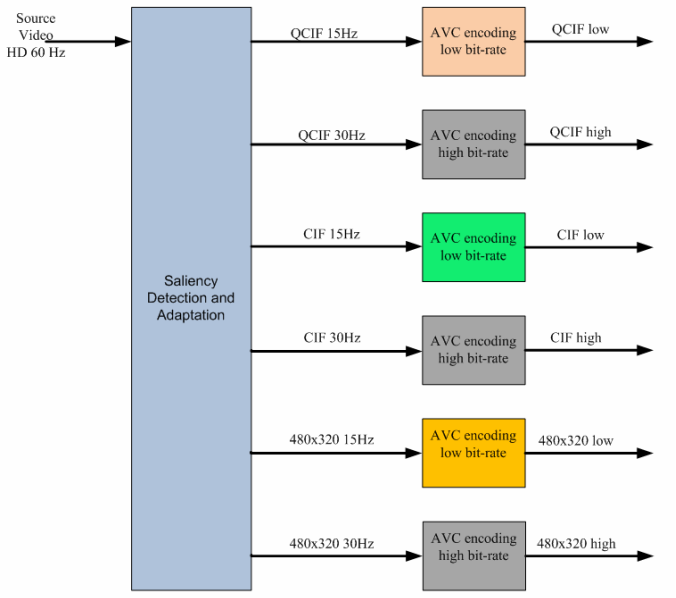

Fig. 4. Exhaustive example of saliency detection and adaptation of incoming high bit-rate HD sequence into output sequences of low bit-rates and high bit-rates of QCIF, CIF and 480x320 resolutions.

The selection of saliency regions also depend upon contents' type of the input video images. For example, sports and movies are two completely different categories. Saliency detection may not always be a consistent across all types of video contents but there is ongoing research in this field to improve the detection process for salient regions. Sometimes the regions detected by the saliency detection process might not be important in the final output image. It is interesting to note, however, that this does not contradict the fact that these "unimportant" features are visually conspicuous.

4. EXPERIMENTAL SETUP AND RESULTS

In order to have some insight in the performance and the consequences of the proposed architecture, a series of tests was set up. The measurements include the impact of the adaptation process on the bitstream and on the receiving decoder. Also an assessment of the performance of the overall adaptation framework is given. To evaluate the performance of our saliency detection framework for exploiting spatio-temporal characteristics of human vision, we performed our experiments on a variety of video contents divided into different categories. Input sequences were selected from 720p and 1080p resolutions. Results are compared for the target resolution of 480x320. This choice was inspired by the display capabilities of most of the smart phones including the popular Apple's iPhone have this resolution. To compare our results, we performed four different types of experiments. At least one bitstream from each category is subjected to 4 different kinds of methods as explained in figure 5.

Table 2. Bitstreams/sequences used for performance evaluation.

Sequence/Bitstream	Category	Frames	Format
Spider Man 3	Movie	500	HD 720p (1280x720)
Dolphin	Documentary	500	HD 720p (1280x720)
NasCar	Sports	500	SD 480p (720x480)
ParkingLot	Surveillance	500	SD 480p (720x480)

In figure 5, an overview of the saliency based framework is given as employed in this paper. We compared the results of four different methodologies in terms of PSNR, bitrate and perceptual quality. The sequences were chosen based on different categories of movies, documentaries, sports, surveillance and general. The input sequences were of either HD or SD quality. For surveillance category, we recorded our own sequence by using HD camcorder. For each bitstream, the corresponding salient locations have to be generated in order to find the salient regions in the given bitstream. The generated saliency regions are subject of a transformation performed by an adaptation engine. This engine typically takes a point (XY coordinates) or a couple of points as input to determine the most interesting part of the video sequence. In our experiments, ROI is either defined around a single pixel coordinates or encompassing a couple of pixel coordinates otherwise known as salient points. Finally, the adapted bitstreams are created by using the modified version of H.264/AVC reference software (JM 13.2). The encoding was done that conforms to Baseline Profile of H.264.

Other relevant encoding parameters are, a constant Quantization Parameter (QP) of 28 and coding pattern of {IPPP…}. All test runs that are mentioned in this results section were performed on a Pentium IV 2.8 GHz Core 2 duo machines with 2GB RAM, running Windows XP SP2. Some properties of the resulting bitstreams are summarized in Table 2. The impact of the adaptation process on the PSNR and perceptual quality of the bitstreams is given in table 3 and table 4.

