Pattern Recognition 48 (2015) 3477–3487

Contents lists available at ScienceDirect

Pattern Recognition

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

Unsupervised segmentation of highly dynamic scenes through global optimization of multiscale cues

Yinhui Zhang [a], Mohamed Abdel-Mottaleb [b,c], Zifen He [a,*]

[a] Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, 727 Jingming South Road, Chenggong, Kunming 650500, China
[b] Department of Electrical and Computer Engineering, University of Miami, 1251 Memorial Drive, Coral Gables, FL 33146, USA
[c] Effat University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 2 July 2014
Received in revised form
25 March 2015
Accepted 20 April 2015
Available online 29 April 2015

Keywords:
Image sequence segmentation
Dynamic scene
Unsupervised segmentation
Global optimization

ABSTRACT

We propose a novel method for highly dynamic scene segmentation by formulating foreground object extraction as a global optimization framework that integrates a set of multiscale spatio-temporal cues. The multiscale features consist of a combination of motion and spectral components at a pixel level as well as spatio-temporal consistency constraints between superpixels. To compensate for the ambiguities of foreground hypothesis due to highly dynamic and cluttered backgrounds, we formulate salient foreground mapping as a convex optimization of weighted total variation energy, which is efficiently solved by using an alternating minimization scheme. Moreover, the appearance and position spatio-temporal consistency constraints between superpixels are explicitly incorporated into a Markov random field energy functional for further refinement of the set of salient pixel-level foreground mapping. This work facilitates sequential integration of multiscale probability constraints into a global optimal segmentation framework to help address object boundary ambiguities in the case of highly dynamic scenes. Extensive experiments on challenging dynamic scene data sets demonstrate the feasibility and superiority of the proposed segmentation approach.

1. Introduction

Reliably segmenting image sequences into a set of regions with distinct homogeneous behavior is a fundamental problem for solving a number of interesting applications, such as object class recognition [1], activity pattern monitoring [2] as well as mobile robot stereo vision [3]. In this paper, we consider the problem of automatically segmenting salient foreground objects from image sequences in the presence of highly dynamic scenes. This is a very difficult problem because objects in such scenes can vary greatly in appearance, scale, position and number. Moreover, the backgrounds of image sequences captured in natural environments are usually cluttered and highly dynamic, which might include, for instance, ripping water, swaying trees, rapidly moving shadows and even flocks of birds.

In general, image sequence segmentation methods can be mainly divided into two categories, namely supervised approaches and unsupervised approaches. Early approaches to image sequence segmentation require annotating [4–6,20] the presence of a foreground object at some given frame locations by a user. These approaches are labor-intensive as they require hand-segmented training data and therefore would not scale to automatic processing of image sequences. Semi-supervised image sequence segmentation methods also require a user to select semantic regions in key frames. These methods achieve impressively accurate results, but the limitation is those results heavily depend on the selection of the labeled key frames, and therefore, cumbersome postprocessing, for example, through random decision forest training [7], is required to correct the results.

More recent image sequence segmentation approaches take into account classic background subtraction techniques [8,23] and make use of dynamic background texture modeling [9] for unsupervised segmentation. Barnichm and Van Droogenbroeck [10] determine whether a pixel belongs to the background by comparing its current value with past ones and propagate the value of a background pixel into background subtraction model. Rather than relying on significantly lower variances of Gaussian Mixture Model (GMM), Haines and Xiang [11] use Dirichlet process Gaussian mixture models to estimate background distributions and use them as an input to a model learning process for continuous update as scene changes. These methods are typically

* Corresponding author.
E-mail addresses: yinhui_z@163.com (Y. Zhang), mottaleb@miami.edu (M. Abdel-Mottaleb), zyhhzf1998@163.com (Z. He).

http://dx.doi.org/10.1016/j.patcog.2015.04.021
0031-3203/© 2015 Elsevier Ltd. All rights reserved.

based on the strong assumption that the dynamic backgrounds are changing slowly, which is not the case for background subtraction of highly dynamic scenes.

The large variations in appearance and location of objects that may appear in image sequence remain a concern for modern object segmentation approaches. Papazoglou and Ferrari [12] propose a Fast Object Segmentation (FOS) algorithm which attempts to build dynamic appearance models of the object and background under the assumption that they change smoothly over time. An advantage of this approach is that it may be possible to handle spatio-temporal cues on image patches in the labeling refinement stage. However, initializing inside object points merely on motion boundaries tends to produce a significant amount of false-positive seeds, especially in the case of highly dynamic scenes. Grundmann et al. [13] combine hierarchical cues by constructing a tree of spatio-temporal segmentation. This approach allows for subsequent selection from varying levels of granularity. Although good for handling hierarchical appearance cues, a strong limitation of this method is that it does not solve the foreground segmentation task on its own due to the oversegmentation of the scene.

A parallel line of recent work in unsupervised image sequence segmentation, which focuses primarily on the ranking of object proposals. Endres and Hoiem [14] generate bag of regions based on seeds and rank them using structured learning. Lee et al. [15] use consistent appearance and motion to rank hypothesis groups among object-like regions. Whereas Ma and Latecki [16] introduce intra and inter-frame constraints in a weighted region graph to find maximum weight superpixels. Alternatively, in [17] segmentation is performed using graph cuts and simple color cues, and the regions are ranked through classification based on gestalt cues with a simple diversity model. Most recently, object models [18] are built based on the primary object hypothesis regions. While the object proposal facilitates patch-level object hypotheses, in practice, all the above methods suffer from the high complexity of choosing object proposals in cases of even moderate video size due to highly redundant segmentations for the different regions of input image sequences.

