J. Vis. Commun. Image R. 24 (2013) 829–837

Contents lists available at SciVerse ScienceDirect

J. Vis. Commun. Image R.

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

Layered moving-object segmentation for stereoscopic video using motion and depth information

Yibin Chen [a], Canhui Cai [a,*], Kai-Kuang Ma [b], Xiaolan Wang [a]

[a] School of Information Science and Technology, Huaqiao University, China
[b] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 15 November 2011
Accepted 20 May 2013
Available online 4 June 2013

Keywords:
Video segmentation
Stereoscopic video
Layered segmentation
Depth-layer mask
Disparity estimation
Motion mask
Moving objects
Higher order statistics
Change detection

ABSTRACT

A novel *layered* stereoscopic moving-object segmentation method is proposed in this paper by exploiting both *motion* information and *depth* information to extract moving objects for each depth layer with high accuracy on their shape boundary. By taking a higher-order statistics on two frame-difference fields across three adjacent frames, the computed motion information are used to conduct change detection and generate one *motion* mask that consists of all the moving objects from all the depth layers involved at each view. It would be highly desirable, and challenging, to further differentiate them according to their residing depth layer to achieve *layered* segmentation. For that, multiple *depth-layer* masks are generated using our proposed disparity estimation method, one for each depth layer. By intersecting the motion mask and one depth-layer mask at any given layer-of-interest, the moving objects associated with the corresponding layer are then extracted. All the above-mentioned processes are repeatedly performed along the video sequence with a sliding window of three frames at a time. For demonstration, only the foreground and the background layers are considered in this paper, while the proposed method is generic and can be straightforwardly extended to more layers, once the corresponding depth-layer masks are made available. Experimental results have shown that the proposed layered moving-object segmentation method is able to segment the foreground and background moving objects separately, with high accuracy on their shape boundary. In addition, the required computational load is considered fairly inexpensive, since our design methodology is to generate masks and perform intersections for extracting the moving objects for each depth layer.

1. Introduction

Segmenting the moving objects of interest from a digital video sequence is a challenging task and instrumental to many object-based video processing applications that are frequently encountered in various application domains—computer vision, pattern recognition, digital entertainment, interactive multimedia, to name a few. For stereoscopic video, the technical challenges are greatly compounded, since the moving objects presented in various depth layers inevitably 'interferes' each other and makes the extraction of moving objects becoming much more difficult. Furthermore, it is quite likely that only the moving objects appeared in a specific depth layer might be needed and extracted for further rendering and manipulation. Such need could frequently arise in interactive video and animation, for example. Therefore, it is our generic goal to extract all the moving objects from a stereoscopic video sequence according to their associated depth layers, respectively; thus, it is coined as *layered moving-object segmentation* in this paper. To our best knowledge, such generic goal and solution can not be found in the literature, since existing moving-object segmentation methods—be it for monocular or stereoscopic video sequences, only demonstrate the extraction of "foreground" moving objects being separated from their background, and oftentimes still background only. It is highly expected that the segmentation performance will be greatly degraded, if the background is not stationary but containing multiple moving objects with large sizes. In this paper, a novel *layered* stereoscopic moving-object segmentation method that is able to extract moving objects appeared in any layer of stereoscopic sequence is proposed.

The key design methodology of our approach is that we shall only exploit minimum number of cues for conducting the pursued layered moving object segmentation task. For that, obviously, *motion* information for detecting 'moving' objects and *depth* information for differentiating which moving objects should be associated with which depth layer are needed, while completely avoiding the processing of those commonly-used cues, such as color, texture, and so on, to substantially reduce computational complexity. In our work, utilizing motion and depth information will generate

* Corresponding author.
E-mail addresses: chenyibin@sptdi.com (Y. Chen), chcai@hqu.edu.cn (C. Cai), ekkma@ntu.edu.sg (K.-K. Ma), 413260175@qq.com (X. Wang).

1047-3203/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2013.05.010

one *motion mask* and multiple *depth-layer masks* for the current frame, and these processing steps will be applied to individual frames along each view, respectively. The generated motion mask indicates where are those moving objects presented in the current frame. It is the individually generated depth-layer mask that provides the opportunity to further extract those moving objects from the motion mask that are considered belonging to a specific depth layer of interest. This can be done by simply intersecting the depth-layer mask under concern with the motion mask through a simple logic AND operation. The proposed method is simple, effective, and accurate—not only the moving objects have been correctly extracted for each depth layer, but also the shape boundary of each extracted moving object is in high accuracy.

The proposed layered moving-object segmentation method has another attractive merit—inexpensive computational load. Note that the proposed method is model-free and thus avoids the time-consuming model-updating process as commonly encountered in several well-known *video object plane* (VOP) segmentation methods. As mentioned earlier, only the minimum number of cues (i.e., motion and disparity) are involved, and no pre- and post-processing steps are required so far to produce those segmentation results as reported in this paper. All these are greatly beneficial to the reduction of computational load.