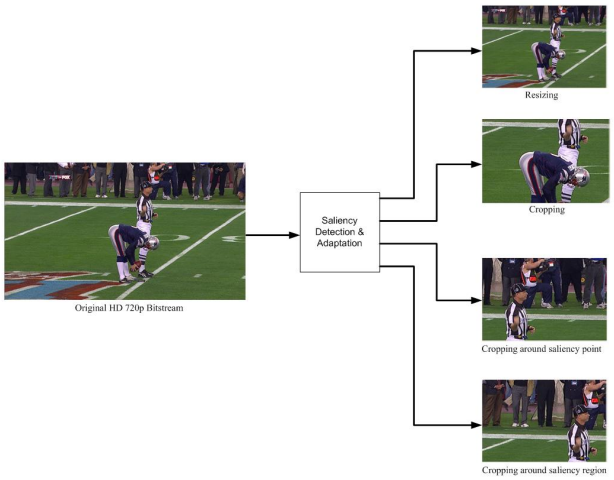

Fig. 5. Saliency based framework for content adaptation. Four resizing, (ii) cropping, (iii) cropping arouns saliency point and saliency region

Table 3. PSNR of bitstreams at bitrate 256kbps.

Sequence	Down Sampling	Centred ROI	Single Point Saliency Based ROI	Saliency Region Based ROI
Spider Man 3	45.073	46.483	47.538	48.691
Dolphin	31.333	34.96	36.373	36.937
NasCar	28.324	28.715	27.979	27.965
ParkingLot	36.7	36.7	36.838	35.572

Table 4. Perceptual rating (Mean Opinion Score) of the bitstreams on scale of 1-5, (1 = Bad, 2 = Poor, 3 = Fair, 4 = Good, 5 = Excellent)

Sequence	Down Sampling	Centered ROI	Single Point Saliency Based ROI	Saliency Region Based ROI
Spider Man 3	3	3	2	2
Dolphin	3	3	4	5
NasCar	3	4	3	4
ParkingLot	4	4	4	4

4.1 Down Sampling

This process when applied to a bitstream, applies down-sampling according to the desired output resolution. In this process the image is essentially miniaturized and if the source image is of type HD or SD, the contents of the image are compressed and the users have usually difficulty in viewing the contents of the video on their small devices' display screens. Figure 6 shows the result of this process. As shown in the table 3, the PSNR of the down-sampled bitstream is usually lower than other adaptation methods described and explained below. Since the whole image is compressed into a smaller size, the details of the video content remain in the smaller though the relative sizes of objects is much scaled down. Therefore, perceptual quality also lies in the middle as shown in table 4. This method is a quick way to reduce the bandwidth requirement. Figure 6 shows some scenarios of the implementation of this approach.

4.2 Centered ROI

This process crops the image from the center according to the output resolution without changing the size of the objects in that area. Video contents where usually the action remains in the center area of the image are ideal for this type of adaptation. After the adaptation of the video contents by employing this process, the perceptual quality of the cropped region remains the same as that of before. The only downside in this process is that all the peripheral details in the source video contents are lost. As compared to the down-sampling, this process maintains the quality of the sequence in the form of PSNR and in most case PSNR in this method is usually is larger than the down-sampling approach. Subjective tests for this approach (Table 4) show that the perceptual quality is somehow up to the viewers' "Fair" level of expectations. Video contents of the categories such as News, Video Conferencing are ideal candidate for this type of adaptation process. Since, usually there are no moving objects in this type of video contents and also the objects appear almost in the middle area of the display screen, therefore, centred ROI is the best adaptation process for these types of video contents. Figure 7 explains the application of this approach.

4.3 Single Point Saliency Based ROI

This process determines a point of saliency in the given image, and based on the coordinates of this points configures the coordinates of the output image such that the point of saliency lies in the middle of the output image. The process of determining the point of saliency is the bottom-up approach, since unlike object identification (top-down processing) the saliency of a given location is defined by the stimulus driven mechanism derived by the intensity, color and orientation of the objects. This approach is based on the idea that usually there is a single most salient point in a given image that draws the viewer's attention while watching the video contents and while watching video contents on mobile devices, the viewer is much more concentrated on a very small area that is interested to him. The idea behind this approach is to

grab that saliency point and put it in the middle of the output image. The observed PSNR obtained from different sequences shows that the theoretically quality of the image does not suffer. Movies earned a low score in the subjective testing for this methodology. This approach has some advantages and disadvantages which are explained in figure 8.

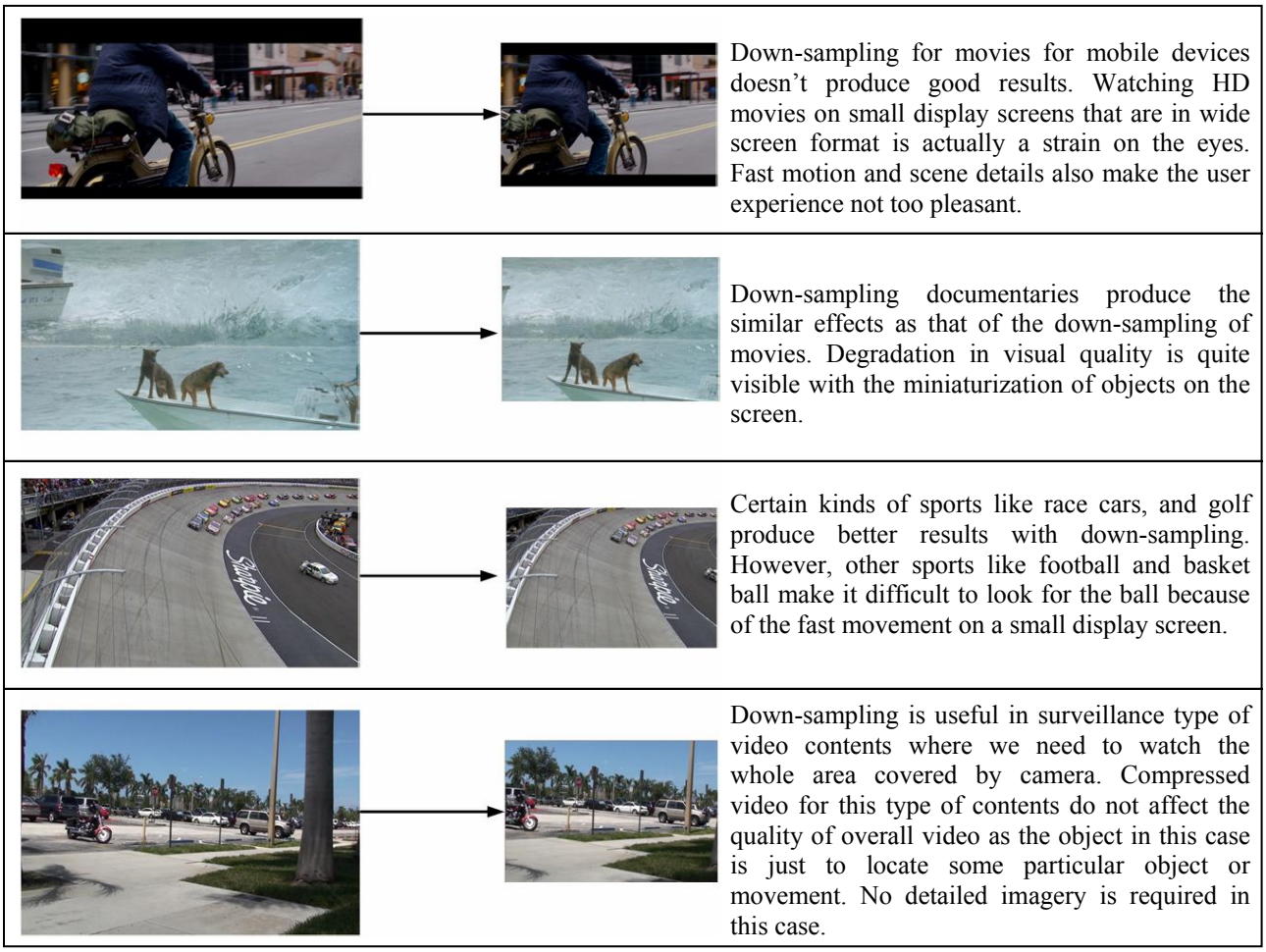

Fig. 6. Down-sampling applied to different bitstreams.

4.4 Saliency Region Based ROI

This process determines multiple points of saliency in a given image; the adaptation engine of this process then identifies a region encompassing all of the already found points of saliency and that region is the output of this process. This adaptation process produces the best results of all of the other methods since it tries to include multiple points of saliency in the output image, therefore the chances of possibility of including user's interested area in the output image also increases. The application and the output of this process is shown in figure 9. PSNR shows better theoretical quality of the

4.5 Discussion

The saliency based framework has many potential applications. One application is knowledge-based video coding. Many different kinds of knowledge about the contents and the contexts of the encoded sequences can be naturally used while adaptation process. This implies that saliency based framework is very good for special-purpose video communication applications such as videoconferencing and telemedicine, where a lot of prior information is available to the encoder. In general, the more we know about the video signal being encoded, the more we can improve the performance of saliency based framework.

Saliency based frameworkis very suitable for dynamic variable bit rate network video transmission. For example, if the available bandwidth drops dramatically on the network, a fixed data rate coding system has to stop transmission. A uniform resolution scalable coding system can still work properly but might transmit completely unacceptable quality video to the client. A system based on saliency detection, however, may still deliver useful information to the client, who might be specifically interested in certain areas in the video frame at different times.

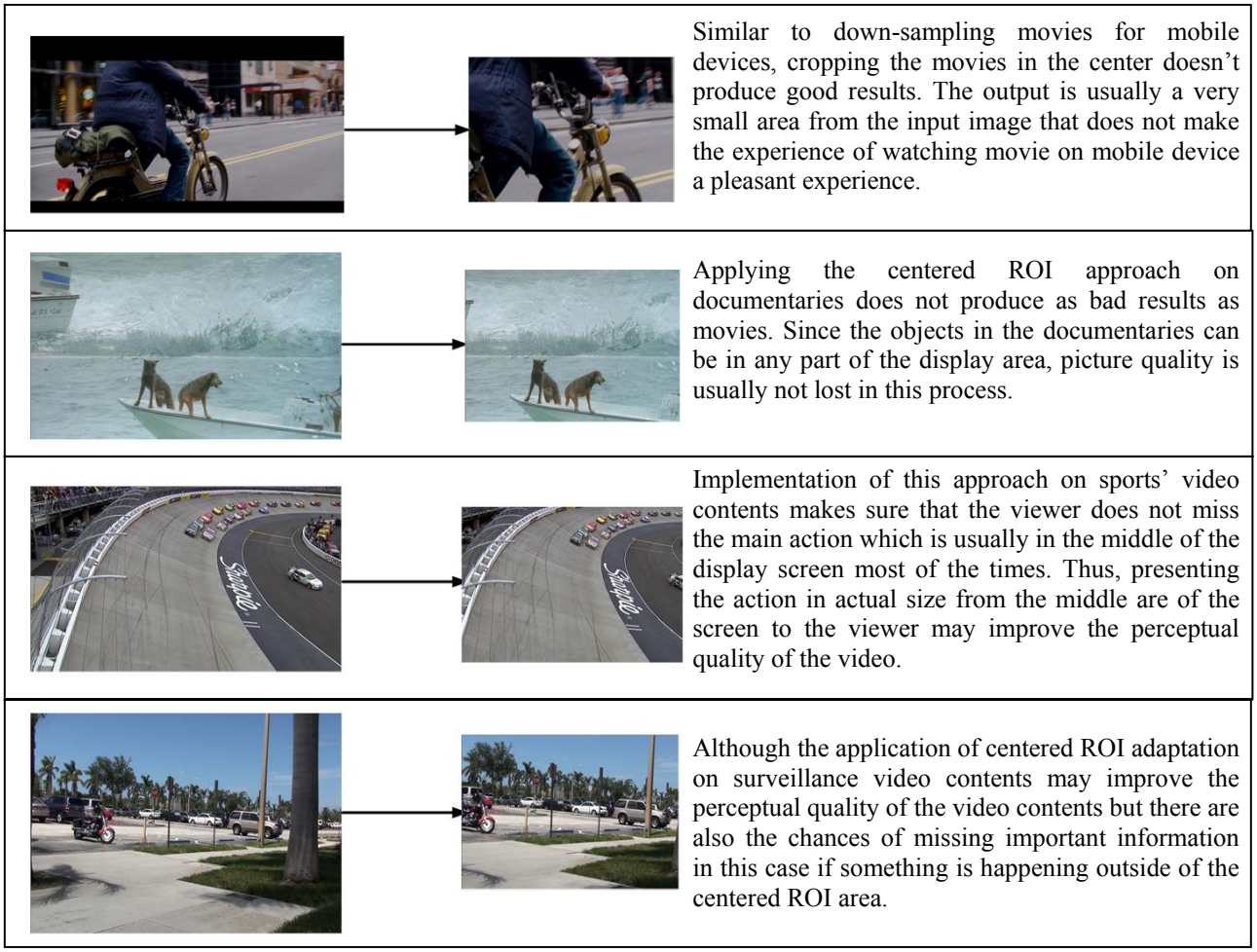

Fig. 7. Centered ROI applied to different bitstreams.

