

Region-of-Interest Based Video Image Transcoding for Heterogenous Client Displays

Karun B. Shimoga

Dept. of Radiation Oncology, Presbyterian University Hospital,
Pittsburgh, PA, USA.

Abstract

Problem:

It is a formidable problem to broadcast quality video to individually suit the variety of clients' display capabilities. As a solution to the problem, we present an algorithm for adapting video image size to multiple heterogeneous clients while maximizing information content in the transcoded image.

Solution:

If an image is decimated, some sub-images will have higher *visual attention* (VA) value than others. Studies show that most useful information is contained in areas with high VA. Given an image, VA of individual sub-images can be determined using any of several existing methods. However, the VA value for individual image-blocks are assigned over range 0.0 to 1.0, depending on the importance of the block's contents. Then, given the quantitative VA values of all blocks in the image, the *Region-of-Interest* (cropped sub-image) will be chosen to encompass as many regions of high VA values as possible, thereby maximizing the information content (VA value) of the final image. This is the crux of our algorithm.

Results: A 512x512 gray scale image was considered for transcoding for five types of client displays: workstation (256x256), desktop (192x192), TV-browser (128x128), hand-held (96x96), and personal digital assistant (PDA) (64x64). For simplicity, the aspect ratios of all displays was assumed to be 1.00. The 512x512 image was divided into 64 sub-images (8-rows, 8-columns). Each 64x64 block was assigned a VA value. The new transcoding algorithm was applied to get the final image for each display size. Results showed that in the workstation image, most important regions of the original image were preserved while containing significant global and local information. In the PDA image, although significant global information was not preservable, the algorithm had retained the best local information under the display-size and compression-ratio constraints.

Conclusions: Results show that region-of-interest based cropping, with maximized visual attention value, is the most natural method for cropping an image for transcoding for display on heterogenous clients.

Keywords: Region of Interest, Visual Attention, Video Transcoding, Image Transcoding, Display Transcoding, Transcoding.

1. Introduction

1.1 Transcoding

The packetvideo applications are being accessed by ever increasing varieties of devices such as Hand-held computers (HHCs), personal digital assistants (PDAs) and video-phones. Being the dominant part of multimedia, the images and videos have the most impact on the quality of any application. However, the display capabilities of the receiving clients vary substantially. Typical workstation or a desktop PC can display a good quality large color image without difficulty although a video display can be slow. Many HHCs, PDAs and video phones, on the other hand, can only display small images. TV-based web browsers are constrained by low-resolution interlaced display. Given the variety of client devices, it is a formidable problem to broadcast images and videos individually suited to the specific client's display capabilities. A solution to this problem is the *transcoding*.

Image transcoding is the process of adapting the image attributes so as to suit the communication, storage, processing and display capabilities of client devices while maximizing the quality of presentation. The image attributes include the resolution, accuracy, number of colors, and the image format. The display capabilities of clients include the pixel resolution, number of colors and color depth and formats. This work focuses on video image transcoding for client display size only.

1.2 Related Work

Several transcoding approaches have been proposed in the literature for adapting images for heterogenous client displays. A novel transcoding approach is to describe each image using the data format of *InfoPyramid*, proposed and detailed in Li et al [1], and Smith et al [2]. In *InfoPyramid*, different media contents such as text, graphics, image, audio and video, are described at different resolutions and multiple abstractions. Then, a rule-based transcoding proxy will dynamically select a combination of resolutions and abstraction of each content to best suit the client's capabilities. This approach is good for a pre-authored image content such as web pages but is unsuitable for streaming video images.

Another novel approach is that each image be decomposed into image objects of specific type and purpose, as suggested by Smith et al [3]. Each image object is then assigned a value of importance based on the type and purpose. Given a client's display size and capabilities, object types of high importance values are retained while objects of low importance are dropped. This approach is also very well suited for pre-annotated web pages where the author can specify the object types and purpose. However, for video images, computation needed for real-time annotation will be prohibitively complex and hence the approach is unsuitable.

