Signal Processing: *Image Communication* 38 (2015) 32–44

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Video saliency detection incorporating temporal information in compressed domain

Qin Tu [a,*], Aidong Men [a], Zhuqing Jiang [a], Feng Ye [b], Jun Xu [a]

[a] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[b] School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350001, China

ARTICLE INFO

Available online 12 August 2015

Keywords:
Compressed domain
Video saliency detection
Visual window
Motion importance factor

ABSTRACT

Saliency detection is widely used to pick out relevant parts of a scene as visual attention regions for various image/video applications. Since video is increasingly being captured, moved and stored in compressed form, there is a need for detecting video saliency directly in compressed domain. In this study, a compressed video saliency detection algorithm is proposed based on discrete cosine transformation (DCT) coefficients and motion information within a visual window. Firstly, DCT coefficients and motion information are extracted from H.264 video bitstream without full decoding. Due to a high quantization parameter setting in encoder, skip/intra is easily chosen as the best prediction mode, resulting in a large number of blocks with zero motion vector and no residual existing in video bitstream. To address these problems, the motion vectors of skip/intra coded blocks are calculated by interpolating its surroundings. In addition, a visual window is constructed to enhance the contrast of features and to avoid being affected by encoder. Secondly, after spatial and temporal saliency maps being generated by the normalized entropy, a motion importance factor is imposed to refine the temporal saliency map. Finally, a variance-like fusion method is proposed to dynamically combine these maps to yield the final video saliency map. Experimental results show that the proposed approach significantly outperforms other state-of-the-art video saliency detection models.

1. Introduction

Video saliency provides an efficient way to filter out redundant visual information and to perceive most relevant parts of a visual field in dynamic scenes, which is imperative to understand the content presented in image/video. Picking out relevant parts of a scene as visual attention regions is a hot research topic and has a wide application prospect, such as content-based image and video retrieval [1], perceptual video compression and coding [2], object detection and segmentation [3–5], video analysis [6]. Recently, to tackle these information overload problems, numerous computational saliency models are proposed to simulate biological vision systems by researchers in physiology, psychology, neural system, and computer vision. Saliency detection aims at extracting salient objects or predicting human fixations in image/video, while various applications lead to different requirements for saliency detection. In this paper, we attempt to combine these two aspects and design a saliency model which can construct a strong correlation between human fixations and saliency maps, and can further assist salient objects extraction in practice applications.

* Corresponding author.
E-mail addresses: tuqin@outlook.com (Q. Tu), javetu@126.com (A. Men), jiangzhuqing@bupt.edu.cn (Z. Jiang), yefeng.bupt@gmail.com (F. Ye), xujun@bupt.edu.cn (J. Xu).

http://dx.doi.org/10.1016/j.image.2015.07.013
0923-5965/

Since Itti et al. [7] proposed a saliency detection model based on the neuronal architecture of the primates' early visual system, a great deal of similar contrast-based works, inspired by Itti's model, were presented using local feature contrasts between image regions and their surroundings. Local filtering [8], Fourier transform [9–11], mutual information formulation [12] or band filtering [13] technologies were regarded as principles in these models, and were implemented to integrate local feature contrasts for image saliency detection by highlighting the pixels with great center-surround differences [14].