Saliency based framework also provides greater flexibility for multi-user and heterogeneous network video communications. If the video server needs to send video signals to different users with very different bandwidth connections, then saliency based framework, with only one-time encoding, supports the possibility to provide every user with the best quality video he/she can get because of the selection of a limited area from the source bitstream.

Finally, saliency based framework is also a good choice for interactive video communications, where the users are involved in giving feedback information to the other side of the communication system. The feedback information may be regions or objects of interest and can be converted into knowledge about the video sequence inside the encoder part of the framework. Consequently, improved video quality can be achieved.

Bitstreams that are adapted by this ROI extraction process by using the saliency based model have a significantly different portion of the video contents selected from the incoming bitstream than in the original case. While this has in general a profound impact on the quality of the decoded video sequence, this impact is marginal in case of a fixed camera and a static background. This observation may lead to new opportunities in the domain of video surveillance or

video conferencing. Next to the decrease in bandwidth, the adaptation process has a positive effect on the receiving decoder: because of the easy processing of salient region the decoding speed increases.

In our future work, we will further investigate the possibilities of adapting other types of video contents, such as video conferencing, in the input bit stream to produce the adapted output bit stream, and avoiding the need of complete decoding for faster ROI transcoding in the encoded domain. We will also investigate the possibility of this content adaptation framework, as presented in this paper, operates in real-time. Because if each component of the framework is able to function in case of actual streaming video, the framework will also be suited for live streaming video applications. As such, the framework can be deployed in an active network node, for instance at the edge of two different networks.