In the past two approaches, the images are eventually reduced by, either sub-sampling and/or by cropping, to suit client's display size. However, if the objects involved are type text or facial images, excessive subsampling will prevent the object from serving any useful purpose. Therefore, Lee et al. [4] suggest that regardless of the level of importance, each object must not be sub-sampled beyond the limit of human perceptual ability and propose a perception-based transcoding method where each block of image is assigned with two attributes: importance and perceptual-hint. Here, perceptual hint is nothing but the maximum reduction in spatial resolution without affecting the human visual perception. This approach is better than those based on the *InfoPyramid* and the content-classification

described above. Once again, annotating each block's importance and perceptual resolution limit prevents the approach from use for video images.

To transcode video images, we require a means of automatically assigning importance and resolution in real-time and this is the approach proposed in this paper.

This paper is organized as follows. Section-2 describes how we can assign the level of importance to each block of an image based on the notion of visual attention. Section-3 shows how the region of interest is decided within an image, based on the level of visual attention and a constant maximum allowable image size reduction. Examples and results are presented in Section-4 while Section-5 concludes the paper.

2. Visual Attention

Visual attention is the ability of a part of an image to attract the attention of the viewer. For instance, in Figure-1, the cars are the areas of high visual attention while the surrounding and the racing track itself is not of much visual attention. Similarly, in an image of an individual, his or her face is often the point of visual attention unless there is more interesting object within the image. Thus, the visually most striking or interesting object or region within an image is the point/area of visual attention. Studies show that the areas of high visual attention value have most of the important information within an image. Therefore, it is necessary that we focus attention on visual attention values of subregions within an image since we are concerned with retaining important information content.

