# FINDING REGIONS OF INTEREST IN HOME VIDEOS BASED ON CAMERA MOTION

*Golnaz Abdollahian and Edward J. Delp*

Video and Image Processing Lab
School of Electrical and Computer Engineering
Purdue University, West Lafayette, IN 47907, USA

## ABSTRACT

In this paper, we propose an algorithm for identifying regions of interest (ROIs) in video, particularly for the keyframes extracted from a home video. The camera motion is introduced as a new factor that can influence the visual saliency. The global motion parameters are used to generate location-based importance maps. These maps can be combined with other saliency maps calculated using other visual and high-level features. Here, we employed the contrast-based saliency as an important low level factor along with face detection as a high level feature in our approach.

***Index Terms***— video content analysis, regions of interest identification, visual saliency map, keyframes

## 1. INTRODUCTION

Searching and browsing through stored video volumes is a time consuming task unless we have more efficient ways of representing a video in an abstract form so that it preserves the "important" contents while removing the redundancy. There are several types of video summarization among which are "representative" frames, keyframes, and key objects.

In our previous work [1], we proposed an approach to employ the intentional camera motion for temporally segmenting the home video clips and extracting the most significant frames, keyframes, from each subsegment. We considered the importance from the camera person's point of view while shooting the video, i.e. what a person considered important while shooting a video. From this viewpoint, camera motion is a good indicator of visual interest. In this paper, we extend the approach to the frame level in order to extract the regions of interest in the keyframes.

Higher-level information can be obtained from the extracted keyframes e.g., regions of interest (ROIs) and key objects. In this paper, we propose an approach to identify the ROIs from the selected keyframes i.e. regions that capture the attention of the viewers while watching the video. Detection of these regions can be used for applications such as video adaptation to small-size screens. One of the challenges in dealing with video contents on mobile devices is how to represent the video or its summary on a small size display. Summarization methods which are based on displaying the video hierarchical trees and storyboards may not be suitable for limited display sizes. Therefore, being able to browse within a frame or zoom into video frames and watch a cropped version of video without loss of important features is an attractive task in mobile video management. Identifying the regions of interest helps to model the information structure within the frames and generate the browsing paths [2].

There has been a great deal of research on identifying the regions of interest in still images [3, 4]. Several factors such as contrast, size, shape, faces, foreground/background and location can influence visual attention [5]. Moreover, there have been some approaches that combine the spatial factors with motion vector fields data to determine the objects with high motion activities and use this information to identify ROIs in video [5, 6].

In this paper, we consider camera motion as another important factor that can be helpful in finding ROIs in video sequences. Based on our preliminary experiments conducted on a group of 15 people using a collection of home video clips, we observed that the direction of the camera motion has a major effect on the regions where the viewers notice the most in the sequence. For home video sequences, in which there is not much object activity, usually the intentional motion of the camera determines the "story" by moving around the scene or zooming in/out. Viewers have the tendency to follow this motion and particularly look for the new objects that are about to enter the camera view. We describe how to make use of this factor to generate location-based saliency maps for the extracted frames based on the global motion information of video at those frames. These maps are overlaid with contrast-based importance maps obtained by an algorithm similar to [3]. Finally, particular objects such as human faces are highlighted in the final saliency map.

The paper is organized as follows. Section 2 describes the method used to identify the saliency map based on contrast. We propose the location-based saliency calculation in Section 3. These maps are combined in Section 4. In Section 5, faces are detected and highlighted in the final importance map. The results for some previously extracted keyframes are illustrated

This work was sponsored by a grant from Motorola. Address all correspondence to E. J. Delp, ace@ecn.purdue.edu

in Section 6. Finally, concluding remarks and the ideas for our future work are discussed in 7.

## 2. CONTRAST-BASED SALIENCY MAP

A *saliency map* or *importance map(IM)* of an image indicates how much the viewers' eyes are visually attracted to different regions in that image. Studies based on eye movements and visual search have identified several factors that can influence visual attention such as motion activity, contrast, size, shape, faces, foreground/background and location [5, 3]. One of the important factors that causes a region to stand out and be more noticeable is the contrast of that region compared to its neighborhood. This includes contrast in luminance and color. Ma et al. [3] use the color components in LUV space as the stimulus on *perceive field*.

We used a similar technique to the one described in [3] however, we employed the RGB color space to generate the contrast-based IM. Our experiments showed better results in this space compared to using UV components since it combines the luminance and color contrasts. The contrast-based saliency map is constructed as follows.

First, the three-dimensional pixel vectors in RGB space are clustered into a small number of color vectors ( 32 in our experiments ) using peer group filtering method in [7]. This method suppresses color clusters in detailed regions where the human perception is less sensitive to the differences. This method can be considered as a weighted version of generalized Lloyd algorithm (GLA) for vector quantization [8] where the expressions for updating the location of cluster centroid and distortion measure are modified to:

$$c_i = \frac{\sum v(n)x(n)}{\sum v(n)},\ x(n) \in C_i \tag{1}$$

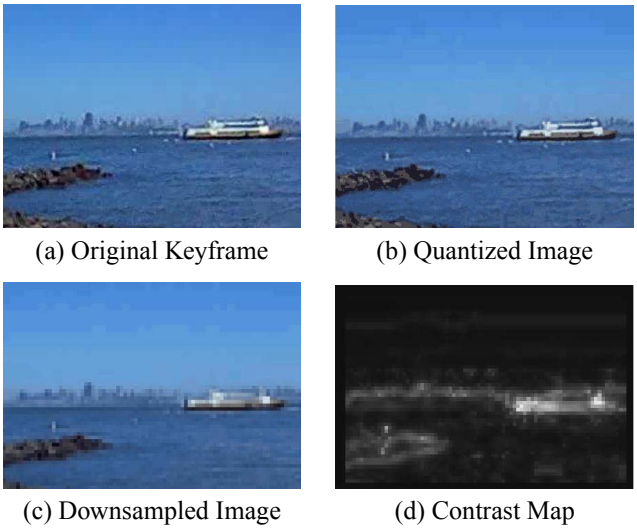
$$D_i = \sum v(n)\,\|x(n) - c_i\|^2,\ x(n) \in C_i \tag{2}$$

In which, $c_i$ and $D_i$ are the centroid and distortion of cluster $i$ respectively. $x(n)$ represents the RGB vector of each pixel and $v(n)$ is the corresponding pixel weight which is computed in such a way that pixels in noisy regions have smaller weights than the ones in the smooth regions [7]. At each iteration, the cluster with maximum distortion value is split into two new clusters until the predetermined number of clusters is obtained. The color-quantized frame is then downsampled by a factor of $n$ in each dimension in order to reduce the computation complexity. Here, the value of $n$ is 10.

The contrast-based saliency value for each pixel $(i, j)$ in the downsampled image (block $(i, j)$ in the original frame) is obtained by

$$S(i,j) = \sum_{q \in \Theta} d(p_{ij}, q) \tag{3}$$

Where $p_{ij}$ and $q$ are the RGB pixel values and $\Theta$ is the neighborhood of pixel (i,j). In our experiment, a $5 \times 5$ neighborhood was used. Figures 1 and 2 show examples of the contrast-based saliency map for some extracted keyframes.



(a) Original Keyframe (b) Quantized Image

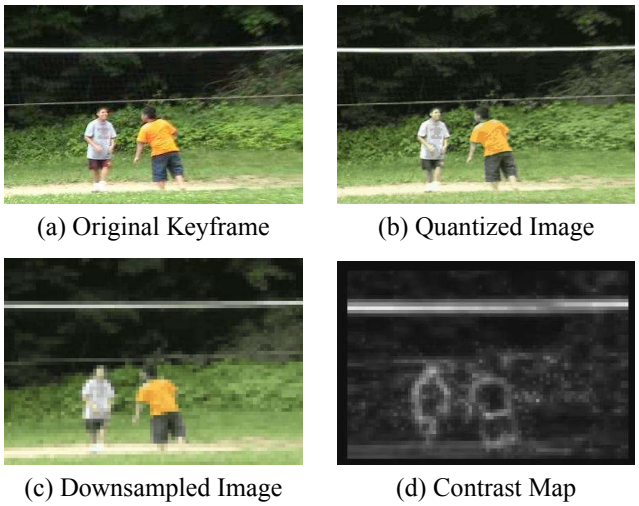
(c) Downsampled Image (d) Contrast Map

**Fig. 1**. Example of contrast-based saliency map for an extracted keyframe. (a) Original keyframe, (b) Quantized frame,(c) Downsampled frame and (d) Saliency map.

## 3. LOCATION SALIENCY MAP BASED ON CAMERA MOTION

Our preliminary study on a group of 15 people using a collection of home video clips showed that in addition to object motions, the direction of the camera motion has a major effect on the regions where the viewers notice the most in the sequence. Most approaches that have considered location saliency are based on the experiments on still images which have resulted in central saliency i.e. the center of the image is considered to be visually more important [5, 4, 6]. For home video sequences, in which there is not much object activity, usually the intentional motion of the camera determines the "story" by moving around the scene or zooming in/out. Human visual system have the tendency to follow this motion and particularly look for the new objects that are about to enter the camera view. For example, if the camera is panning towards the right, the viewers are more attracted to the right side of the scene or when the camera starts to zoom out, the attention to the borders increases. In the case of zoom-in or still camera, the location saliency is similar to the one for still pictures and the attention is more concentrated on the center of the frames.