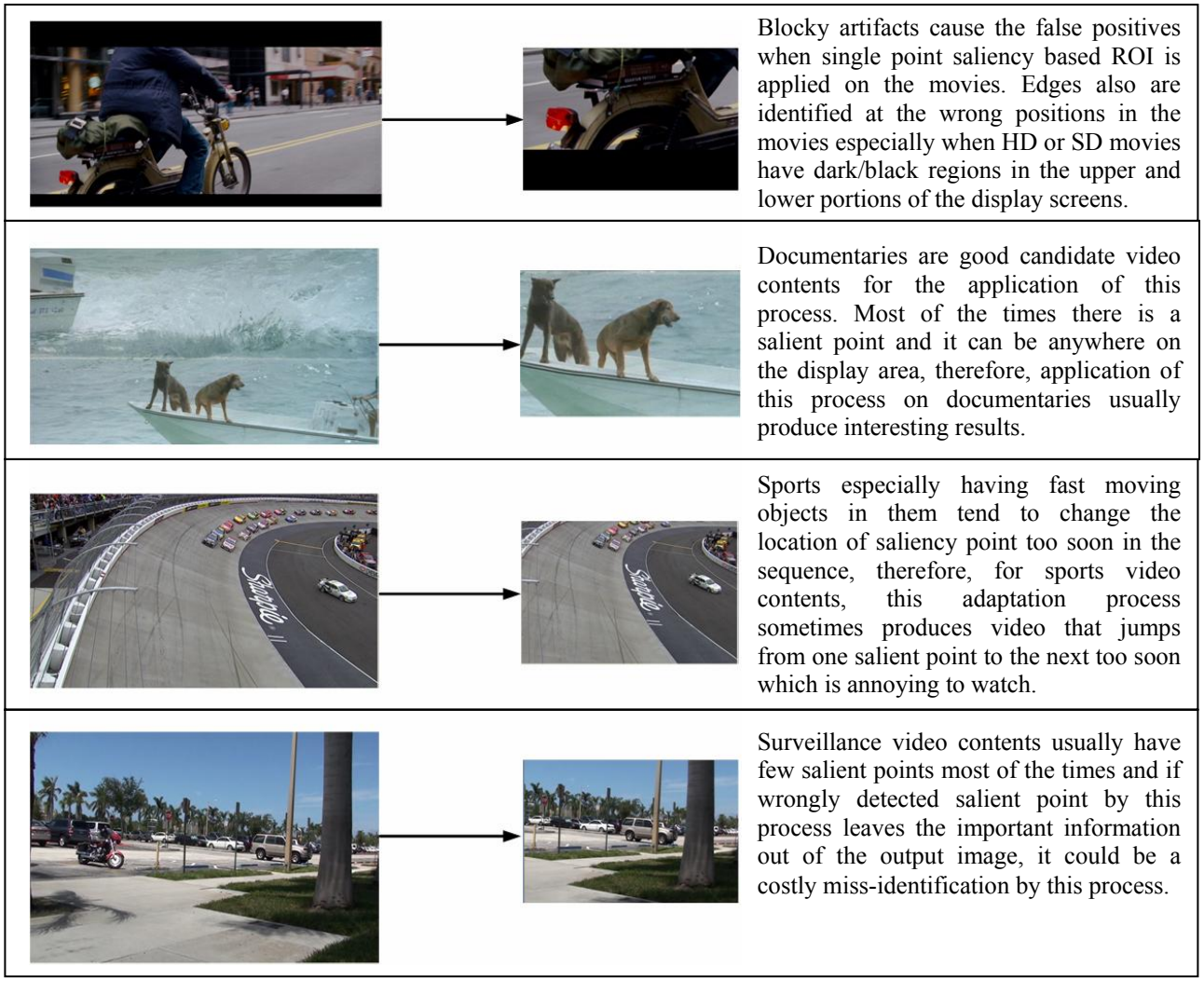

Fig. 8. Single Point Based Saliency applied to different bitstreams.

5. CONCLUSIONS

This paper discusses the feasibility of exploiting spatio-temporal characteristics of human vision to adapt video contents for small mobile devices. The proposed approach helps the adaptation engines focus on the visually important components in the video sequence. In this paper we have evaluated the automatic creation of ROI using various methods. Structure and composition in professionally created content can be exploited to create ROIs around the center of a frame. The saliency based framework evaluated can be used to determine the ROI using computational models of

visual attention. We also presented the effects of different kinds of video contents and their suitability for the application of a particular approach. Four methods for ROI extraction were implemented for this framework that exploit human visual attention model based on spatio-temporal characteristics of the video contents. Finding the ROIs from video contents can be a complex task but if the video contents' category is known and given that the bitstream is targeted for mobile video applications, certain assumptions can be made to adapt the incoming bitstream according to the target devices and their preferences and ultimately, it can help to make the experience of watching video contents on mobile devices more enjoyable.

Fig. 9. Saliency Region Based ROI applied to different bitstreams.

REFERENCES

[1] Kalva, Hari., "The H.264/AVC Video Coding Standard," IEEE Multimedia, Vol. 13, No 4, Oct.-Dec. 2006, pp. 86-90.