The rest of the paper is organized as follows. Section 2 provides an overview of relevant video object segmentation methods. Section 3 describes the proposed layered moving-object segmentation method for stereoscopic video sequence in detail, including how to generate motion masks and depth-layer masks, as well as a new disparity estimation that contributes the generation of multiple depth-layer masks. Section 4 presents some experimental results to demonstrate the efficiency and efficacy of the proposed method. The extracted video objects from the foreground layer and from the background layer, respectively, are served as performance evaluation results. Conclusion is drawn in Section 5.

2. Overview of moving-object segmentation

Conventional video object segmentation algorithms can be classified into three categories: (1) *region*-based spatial segmentation method [1,2], (2) *motion*-based temporal segmentation method [3,4], and (3) *spatio-temporal* segmentation method [7,8]. The common objective shared by these methods is to generate VOPs.

In the first category, Patras et al. [1] exploited watershed segmentation approach to generate a number of initial segments that were further labeled according to the computed motion information. The label field is then modeled as a Markov random field (MRF) for further conducting video-segmentation optimization through their proposed iterative motion estimation-labeling algorithm. Tsaig et al. [2] proposed a region labeling approach for automatically segmenting the foreground and background moving objects through extracted video object planes (VOPs). The segmentation problem is first formulated as graph labeling over a *region adjacency graph* (RAG), based on motion information. Like Patras et al. [1], the label field is also modeled as an MRF. However, both approaches tend to yield over-segmentation, which makes region classification more difficult and incurs higher computational load.

For the motion-based temporal segmentation approach, Meier and Ngan [3] presented a video sequence segmentation algorithm by utilizing a Hausdorff object tracker that matches a two-dimensional (2-D) binary model of the object against subsequent frames based on the Hausdorff distance. In this approach, an initial model is derived first, followed by iterative updating process. A morphological motion filter is then exploited for refining the generated moving-object masks, while removing those edges that, in fact, belong to the stationary background. However, the computational load yielded by this approach is fairly high, besides it may suffer from error propagation and accumulation issue due to its sequential frame-by-frame processing nature [5].

Chien et al. [4] proposed a fast and real-time (subjected to optimized parallel processing implementation) moving object segmentation method by developing a background registration technique. First, a reliable background image is constructed based on the accumulated frame difference information. The moving foreground is then separated from the background region by comparing the current frame with the constructed background image. A morphological gradient filter is applied to the input frames as a pre-processing step for reducing shadow effects once incurred in the background of the scene, while a post-processing step can be applied to the obtained moving-object mask for removing noise regions and smoothing shape boundary further. However, if the object possesses still parts and sparse texture, or if the color of object is similar to that of background, the segmentation result will be significantly degraded.

For the last category of approach on spatio-temporal segmentation, Criminisi et al. [7] proposed a bi-layer segmentation algorithm by fusing motion, color, and contrast cues together with spatial and temporal priors through a probabilistic approach. A *conditional random field* (CRF) is formed, and its parameters are determined by training the CRF model. The segmentation is performed by applying the *binary min-cut* operation with the use of prior knowledge provided from a second-order temporal hidden Markov chain. However, its segmentation performance will be degraded when the background contains large moving objects. Yin et al. [8] developed a bi-layer segmentation algorithm by incorporating more cues—fusing depth, color, texture, and motion information. The segmentation is also using the *binary min-cut* operation. Since all the CRF-based algorithms exploit the trained object detectors to extract shape or motion features, its performance strongly depends on the training process.

Note that all the above-mentioned methods were proposed for *monocular* video segmentation. For *stereoscopic* video segmentation, the depth (or disparity) information can be further utilized and incorporated with spatial and temporal information for segmenting moving objects for each depth layer. For that, Kolmogorov et al. [6–8] have shown that the foreground layer can be efficiently (near real-time) extracted from a binocular stereo video sequence by fusing depth and color/contrast cues; however, ambiguities encountered in stereo matching might yield unwanted artifacts. Wu et al. [9] proposed another bi-layer segmentation for stereo video sequences by utilizing color and texture information for obtaining initial spatial regions, followed by further utilizing optical-flow-based motion vectors to detect moving regions. The depth information and a spatial–temporal coherence check are further used to refine the previously obtained foreground and background regions. However, only the extraction of foreground object is discussed and demonstrated in [9]. The computational load is expected quite high as multiple cues and refinement processes are required. Also, the accuracy of shape boundary is quite poor.

A bi-layer segmentation method using motion and depth cues to detect and extract the moving object from a video sequence taken by a moving camera was proposed in [10]. The shortcoming of this method is that it is an offline approach and impossible to be used for conducting real-time video object segmentation. Since the background images of video frames are highly correlated among each other while the camera is a still one, they can be approximately represented by a low-ranked matrix. Furthermore, since the background images tend to be stationary, moving objects can be detected as outliers through this low-rank representation. For that, a segmentation method called *moving object detection* that is able to detect contiguous outliers in the low-rank representation