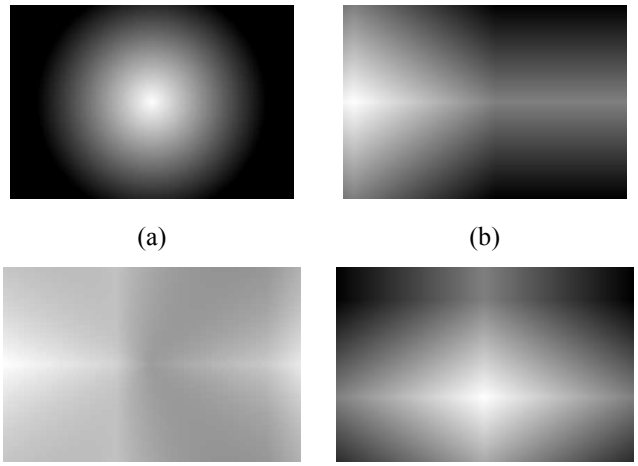
Since our strategy for selecting keyframes [1] were based on camera motion patterns, we can use the available motion information to generate the location maps for the extracted keyframe. The 3-parameter motion model was used, in which $H$, $V$ and $R$ represent horizontal, vertical and radial motion respectively and are estimated using Integral Template



(a) Original Keyframe (b) Quantized Image

(c) Downsampled Image (d) Contrast Map

**Fig. 2**. Example of contrast-based saliency map for an extracted keyframe. (a) Original keyframe, (b) Quantized frame, (c) Downsampled frame and (d) Saliency map.



(a) (b)

(c) (d)

**Fig. 3**. Examples of motion-based location masks for: (a) Zoom-in or camera hold , (b) Large panning toward left, (c) Zoom-out w/ panning left and (d) Tilting down.

Matching technique [9]. Three individual maps for $H$, $V$ and $R$ directions are generated (4), (5), (6) and combined to form the location saliency map (8).

$$Map_H(i,j) = max(0, 1 - \frac{|j - \frac{width}{2} - k_H * H|}{\frac{width}{2}}) \quad (4)$$

$$Map_V(i,j) = max(0, 1 - \frac{|i - \frac{height}{2} - k_V * V|}{\frac{height}{2}}) \quad (5)$$

$$Map_R(i,j) = \begin{cases} 1 - \frac{r}{r_{max}} & R \geq 0 \\ -k_r * \frac{r}{r_{max}} & R < 0. \end{cases} \quad (6)$$

$$Map = Map_H + Map_V + Map_R \quad (7)$$

Where $(i,j)$ is the pixel location, $k_H$, $k_V$ and $k_r$ are constants, whose values were experimentally found to be optimum at 10, 5 and 0.5, respectively. $r$ represents the distance of pixel from the center of the frame and $r_{max}$ is its maximum value in the frame. After combining $H$ and $V$ maps, the peak of the map function is at $(k_H * H, k_V * V)$. If there is no translational motion, the map function will have peak at the center of the frame. For the radial map, $Map_R$, the function is either decreasing or increasing as we move from the center to the borders, depending on whether the camera has a zoom-in/no-zoom or zoom-out operation. Some examples of the generated masks for various operations are shown in Figure 3.

## 4. COMBINING THE SALIENCY MAPS

The overall IM is generated by multiplying the corresponding pixel values of the normalized contrast-based IM ($S$) and location-based IM ($Map$).

$$IM = S \,.\, Map \quad (8)$$

The values of $IM$ are normalized to [0,1]. Examples of the generated saliency maps for several keyframes are illustrated in Section 6.

## 5. HIGHLIGHTING FACES IN THE SALIENCY MAP

In addition to low-level visual features mentioned above, specific objects such as faces, hands and texts can draw the viewers' attention. In our system, faces are detected using the online face detection in [10]. The detected faces are highlighted in the saliency map by assigning the saliency value $S = 1$ to pixels inside the face regions (Figure 4) since these regions are of high semantic importance and may have been overlooked in the low-level saliency maps.

