Neurocomputing 131 (2014) 227–236

Contents lists available at ScienceDirect

Neurocomputing

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

Multi-cue based tracking

Qi Wang [a], Jianwu Fang [b], Yuan Yuan [b,*]

[a] Northwestern Polytechnical University, Xi'an 710072, Shaanxi, PR China
[b] Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, PR China

ARTICLE INFO

Article history:
Received 30 March 2013
Received in revised form
9 August 2013
Accepted 19 October 2013
Communicated by Xianbin Cao
Available online 27 December 2013

Keywords:
Computer vision
Kinect
Optical flow
Depth
Tracking

ABSTRACT

Visual tracking is a central topic in computer vision. However, the accurate localization of target object in extreme conditions (such as occlusion, scaling, illumination change, and shape transformation) still remains a challenge. In this paper, we explore utilizing multi-cue information to ensure a robust tracking. Optical flow, color and depth clues are simultaneously incorporated in our framework. The optical flow can get a rough estimation of the target location. Then the part-based structure is adopted to establish the precise position, combining both color and depth statistics. In order to validate the robustness of the proposed method, we take four video sequences of different demanding situations and compare our method with five competitive ones representing state of the arts. Experiments prove the effectiveness of the proposed method.

1. Introduction

Visual tracking is the central topic in computer vision. The aim of this operation is to identify the examined target in consecutive video frames consistently. To achieve this goal, the target object is usually labeled in the first frame by hand. Then its size and location are automatically determined in the following frames according to the initially labeled property. Since tracking is widely used in applications such as motion-based recognition, automated surveillance, vehicle navigation,human–computer interaction, and video content analysis, a great deal of efforts have been spent to develop various tracking algorithms [1–4].

Generally, approaches for tracking can be classified into three categories, *appearance* based, *motion* based and a *combination* of them. (1) For appearance based methods, the object is first described by the statistics of a predefined target template. This prior information could be global histograms or local keypoint descriptors. Then the tracker searches for a candidate target that is most similar to the template. Popular tracking strategies include point/region matching (SURF tracking [5,6], SIFT tracking [7], part-based tracking [8], sparse representation [9]), kernel tracking (mean-shift tracker [10], eigentracker [11]), and classifier-based tracking (MIL tracker [12], TLD tracker [13], SVM tracker [14]). (2) For motion based methods, the target movement is estimated in the first place. Then tracking is conducted according to the motion field. Typical example is the optical flow based tracking [15], where a dense velocity field is calculated from adjacent frames. Another example focuses on methodologies tailored for tracking specific objects, mostly humans [16]. In this case, human kinematic motions, such as jogging, running and stretching, are modeled particularly and they cannot be extended to other situations. (3) Though the tracking algorithms are classified into the above two categories, there are still a number of methods that do not correspond to any single prototype, but a combination of them [17–19]. These techniques consider the appearance and motion simultaneously and the tracking performance is much more promising.

However, the abundance of emerging tracking algorithms does not mean that this field has achieved perfect success. When problems of occlusion, illumination change and viewpoint variation occur, the accurate tracking in real applications still remains a challenge. This is because the appearance and motion properties at such conditions are different from their corresponding templates, which will cause the difficulty of between-frame association and lead to the drifting problem. Typical examples of these situations are illustrated in Fig. 1. Actually, these exceptional conditions might not be a problem for our human vision system. But the computer is incapable of such tasks. The reason, we think, derives from two aspects. (1) Firstly, the designed algorithms have no comparable learning ability with humans. Though popular machine learning techniques [22] demonstrate certain level of

* Corresponding author.
E-mail addresses: crabwq@gmail.com (Q. Wang), fangjianwu@opt.ac.cn (J. Fang), yuany@opt.ac.cn (Y. Yuan).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.10.021

Fig. 1. Typical examples of challenging video sequences from [8,20,21]. There are occlusion, illumination change, scale variation, and rotation in these sequences, which make the tracking task difficult.

generalizing and incremental ability, they are still limited. (2) Secondly, not all useful clues are properly utilized for processing [23,24]. This is similarly indicated by human visual mechanism that people make decisions based on multiple clues, such as color, texture, motion, depth and other prior information. Nevertheless, most algorithms employ not so many clues, which lead to a confused tracking result when the environment changes. For example, when a person walks on the street, the lighting condition may change from one place to another. If only color appearance is considered, the tracker cannot work efficiently. But once the motion or depth continuity is involved, the task becomes easier.

Based on the above considerations, we propose a tracking method based on multiple cues combination (TMC) in this paper. Optical flow, color and depth information are involved simultaneously. Our assumption is that different features can provide complementary supporting information. When a feature fails to track the target, the other features might act as supplemental evidences; or these features may enhance their individual effect together. The general idea of our method is as follows. In the beginning, the target object is manually labeled by a surrounding rectangle. The obtained template is used to determine the promising candidate target in the following frames. Then the optical flow field is calculated based on the two adjacent frames. The obtained displacement for each pixel can provide an estimation of its corresponding position in the next frame. After that, the target candidate is searched in the neighborhood of the estimated location. For every possible location, its appearance statistics is compared with that of the initially labeled template by a part-based model. The depth continuity is also considered in this process to make the result robust to noises from occlusion and illumination.

The main contributions of this work are as follows:

(1) Depth maps from Kinect sensor is utilized as a valuable clue for tracking objects. Though depth maps have been applied to existing applications as introduced in Section 2, most of them are based on stereo rig. This makes the speed less efficient because stereo algorithms usually employ an optimization process. After the introduction of Microsoft Kinect sensor, the situation has changed because the depth maps can be obtained in real time. But most related works based on Kinect are focused on the tracking of particular objects, such as hands and face. As for the general tracking problem, few works have been reported. Based on this consideration, we present a tracking method for general objects. No specific prior information is included for the tracking procedure, which makes the proposed tracker with wider application scope.

(2) Part-based model is rephrased in the context of depth information. For tracking problem, a part-based method has been recognized for its ability to restrain from occlusion. But this ability still has its limitation when confronting with challenging video sequences. In this work, we propose to employ depth information in the part-based method, together with traditional color statistics. To the best of the authors′ knowledge, this is the first time to extent the part-based tracking in this way.

(3) Several video sequences are recorded and labeled as the benchmark ground truth for depth-based tracking. Though there are data sets publicly available for tracking research [8,20,21], they only have the traditional RGB channels. On the contrary, the constructed data set in this work has both RGB and depth information for each frame.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 introduces the proposed multi-cue tracking model. Section 4 conducts extensive experiments to evaluate the presented method, based on the four video sequences taken by ourselves. In the end, conclusion is made in Section 5.