Pattern Recognition Letters 32 (2011) 1564–1571

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A new framework for on-line object tracking based on SURF

Quan Miao, Guijin Wang *, Chenbo Shi, Xinggang Lin, Zhiwei Ruan

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 15 October 2010
Available online 30 May 2011
Communicated by Y. Liu

Keywords:
Object tracking
Keypoint matching
On-line boosting
Adaptive classifiers
SURF

ABSTRACT

We present a new object tracking scheme by employing adaptive classifiers to match the corresponding keypoints between consecutive frames. The detection of interest points is a critical step in obtaining robust local descriptions. This paper proposes an efficient feature detector based on SURF, by incrementally predicting the search space, to enhance the repeatability of the tracked interest points. Instead of computing the SURF descriptor, we construct a classifier-based descriptor using on-line boosting. With on-line learning ability based on our sample weighting mechanism, the classifier maintains its discriminative power to establish robust feature description and reliable points matching for subsequent tracking. In addition, matching candidates are validated using improved RANSAC to ensure correct updates and accurate tracking. All of these ingredients contribute measurably to improving overall tracking performance. Experimental results demonstrate the robustness and accuracy of our proposed technique.

1. Introduction

Robust object tracking under real-world conditions is still an open issue and limits the use of state-of-the-art methods in applications ranging from visual surveillance to human–computer interfaces. The difficulties of object tracking include complicated object appearance variations, illumination change, partial occlusions and cluttered scenes.

Recently, tracking formulated as a classification problem has received a lot of attention due to its promising results. The classification-based tracking algorithms can be classified into two categories: region-based methods and feature-based methods. In case of the region-based methods (Avidan, 2004, 2005; Collins and Liu, 2005; Grabner et al., 2008), the basic idea is to learn a binary classifier which distinguishes the object from the background. The main advantage of region-based method is its relative robustness against illumination change, occlusion and cluttered scenes. However, these approaches have problems with complex transformations of the target object. In contrast, feature-based trackers (Grabner et al., 2007; Lepetit et al., 2005; Meltzer et al., 2004) are more adaptive to the object transformations. In (Lepetit et al., 2005; Lepetit and Fua, 2006), a feature-based tracker proposes randomized trees and ferns to discriminate keypoints from each other by classifiers. Although their algorithm demonstrates excellent empirical results, it entails learning a set of object changes before the tracking task begins. To achieve robust tracking with this method, it is imperative to collect a large set of training images covering the range of possible appearance variation, costing a considerable amount of time.

To cope with these problems, Grabner et al. (2007) propose an efficient tracking approach which employs the on-line boosting algorithm (Grabner and Bischof, 2006). However, such approaches typically operate on the premise that the model of the target object does not change drastically over time. The keypoints are detected using Harris corner which is sensitive to scale changes, not to mention more complex transformations. The tracker is prone to failure when significant appearance variations such as affine transformation and viewpoint change arise.

By contrast, in our earlier work (Miao et al., 2010), we propose a rough-but-robust feature-based tracking algorithm which fuses the keypoints' scale and rotation information into the on-line boosting technique. This paper further expands the original idea and thus provides in detail a new framework which fully improves the robustness of object tracking. Our contributions can be summarized as follows:

(1) To exploit the sequential patterns in the data, such as correlations between observations close in the sequence, we efficiently compute the SURF features in each video frame by incrementally predicting the object region.
(2) We employ the scale information and the dominant orientation of SURF feature to guide the discriminative learning process of the keypoints' description. This leads to a series of scale- and rotation-invariant classifiers that are able to cope with significant appearance variations between frames.
(3) Unlike standard RANSAC (Hartley and Zisserman, 2004), we employ a non-uniform sampling strategy according to the matching score of each classifier. That is, we consider the

* Corresponding author. Tel.: +86 10 62781291; fax: +86 10 62780317.
E-mail address: wangguijin@tsinghua.edu.cn (G. Wang).

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.05.017

matches with higher matching score more reliable and give them larger weight, to achieve efficient verification and robust estimation of the homography.

(4) We improve the on-line boosting technique by adaptively updating the classifiers. Discriminative samples are selected and assigned higher importance weights.

The remainder of this paper is organized as follows. Section 2 reviews the on-line boosting technique and gives a short survey on existing local feature detectors. A detailed description of the whole tracking framework is presented in Section 3. In Section 4, we illustrate how to implement the proposed algorithm and give a brief analysis of parameters setting. Section 5 is dedicated to experimental validations and Section 6 concludes this paper with remarks on potential extensions for future work.