This paper describes a new method for unsupervised segmentation by formulating foreground object extraction as a global optimization over a set of multiscale spatio-temporal features. The multiscale cues consist of a combination of motion and spectral components at pixel level as well as spatio-temporal consistency constraints between superpixels. Our contributions with respect to earlier work can be summarized as follows:

(1) We introduce a global optimization formulation of weighted total variation energy functional which combines both motion and spectral boundaries with object inside mappings. An effective alternating direction convex optimization scheme is adapted to solve the weighted total variation energy functional and provide high-quality pixel-level salient object mapping.
(2) Integrating pixel-level salient object mapping into a superpixel-based Markov random field facilitates sequential combination of multiscale spatio-temporal oriented probability constraints into the global optimal segmentation

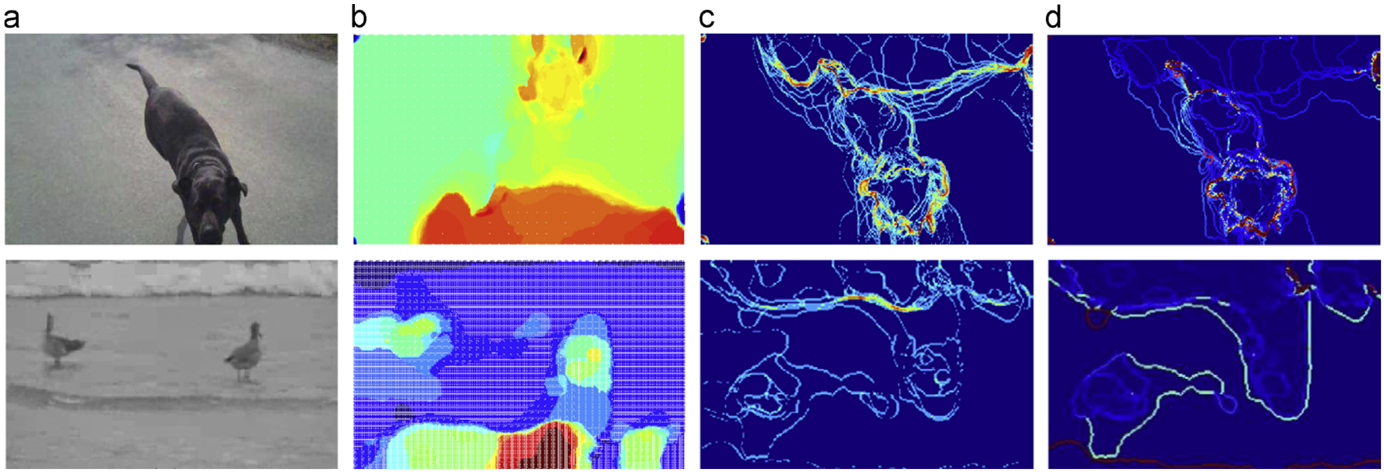

Fig. 1. Motion boundaries by optical flow on two image sequences. (a) The first frame of the *animal* and *birds* sequences. (b) The optical flow vector field $\overrightarrow{f}$ between two consecutive frames. (c) The gradient magnitude of optical flow vector $||\nabla\overrightarrow{f}||$. (d) Maximum motion angle between optical flow vectors $\overrightarrow{f}$ at pixel p and its neighbor q. Note that the values are linearly normalized to span the range [0, 1].

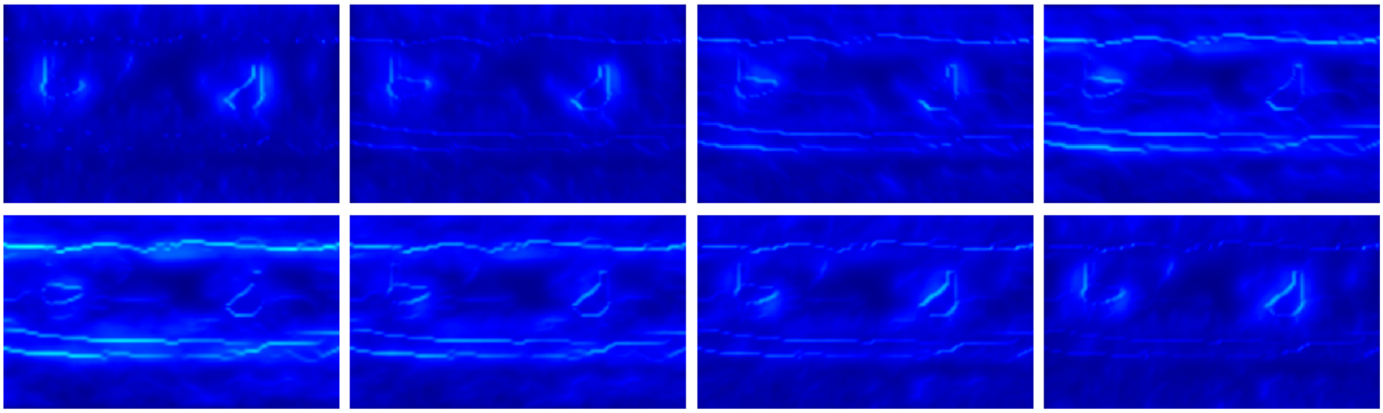

Fig. 2. Spectral boundaries of the first frame in the *birds* sequence at eight equally spaced orientations in the interval $\theta \in [0, \pi)$.