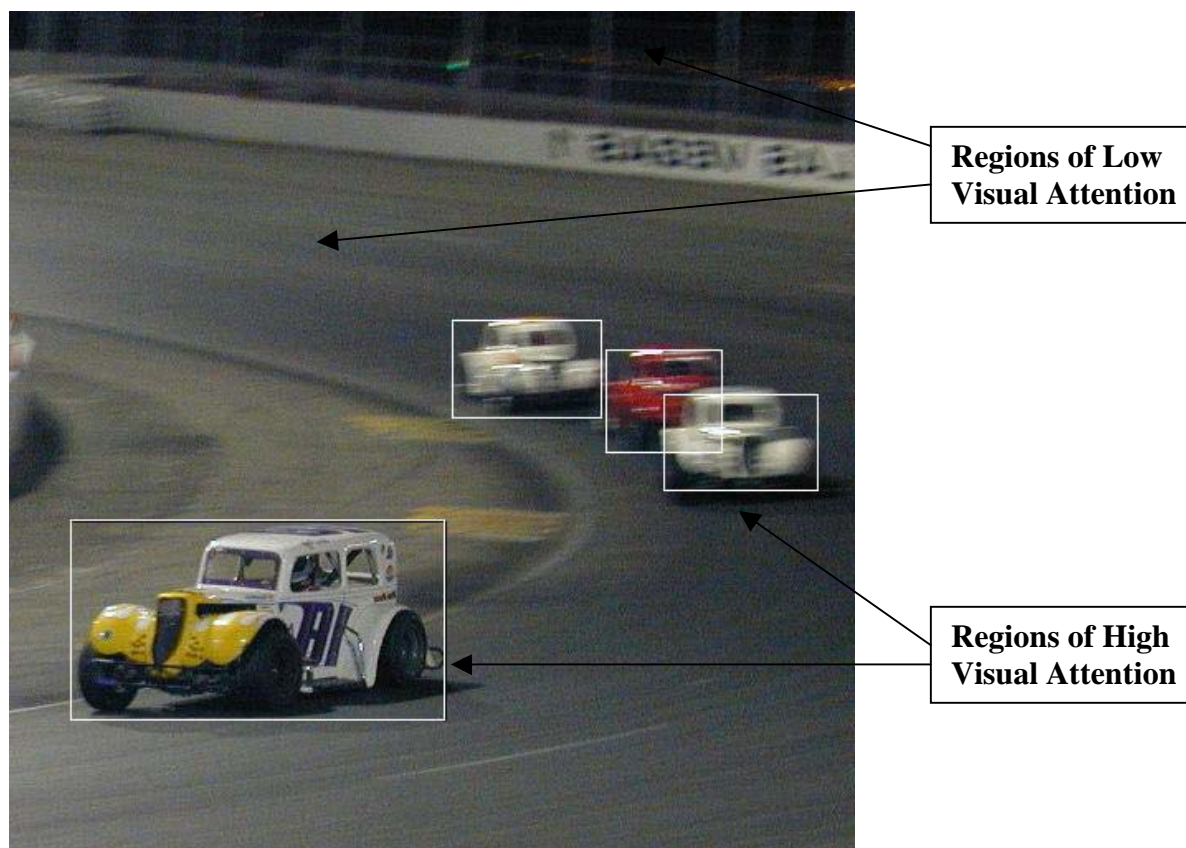


Figure-1: Examples of areas of HIGH and LOW visual attention

Several approaches exist for determining visual attention of a given image. These approaches are, in general, data driven [5] and based on the principle of feature integration. Feature integration is an approach that uses the psychophysical characteristics of human visual attention to compute the regions of visual attention [6,7].

Using any suitable method, if we can determine the quantitative visual attention value of each region in an image, we can then choose the best regions of visual attention and accommodate as many of them as possible within the transcoded image, thereby maximizing the information content of the transcoded image.

3. Region of Interest (ROI)

A region of interest (ROI) in this paper, is defined as a rectangular subimage of desired size. Given the desired size of the ROI, our goal is to choose the ROI with the highest total visual attention value.

Let I be the given image dicimated into $N \times M$ image blocks as shown in Figure-2. Let V_{ij} be the visual attention value of the block I_{ij} . Let the desired size of ROI be $n \times m$ where $n \leq N$ and $m \leq M$. Then, the visual attention value V_{ROI} of such region of interest window I_{ROI} . I_{ROI} positioned arbitrarily over the image I will be the sum of VA values of all blocks contained within the ROI window.

$$V_{ROI} = V(I_{ROI}) = \sum_i \sum_j V_{ij} \text{ for } i = 1 \text{ to } N; j = 1 \text{ to } M. \text{-----} (1)$$

