Unsupervised Video Object Foreground Segmentation and Co-Localization by Combining Motion Boundaries and Actual Frame Edges

Chao Zhang, University of Nottingham, Ningbo, China
https://orcid.org/0000-0002-9218-9028
Guoping Qiu, University of Nottingham, Nottingham, UK

ABSTRACT

In this article the authors proposed a fast and fully unsupervised approach for a foreground object co-localization and segmentation of unconstrained videos. This article first computes both the actual edges and motion boundaries of the video frames, and then aligns them by the proposed HOG affinity map approach. Then, by filling the occlusions generated by the aligned edges, the paper obtained more precise masks about the foreground object. With an accumulation process, these masks could be derived as the motion-based likelihood, which is used as a unary term in the proposed graph model. Another unary term is called color-based likelihood, which is computed by the color distribution of foreground and background. Experiment results shows the method is fast and effective to detect and segment foreground objects.

KEYWORDS
Activity Co-Localisation, HoG Affinity, Motion Boundaries, Salient Object Detection, Video Foreground Segmentation, Video Object Co-Localization

1. INTRODUCTION

The detection of video foreground objects is one of the key problems in video processing, which could directly facilitate the applications such as video object recognition (Prest, Leistner, Civera, Schmid, & Ferrari, 2012), action segmentation (Lu & Jason, 2015) and recognition (Wang & Schmid, 2013). In recent years, many
successful video foreground segmentation approaches (Grundmann, Kwatra, Han, & Essa, 2010), (Jang, Lee, & Kim, 2016) and object detectors (Cho, Kwak, Schmid, & Ponce, 2015; Girshick, 2015), are proposed, which leveraged the understanding of high-level video content.

However, building a fully unsupervised model that to segment the foreground object of unconstraint videos is still challenging, as no additional information about the foreground is provided; and the videos may be influenced by the factors such as dynamic background, motion blurry, light condition changes, or even the editing factors (e.g. subtitles or flying logos) (Papazoglou & Ferrari, 2013). Therefore, the dependencies for segmentation become more significant, including the motion cues and appearance cues. The optical flow is a popular example of motion cues (Figure 1).

With adequate alignments, the motion boundary by optical flow shall be more effective since it often related to the actual edges of foreground objects (Li, Kim, Humayun, Tsai, & Rehg, 2013). On the other hand, unsupervised models such as the PHM (Cho, Kwak, Schmid, & Ponce, 2015) could be regarded as good appearance model, while the colour distribution seems to be a more simple but efficient option (Koh, Jang, & Kim, 2016). In fact, a more precise foreground could also significantly benefit the foreground prediction with colour model (Stretcu & Leordeanu, 2015).

In this paper, we address the problem of automatically locating and segmenting the foreground object for unconstrained videos and proposed a fully unsupervised approach based on both motion and appearance cues of object. The key contributions of our work are: (1) we proposed a fully unsupervised approach for video foreground object segmentation, with competitive performances on three datasets; (2) we obtained more precise motion-based foreground predictions with a novel HOG affinity map; and (3) we show that the shallow image processing algorithms are still capable for complex vision tasks such as video foreground segmentation.

Figure 1. An illustration of proposed Motion-based likelihood: (a) The original frame; (b) The actual edges of current frame (green) and the motion boundaries (red); (c) The HOG affinity map; where the motion boundaries agreed more with edge responses have higher scores; (d) The edges that aligned by motion; (e) Initial prediction of foreground by Inside-outside-map (Papazoglou & Ferrari, 2013); (f) The motion-based likelihood by accumulating the masks obtained by step (e)
Comparison of Image Decompositions Through Inverse Difference and Laplacian Pyramids
[www.igi-global.com/article/comparison-of-image-decompositions-through-inverse-difference-and-laplacian-pyramids/124243?camid=4v1a](www.igi-global.com/article/comparison-of-image-decompositions-through-inverse-difference-and-laplacian-pyramids/124243?camid=4v1a)

Multi-Sensor Motion Fusion Using Deep Neural Network Learning
[www.igi-global.com/article/multi-sensor-motion-fusion-using-deep-neural-network-learning/187137?camid=4v1a](www.igi-global.com/article/multi-sensor-motion-fusion-using-deep-neural-network-learning/187137?camid=4v1a)

Design and Evaluation of a Content-Based Image Retrieval System
[www.igi-global.com/chapter/design-evaluation-content-based-image/8110?camid=4v1a](www.igi-global.com/chapter/design-evaluation-content-based-image/8110?camid=4v1a)