In existing video saliency detection schemes [8,15,10,11,16–19], the image saliency detection approaches are still available for performing spatial saliency detection, and then motion information is fused to generate the saliency map of each frame. In GBVS [15], Harel et al. adopted graph theory to form saliency maps by using a better measure of dissimilarity based on [7]. In [8], the center-surrounding based saliency detection algorithm was inspired by biological vision, namely the psychophysics of motion-based perceptual grouping, and combined spatial and temporal components of saliency in a principled manner with a completely unsupervised method. Guo et al. [10] provided some examples to show that phase spectrum of Fourier transform (PFT) can get better results in comparison with spectral residual (SR) [9], and then extended PFT to QFT (Quaternion Fourier Transform) which was implemented to generate video saliency maps with color, intensity and motion features that were achieved by differing adjacent frames. Hu et al. [11] extended the PFT method in [10] to generate video saliency maps with pixel information in CIELab color space and motion information which was processed by normalization and global motion compensation. Li et al. [16] measured temporal saliency maps using an optical flow technology and the rank deficiency of gray-scale gradient tensors, and then used color, intensity, and orientation features to generate static saliency maps by multi-scale center-surround differences, and finally combined the above two maps using an adaptive linear combination. Bin et al. [17] constructed saliency features from video sequences, including color, texture and motion features. The GPU-based mean-shift segmentation and the region-based matching scheme were applied to the feature space for video saliency computation. Seo et al. [18] computed the so-called space–time local steering kernels from the given video, and then introduced the notion of self-resemblance to measure the visual saliency map of each frame. Zhai et al. [19] defined motion contrast by calculating projection errors with multiple homographies, and used color histograms of images to measure static saliency maps, and fused two maps using a linear combination.

In the above-mentioned methods, considerable efforts on finding video saliency are focused on pixel domain or uncompressed domain, luminance, color, texture and edge features are obtained from RGB or CIELab color space, similar to that extracted from still images. However, these schemes completely neglect motion information or deduce motion information from adjacent frames using an optical flow approach, which can implicitly or explicitly represent the characteristics of moving object, but is also affected by luminance and noise due to the assumption of luminance constancy [20]. With the rapid development of video services over internet or mobile network, more and more videos are encoded with standard video compression technologies which can enhance channel utilization. Thus, how to precisely and effectively measure conspicuous maps in compressed domain becomes an imperative and tough problem. So far, only a limited number of algorithms have been proposed for video saliency detection in compressed domain [21–24]. In [21], Fang et al. compiled saliency maps with adaptive entropy-based spatial and temporal uncertainty weighting. In this model, DCT coefficients, extracted from video bitstream, were used to detect spatial saliency maps while optical flow based motion information was obtained to detect temporal saliency maps. In [22], Fang et al. proposed a video saliency detection algorithm using center-surround differences, in which the features, extracted from MPEG4 ASP video, were used to compile saliency maps. Then, Otsu's threshold method was applied to adaptively combine spatial saliency maps and temporal saliency maps. Muthuswamy et al. [23] used the DCT coefficients of luminance and chroma components to calculate spatial saliency maps, and motion vectors (MVs) were utilized to refine it with a component-wise multiplication. In [24], motion information was used to detect temporal saliency maps by entropy and l_0 norm of DCT coefficients was used to detect spatial saliency maps, and then the final video saliency maps were generated by both additive and multiplicative combination. From these compressed video saliency detection algorithms, it seems obvious that the performances of saliency detection in compressed domain are better than that of saliency detection in uncompressed domain. The major causes for this end are that compressed video saliency models are not a stand-alone entities, the features used in these models are the best choice from video encoder whose aim is to provide the most compact representation of video, such as motion tracking and conspicuous objects selecting [24].

However, in these early compressed video saliency models, the features, extracted from video bitstream, are not fully taken advantage of, especially motion information and encoder parameters, such as skip/intra prediction mode in P frame. In a real scene, it always contains highly textured and clutter background or there are many moving objects in foreground. Some of these cases might cause the coarse saliency detection. It is to say that not all motions in a scene are salient, we should carefully handle the motions induced by the independent moving objects. Another important problem is the fusion of different saliency maps. Although the features used in [23,24] are obtained from several consecutive frames to reduce the influence of encoder, the fusion method, namely component-wise multiplication, cannot effectively reflect which saliency map is more important, resulting in inaccurate saliency detection, such as the incomplete contour of foreground objects and background noise. Therefore, to overcome these shortcomings, we propose a video saliency detection model in compressed domain illustrated in Fig. 1. The main properties of our algorithm include the following: