J. Vis. Commun. Image R. 24 (2013) 232–243

Contents lists available at SciVerse ScienceDirect

J. Vis. Commun. Image R.

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

Moving foreground object detection via robust SIFT trajectories

Shih-Wei Sun [a,b], Yu-Chiang Frank Wang [c,d,*], Fay Huang [e], Hong-Yuan Mark Liao [d,f]

[a] Dept. of New Media Art, Taipei National University of the Arts, Taipei, Taiwan
[b] Center for Art and Technology, Taipei National University of the Arts, Taipei, Taiwan
[c] Research Center for IT Innovation, Academia Sinica, Taipei, Taiwan
[d] Inst. of Information Science, Academia Sinica, Taipei, Taiwan
[e] Inst. of Computer Science and Info. Engineering, National Ilan University, Yi-Lan, Taiwan
[f] Dept. of Computer Science and Info. Engineering, National Chiao Tung University, Hsinchu, Taiwan

ARTICLE INFO

Article history:
Received 20 March 2012
Accepted 12 December 2012
Available online 21 December 2012

Keywords:
Template matching
Object tracking
Video object segmentation
Foreground segmentation
Background subtraction

ABSTRACT

In this paper, we present an automatic foreground object detection method for videos captured by freely moving cameras. While we focus on extracting a single foreground object of interest throughout a video sequence, our approach does not require any training data nor the interaction by the users. Based on the SIFT correspondence across video frames, we construct robust SIFT trajectories in terms of the calculated foreground feature point probability. Our foreground feature point probability is able to determine candidate foreground feature points in each frame, without the need of user interaction such as parameter or threshold tuning. Furthermore, we propose a probabilistic consensus foreground object template (CFOT), which is directly applied to the input video for moving object detection via template matching. Our CFOT can be used to detect the foreground object in videos captured by a fast moving camera, even if the contrast between the foreground and background regions is low. Moreover, our proposed method can be generalized to foreground object detection in dynamic backgrounds, and is robust to viewpoint changes across video frames. The contribution of this paper is trifold: (1) we provide a robust decision process to detect the foreground object of interest in videos with contrast and viewpoint variations; (2) our proposed method builds longer SIFT trajectories, and this is shown to be robust and effective for object detection tasks; and (3) the construction of our CFOT is not sensitive to the initial estimation of the foreground region of interest, while its use can achieve excellent foreground object detection results on real-world video data.

1. Introduction

Detecting moving foreground objects from a video taken by a non-stationary camera attracts intensive attention from researchers and engineers in the field of image and video processing. This is of particular interest for applications such as action and event recognition, and automatic annotation of videos. However, moving foreground detection has been a challenging task since the moving foreground object in real-world videos is often highly articulated or even non-rigid. Without prior knowledge (e.g., training data) on the foreground object of interest, it is difficult to model the object information even with user interaction. In practice, the camera is not fixed and thus conventional object detection methods based on frame differences cannot be applied, which makes background modeling approaches fail. In [1], Patwardhan et al. pointed out the three scenarios which make video foreground object detection very difficult. The first scenario is the presence of complex background which contains moving components such as water ripples or swaying trees. The second case is background motion caused by camera motion (e.g., shaky tripod in windy days), which rules out the use of conventional reconstruction-based approaches for object detection. Finally, most existing works for video object detection require training data or user interaction (e.g., at the first frame). This might not be practical and will result in increased processing time.

1.1. Related work

The history of video-based object detection starts from detection of moving objects in videos captured by a stationary camera. Jain and Nagel [2] proposed the frame difference scheme to detect a foreground object. Wren et al. [3] proposed the use of a Gaussian model, Stauffer and Grimson [4] proposed a GMM-based approach, and Elgammal et al. [5] applied kernel density estimation for background modeling. Unfortunately, the above methods cannot serve

* Corresponding author at: Research Center for IT Innovation, Academia Sinica, Taipei, Taiwan.
E-mail addresses: swsun@newmedia.tnua.edu.tw (S.-W. Sun), ycwang@citi.sinica.edu.tw (Y.-C.F. Wang), fay@niu.edu.tw (F. Huang), liao@iis.sinica.edu.tw (H.-Y.M. Liao).

1047-3203/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2012.12.003

well for scenarios in which the camera is moving (even with nominal motion). Recent researchers focus more on foreground object detection in videos captured by freely moving cameras. In [6], Sheikh and Shah proposed to build foreground and background models using a joint representation of pixel color and spatial structures between them. In [1], Patwardhan et al. decomposed a scene and used maximum-likelihood estimation to assign pixels into layers. From their experimental results, only moving foreground objects with the average velocity up to 12–15 pixels per frame can be detected. As a result, their approach is only capable of handling videos captured by a camera with mild camera motions.