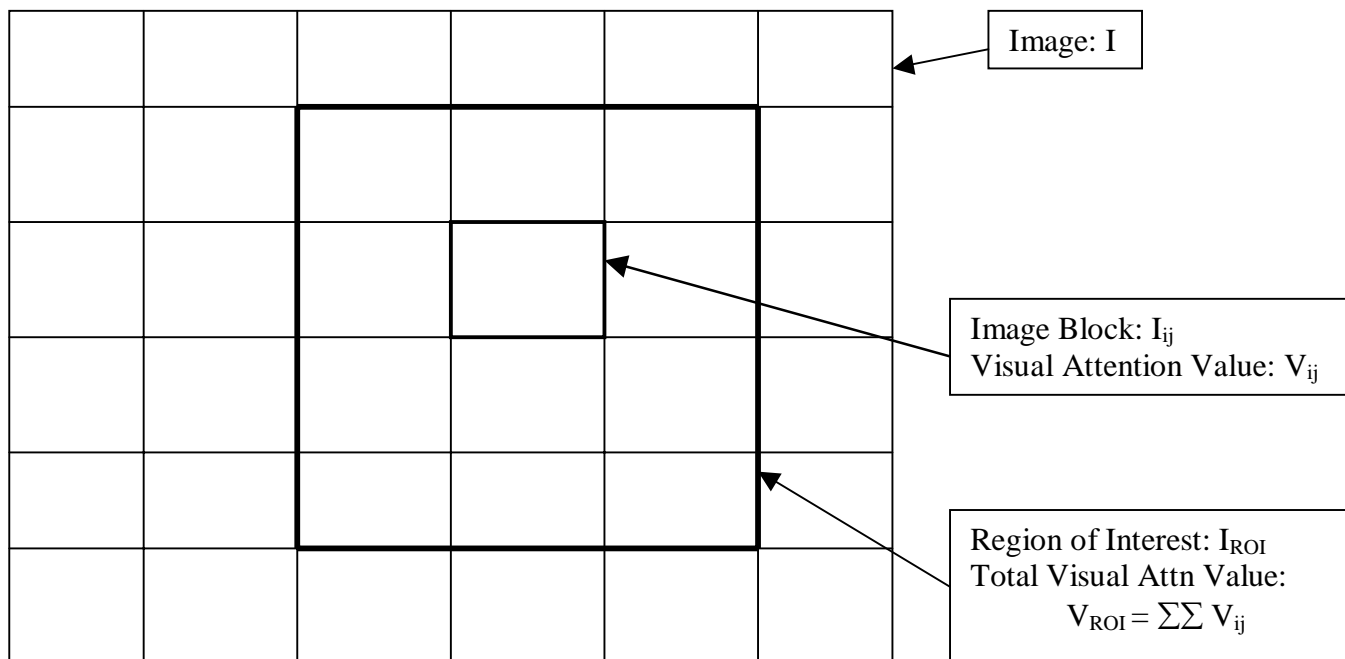


Figure-2: Definitions of Image Block and Region-of-Interest (ROI). The total visual attention value of a ROI is the sum of the visual attention values of all image blocks contained within the ROI.

Now, the problem of choosing the best region of interest I^{\wedge} of a given size reduces to that of searching a ROI window with largest V_{ROI} within the image I .

$$\hat{I} = I_{ROI} \text{ such that } V(I_{ROI}) = V_{max} \text{ ----- (2)}$$

The next question is: how do we determine the desired size of the ROI window? Consider that R^* is the maximum allowable compression ratio on any part of the video so as not to exceed the human perceptual ability limit. Then, the size of the ROI window must be, at the most, $1/R^*$ times larger than the target device display size.

$$(n, m) = (n^*, m^*) \times R^* \text{ ----- (3)}$$

Thus, given the display size (n^*, m^*) of the device and the maximum allowable resolution reduction ratio (R^*), the desired size of ROI is computed from Equation-3.

The ROI-based transcoding algorithm can now be summarized as follows:

- **Step #1:** Specify (a) source image I and its size $(N \times M)$ pixels; (b) target device display size $(n^* \times m^*)$ pixels; (c) maximum allowable compression ratio (R^*) so as not to exceed human perceptual limit; and (d) Decimation blocking rows (K) and columns (L).
- **Step #2:** Decimate image I into $K \times L$ blocks.
- **Step #3:** Determine the visual attention value V_{ij} of each block using any of the existing techniques like the ones listed in Section-2.
- **Step #4:** Use device display size (n^*, m^*) and maximum allowable resolution reduction ratio R^* to calculate maximum size of the ROI from Equation-3.
- **Step #5:** Calculated visual attention value of each ROI using Equation-1.
- **Step #6:** Choose the best ROI, \hat{I} as in Equation-2.
- **Step #7:** Crop \hat{I} from I and reduce \hat{I} by a compression ratio of R^* .
- **Step #8:** Resulting image I^* of size $(n^* \times m^*)$ is the transcoded image for the target device display.

4. Results

A 512 x 512 gray scale image, shown in Figure-1, was considered for transcoding to fit display sizes of five types of client devices: work-station (256x256), desktop (192x192), TV-browser (128x128), handheld (96x96), and personal digital assistant (PDA) (64x64). Two sets of results were obtained. The first set contained direct scaling down by subsampling the original image without any preferential consideration for its contents. The second set was the result of applying the ROI-based transcoding algorithm to the given image.

4.1 Direct Scaling

In this approach, the entire 512 x 512 image in Figure-1 was reduced to fit the display sizes of the five devices. Consequently, the reduction ratios needed for each image were different: workstation (2.0), PC (2.67), TV browser (4.0), HHD (5.33), and PDA (8.0). Results are shown in Figure-3. Obviously, all global content was retained in each transcoded image but the local details were lost depending on the reduction ratio. The workstation image has good local and global contents. However, the PDA image hardly has adequate spatial resolution to infer any local detail. Thus, the direct scaling is useful only if the global information is important and local information is irrelevant.

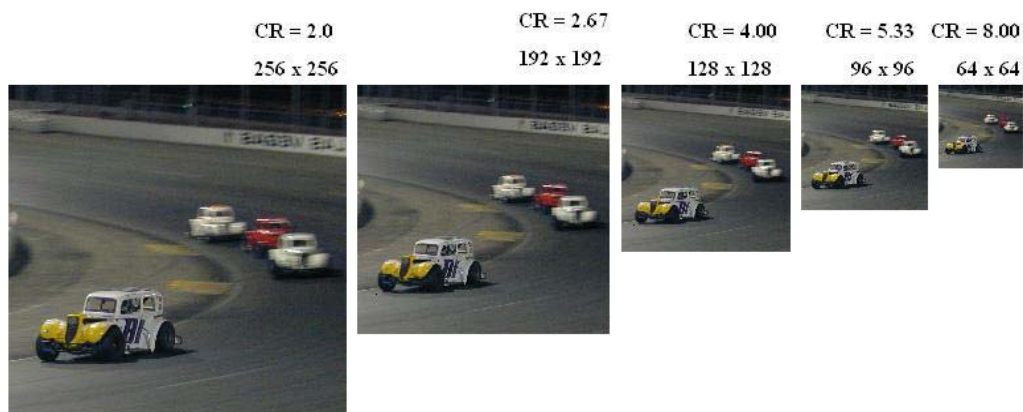


Figure-3: Results of using direct scaling for transcoding. Notice that, as we progress toward 64x64 image, local details become less and less visible.

4.2 ROI-based Algorithm

In this case, the parameters used in Step #1 of the algorithm were as follows: Source image I was as shown in Figure-1 (512 x 512 pixels). The display sizes of the five devices were as listed at the start of Section-4. Maximum allowable reduction ratio is 2.0 and the decimation grid dimensions are 8 rows x 8 columns, 64 blocks in all. For simplicity, the visual attention values were assigned to the 64 blocks and were shown in Table-1. In a real-time implementation, however, the visual attention values must be computed on-line.

Transcoded images are shown in Figure-4. Obviously, the local details are well preserved in all images. However, due to the reduction ratio constraints, not all global information was preserved. Despite reduction in global content, the overall content is well preserved in most images. So, it is needless to say that the ROI-based transcoding algorithm has been able to retain higher local content, compared to direct scaling.

Columns →

0.3	0.3	0.3	0.3	0.1	0.1	0.1	0.1
0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3
0.2	0.2	0.2	0.2	0.8	0.8	0.1	0.1
0.3	0.3	0.3	0.3	0.8	0.8	0.9	0.9
0.4	0.3	0.3	0.4	0.4	0.4	0.9	0.9
0.4	1.0	1.0	1.0	0.4	0.4	0.2	0.2
0.4	1.0	1.0	1.0	0.2	0.2	0.2	0.2
0.2	0.3	0.3	0.3	0.2	0.2	0.2	0.2

↓ Rows

Table-1: Visual Attention Values assigned to the 64 sub-image blocks. Imagine the grid to be superimposed on the given image shown in Figure-1.