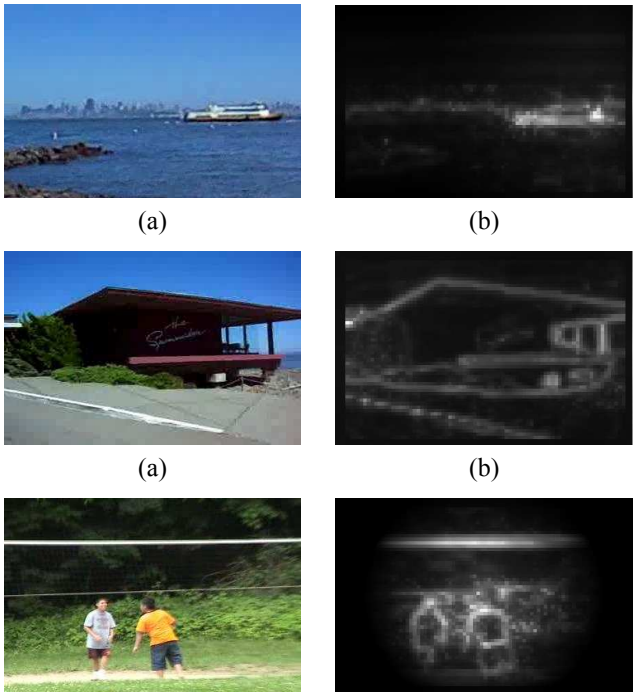
(a) (b)

**Fig. 4**. Highlighting face areas in the saliency map.(a) Original frame and (b) final saliency map.

## 6. EXPERIMENTAL RESULTS

We analyzed the extracted frames from various home video sequences, in particular, keyframes obtained in [1], using our algorithm. Some examples are illustrated in Figure 5. In the first row, the camera is panning towards the right and the motion parameters are $H = 10, V = 0, R = 0$. The final saliency map is brighter at the right side of the frame, which indicates more importance. The second keyframe (second row in Figure 5) is extracted after a large zoom-out ($H = 0, V = 0, R = -4$) so the map shows an emphasis on borders of the frames. The last example is a selected frame after a zoom-in ($H = 0, V = 0, R = 3$) so the attention is concentrated at the center of the frame.

Our preliminary experiments were done in a way that when we played the video clips for the viewers, they indicated what captures their attention the most in every view. The selected videos did not contain a great amount of object activities. We observed high correlation between the generated saliency and the user indicated areas.



(a) (b)

(a) (b)

(a) (b)

**Fig. 5**. Examples of saliency maps for extracted keyframes: (a) Original frames and (b) Saliency maps.

## 7. CONCLUSION AND FUTURE WORK

In this paper, we proposed a method for generating saliency maps for frames extracted from a video sequence. The camera motion has been used as an important factor that can affect the location of regions of interest in a video. First, a contrast-based saliency map is attained. Then, the global motion parameters are used to generate the location-based masks for each extracted frame. These masks are correlated with the initially obtained contrast images. Finally, the faces are detected and highlighted in the final maps. The results based on our experiments are promising. As future work, we will conduct more user studies for objective evaluation of our approach. Moreover, we would like to examine other important factors in attracting human visual attention in video.

## 8. REFERENCES


[1] G. Abdollahian and E. J. Delp, "Analysis of unstructured video based on camera motion," *Proceedings of SPIE International Conference on Multimedia Content Access: Algorithms and Systems*, San Jose, California, January 2007.

[2] X. Xie, H. Liu, W. Ma, and H. Zhang, "Browsing large pictures under limited display sizes," *IEEE Transactions on Multimedia*, vol. 8, no. 4, pp. 707–715, August 2006.

[3] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," *Proceedings of the eleventh ACM international conference on Multimedia*, 2003, pp. 374–381.

[4] F. Liu and M. Gleicher, "Automatic image retargeting with fisheye-view warping," *Proceedings of ACM UIST*, October 2005, pp. 153–162.

[5] W. Osberger and A. Rohaly, "Automatic detection of regions of interest in complex video sequences," *Proceeding of SPIE, Human Vision and Electronic Imaging VI*, vol. 4299, Bellingham, USA, 2001, pp. 361–372.

[6] Y. Hu, L.-T. Chia, and D. Rajan, "Region-of-interest based image resolution adaptation for mpeg-21 digital item," *Proceedings of ACM Multimedia*, New York, USA, October 2004.

[7] Y. Deng, M. M. C. Kenney, and B. Manjunath, "Peer group filtering and perceptual color image quantization," *IEEE International Symposium on Circuits and Systems*, vol. 4, pp. 21–24, 1999.

[8] R. M. Gray, "Vector quantization," *IEEE ASSP Magazine*, vol. 1, no. 2, pp. 4–29, April 1984.

[9] D. Lan, Y. Ma, and H. Zhang, "A novel motion-based representation for video mining," *Proceedings of IEEE International Conference on Multimedia and Expo*, Baltimore,Maryland, July 2003.

[10] Pittsburgh pattern recognition, demonstration: Face detection in photographs. [Online]. Available: http://demo.pittpatt.com/