In this paper, we address automatic video foreground object detection problems under arbitrary camera motion (e.g., panning, tilting, zooming, and translation). Prior methods focusing on this type of problem can be classified into two categories. The first category (e.g., Meng and Chang's method [7]) is to detect moving foreground object as the outliers, and thus to estimate the global motion of the camera [8]. Irani and Anandan [9] proposed a parametric estimation method for detecting the moving objects, and Wang et al. [10] also approached this problem in a similar setting. Furthermore, Bugeau and Perez [11] proposed motion and feature clustering techniques for estimating foreground object regions. The second category of moving object detection algorithms aims to model the reference background image. Felip et al. [12] proposed to estimate the dominant motion from the sampled motion vectors. Zhao et al. [13] proposed to detect objects present or removed from a non-static camera for indoor scenes based on the calculation of SIFT features [14] and homography. While it is possible to model the scene as background information for videos captured from an indoor scene or a closed area scene, modeling outdoor scene or more complicated background remains a very difficult problem.

The correspondences of feature points are widely used for linking the relationship between pairs of video frames. SIFT flow is recently proposed by Liu et al. [15] to determine the dense correspondences between image pairs for the retrieval of similar scenes. On the other hand, Sand and Teller [16] proposed a particle video approach, which is able to construct the long trajectory based on the optical flow correspondences, and thus provides more chances to detect and track foreground objects. We note that, although the method of Liu et al. [15] is able to determine dense corresponding SIFT points, it would be impractical to enforce the trajectory across all video frames, which results in linking SIFT points in dissimilar pairs of video frames. While the approach of Sand and Teller [16] better links corresponding particle points, its high computational cost would prohibit future speed-up or higher-level processing or learning tasks. In [17], a motion-flow based approach was proposed to analyze MPEG bitstreams for moving objects using the associated trajectories. Motivated by the above methods, we advance the context and spatiotemporal information of moving foreground objects, and we advocate the use of the trajectories to provide rich information in detecting foreground objects in videos.

1.2. Our contributions

We present a novel foreground object detection approach in this paper. We focus on detecting and tracking a single and dominant foreground object in uncontrolled videos, i.e., videos captured by freely moving cameras or those downloaded from the Internet. Based on the SIFT matching strategy, we calculate the foreground feature point probability for constructing robust SIFT trajectories, which imply the foreground candidate region across video frames without the need of user interaction such as parameter or threshold tuning. To perform foreground object detection, we propose a consensus foreground object template (CFOT) based on the extraction and association of SIFT points with longer trajectories and higher confidence. We note that our CFOT is not only derived by integrating the information of the candidate foreground regions in a probabilistic way, we also advance an adaptive re-start scheme to handle false object detection results when tracking the foreground object. This makes our CFOT more robust in detecting objects in real-world uncontrolled videos, and thus we are able to extract and track the foreground objects even when the contrast between the foreground and background regions is low (spatially or temporally). This is why our proposed method can be generalized to videos with dynamic backgrounds, and is robust to viewpoint changes across video frames.

The contribution of our proposed method is trifold: (1) we provide a probabilistic self-decision framework to determine the moving foreground objects in videos, while no user interaction or parameter tuning is required; (2) the extracted SIFT points across video frames allow us to associate the candidate foreground interest points and to calculate the foreground feature point probability for robust object detection; and (3) our CFOT results in a compact representation of the foreground object of interest, while the construction of CFOT is not sensitive to the initial estimation of foreground region of interest due to the re-start mechanism when necessary. From our experimental results, it can be verified that the use of our CFOT produces excellent foreground object detection results in real-world video data.

2. Foreground object detection

Fig. 1 shows the proposed framework for video foreground object detection. This framework consists of two steps: the construction of CFOT and the use of CFOT for foreground object detection, which will be discussed in Sections 2.1 and 2.2, respectively.

2.1. Construction of consensus foreground object template (CFOT)

Fig. 2 depicts the process for constructing the consensus foreground object template (CFOT) for detection purposes. We now detail each step in this subsection.

2.1.1. Foreground key point extraction

Scale-invariant feature transform (SIFT) [14] is a popular computer vision algorithm, which can be used to detect local interest points in an image. As an initial stage of our foreground feature point extraction, we apply the SIFT feature detector in each frame of a video sequence. The goal for this step is to obtain a set of foreground key points which most likely belong to the foreground object of interest, which is achieved by identifying the SIFT key points across video frames in a probabilistic point of view, as we discuss below.

As the initialization stage of this step, a new key point p_i^t is detected as a SIFT point for the first time at time t, and its corresponding probability $f_{point}(i,t)$ will be set as $\alpha = 0.5$ since we have no prior knowledge that whether this key point belongs to foreground or background. The foreground point probability function is defined as:

$$f_{point}(i,t) = \begin{cases} f_{point}(i,t-1)\cdot\lambda + 1\cdot(1-\lambda), & \text{if } p_i^t \neq \emptyset; \\ f_{point}(i,t-1)\cdot\lambda, & \text{if } p_i^t = \emptyset, \end{cases} \quad (1)$$

where λ is an update factor and is set to 0.95 as suggested in [18].

The above equation provides a probabilistic way to update the probability of assigning an extracted key point as foreground, depending on its key point matching history. To be more precise, if a SIFT key point is consistently identified across video frames (by SIFT matching), it is more likely to belong to the foreground object and thus a higher probability value will be assigned. Fig. 3a