[2] A. Luthra, R. Gandhi, K. Mckeon, Y. Yu, K. Panusopone, D. Baylon, L. Wang. "Performance of MPEG-4 Profiles used for Streaming Video and Comparison with H.26L". M7227, Motorola, Broadband Communications Sector, July 2001.
[3] G. Fernandez-Escribano, H. Kalva, P. Cuenca, and L. Orozco-Barbosa, "Very Low Complexity MPEG-2 to H.264 Transcoding Using Machine Learning," Proceedings of the ACM Multimedia 2007, Santa Barbara, CA, October 2006, pp. 931-940.
[4] D. Wu, T. Hou, Y. –Q. Zhang, "Transporting real time video over the Internet: challenges and approaches", Proc. IEEE 88 (2000) 1855-1877.
[5] Siddhartha Chattopadhyay, Suchendra M. Bhandarkar, Kang Li, "FGS-MR: MPEG4 Fine Grained Scalable Multi Resolution Layered Video Encoding", Proc. Of ACM NOSSDAV'06, pp. 5, New Port, Rhode Island, May 22-23, 2006.
[6] Schaar, M. van der, and Lin, Y.-T., "Content-based selective enhancement for streaming video", in Proc. IEEE International Conference on Image Processing, Vol. 2, pp. 977-980, 2001.
[7] J. B. Lee and A. Eleftheriadis, "Spatio-temporal model-assisted compatible coding for low and very low bit rate video-telephony," Proc. IEEE Int. Conf. Image Processing Lausanne, Switzerland, pp. 429-432, Oct. 1996.
[8] Z. Wang, L. G. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection", IEEE Trans. on Image Processing, vol. 12, pp. 243-254, Feb. 2003.
[9] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention", IEEE Trans. on Image Processing, 2004.
[10] Mobile Devices and Content, http://www.dvdinformation.com/TechResources/images/DEG_MobileDevices_FINAL.pdf
[11] T. Wiegand, G. Sullivan, J. Reichel, H. Schwarz, and M. Wien, "Joint draft 8 of SVC amendment", ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6 9 (JVT-U201), 21st Meeting, Hangzhou, China, Oct. 2006.
[12] M. H. Lee, H. W. Sun, D. Ichimura, Y. Honda, and S. M. Shen, "ROI slice SEI message", JVT-S054, Input Document Joint Video Team (JVT), Geneva, Switzerland, Apr. 2006.
[13] ISO/IEC JTC/SC29/WG11, "Applications and requirements for Scalable Video Coding", N6880, Jan. 2005.
[14] Video Codec for Audiovisual Services at p _ 64 kbit/s, ITU-T Rec. H.261, ITU-T, Version 1: Nov. 1990, Version 2: Mar. 1993.
[15] Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 Mbit/s—Part 2: Video, ISO/IEC 11172-2 (MPEG-1 Video), ISO/IEC JTC 1, Mar. 1993.
[16] Generic Coding of Moving Pictures and Associated Audio Information—Part 2: Video, ITU-T Rec. H.262 and ISO/IEC 13818-2 (MPEG-2 Video), ITU-T and ISO/IEC JTC 1, Nov. 1994.
[17] Video Coding for Low Bit Rate communication, ITU-T Rec. H.263, ITU-T, Version 1: Nov. 1995, Version 2: Jan. 1998, Version 3: Nov. 2000.
[18] Coding of audio-visual objects—Part 2: Visual, ISO/IEC 14492-2 (MPEG-4 Visual), ISO/IEC JTC 1, Version 1: Apr. 1999, Version 2: Feb. 2000, Version 3: May 2004.
[19] Advanced Video Coding for Generic Audiovisual Services, ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), ITU-T and ISO/IEC JTC 1, Version 1: May 2003, Version 2: May 2004, Version 3: Mar. 2005, Version 4: Sept. 2005, Version 5 and Version 6: June 2006, Version 7: Apr. 2007, Version 8 (including SVC extension): Consented in July 2007.
[20] Wei Liu; Wei Wang; Guohui Li, "A FOA-based error resiliency scheme for video transmission over unreliable channels," Wireless Communications, Networking and Mobile Computing, 2005. Proceedings. 2005 International Conference on , vol.2, no., pp. 1265-1270, 23-26 Sept. 2005.
[21] P. Lambert, D. De Schrijver, D. Van Deursen, W. De Neve, Y. Dhondt, R. Van de Walle, "A real-time content adaptation framework for exploiting ROI scalability in H.264/AVC", in: Lecture Notes in Computer Science (8th international conference on Advanced Concepts for Intelligent Vision Systems), vol. 4179, Antwerp, Belgium, 2006, pp. 442–453.
[22] H. Kodikara Arachchi, S. Dogan, H. Uzuner, A. M. Kondoz, "Utilising Macroblock SKIP Mode Information to Accelerate Cropping of an H.264/AVC Encoded Video Sequence for User Centric Content Adaptation," *axmedis*, pp. 3-6, Third International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution (AXMEDIS'07), 2007.
[23] D. De Schrijver, W. De Neve, D. Van Deursen, S. De Bruyne, and R. Van de Walle, "Exploitation of interactive region of interest scalability in scalable video coding by using an XML-driven adaptation framework", in *Proc AXMEDIS'2006*, Leeds, UK, Dec. 2006.