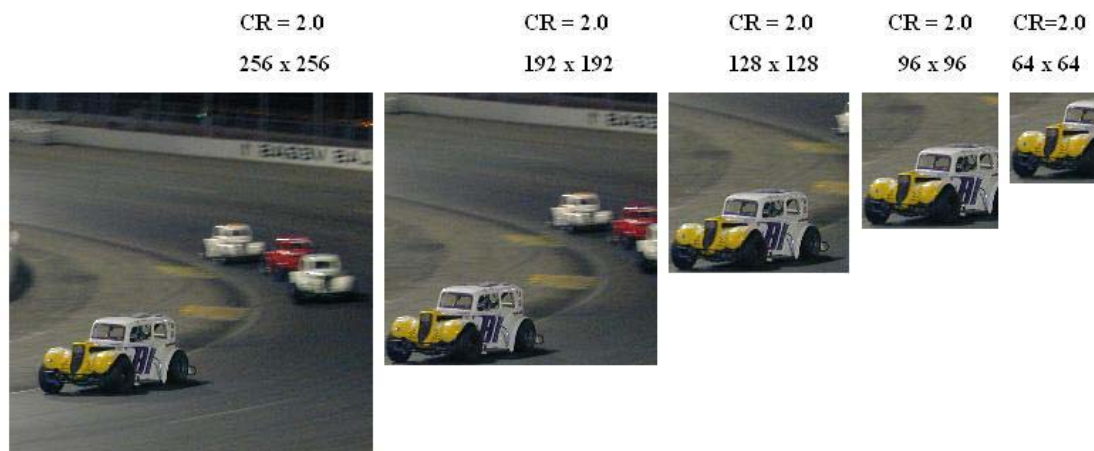


Figure-4: Results of applying the ROI-based transcoding algorithm to the reference image. Notice how the local details are very much retained even in 64x64 image.

5. Summary

This paper presented and demonstrated a region-of-interest based algorithm for transcoding video images to suit five types of client device display sizes. The algorithm is based on choosing the subimage with highest information content by maximizing the total visual attention value before subsampling the image to suit device display size.

Results of applying the algorithm to a 512 x 512 gray scale image showed significant retention of both global and local details in transcoded images as compared to simple scaling approach which is currently used by many.

In conclusion, the ROI-based transcoding, with maximized visual attention value, is a natural method for transcoding images for heterogeneous client displays typical of wireless networks.

6. References

1. Li, C.S., Mohan, R. and Smith, J.R., Multimedia content description in the infopyramid, Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, pp. 3789-92, 1998.
2. Smith, J.R.; Mohan, R.; Chung-Sheng Li, Transcoding Internet content for heterogeneous client devices, Proc. IEEE Intl. Symp. on Circuits and Systems, vol. 3, pp. 599-602, 1998.
3. Smith, J.R., Mohan, R., and Li, C.S., Content-based transcoding of images in the internet, IEEE Intl. Conf. on Image Processing, pp. 7-11, 1998.
4. Lee, K., Chang, H.S., Chun, S.S., Choi, H., and Sull, S., Perception-based image transcoding for universal multimedia access, IEEE Intl. Conf. on Image Processing, vol. 2, pp. 475-478, 2001.
5. Itti, L., Koch, C. and Niebur, E., A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Analysis and Machine Intelligence, 20(11):1254-1259, 1998.
6. Privitera, C., and Stark, L., Focused JPEG encoding based upon automated pre-identified regions of interests, SPIE vol. 3644, pp. 552-58, 1999.
7. Ouerhani, N., Bracamonte, J., Hugli, H., Ansorge, M. and Pellandini, F., Adaptive color image compression based on visual attention., 11th IEEE Intl. Conf. On Image Analysis and Processing, pp. 416-21, 2001.