2. Background and related work

2.1. On-line boosting

The underlying idea of boosting is to combine a set of well selected weak classifiers (Freund and Schapire, 1997) to form a strong classifier. After the seminal work of Viola and Jones (2001), boosting has been successfully used in many computer vision problems, such as human detection (Laptev, 2006), image retrieval (Tieu and Viola, 2004), face detection (Viola and Jones, 2004), etc.

Recently, there has been considerable research interest in on-line vision applications, in which the learning and updating phase are performed on-line as new samples arrive. Oza and Russell (2001) make the primary efforts on studying on-line boosting and demonstrate their equivalence to the off-line counterparts under particular conditions. Based on Oza and Russell (2001), Nair and Clark (2002) employ on-line boosting in a co-training framework for object detection, Collins and Liu (2005) apply on-line discriminative learning in object tracking and Grabner and Bischof (2006) propose a novel on-line boosting for feature selection, etc.

There is a rich literature in on-line boosting and a thorough discussion on this topic is beyond the scope of this paper. Here, we briefly review the most relevant on-line boosting algorithm (Grabner and Bischof, 2006) in which the strong boosted classifier C is composed of J selectors h_j^{sel}. Each classifier holds a binary weak classifier pool X from which the training procedure selects the ones with the minimal estimated error. The strong classifier wishes to predict the matching confidence measure of an unknown point $\boldsymbol{x}$ by:

$$C(\boldsymbol{x}) = conf(\boldsymbol{x}), \tag{1}$$

$$conf(\boldsymbol{x}) = \sum_{j=1}^{J} \alpha_j \cdot h_j^{sel}(\boldsymbol{x}) \Bigg/ \sum_{j=1}^{J} \alpha_j, \tag{2}$$

where the value $conf(\bullet)$ denotes the confidence measure. As new samples arrive sequentially, each selector h_j^{sel} is responsible for reselecting the best weak classifier and the corresponding voting weight α_j is updated.

During boosting learning, how to construct a robust weak classifier pool is an important issue. The method described in (Grabner and Bischof, 2006) uses the standard Haar-like features (Viola and Jones, 2001) computed in a fixed bounding patch centered at the corresponding keypoint, which can only deal with pure translations and slight rotations. This paper employs the scheme we proposed in (Miao et al., 2010) where the scale and the dominant orientation of the keypoint are incorporated in the weak classifier pool. In addition, each sample should bear an importance weight to indicate its contribution to the classifier update. Grabner's method gives all the samples equal weight. We emphasize the negative samples that are "similar" to the positive one, to make the updated classifiers more discriminative.

2.2. Feature detectors

Feature detectors, which provide the feature points to be matched (Li and Allinson, 2008), are widely utilized in a large number of applications such as image retrieval (Tuytelaars and Van Gool, 2004), image registration (Brown and Lowe, 2007), and object recognition (Lowe, 2004). Feature detectors can be traced back to the Moravec's corner detector (Moravec, 1977), and improved by Harris and Stephens (1988) to make it more repeatable under small image variations. However, Harris corners are very sensitive to changes in image scale, so it does not provide a good basis for matching images of different sizes. Lindeberg (1998) introduces the concept of automatic scale selection. Based on Lindeberg (1998), several approaches to scale-invariant interest point detection have been proposed, such as the detector based on Harris-Laplace and Hessian-Laplace by Mikolajczyk and Schmid (2001), Difference of Gaussians (DoG) in SIFT by Lowe (2004), and Hessians approximated in SURF by Bay et al. (2006). Matas et al. (2002) have also developed the maximally stable extremal region (MSER) detector, which is a watershed-like method.

In this paper, we use the SURF detector (Bay et al., 2006) to extract keypoints because of its high detection accuracy and full invariance to rotation and scale changes. Furthermore, it can be computed efficiently due to the use of integral images.

3. Proposed algorithm

Feature-based object tracking involves three consecutive steps: feature detection, feature description and feature matching. In feature detection, we incrementally detect keypoints based on SURF. Then we compute the classifier-based descriptions, followed by feature matching in which adaptive classifiers are employed.

The target object region is located in the first frame, either manually or by using an automated detector. When a new frame arrives, we establish matching candidates with the previous frame by means of the feature-based scheme mentioned above. The homography H is estimated using weighted RANSAC over the set of matching candidates. The on-line classifiers are updated to perform further target tracking in the subsequent frame. In the remainder of this section we will describe the algorithm shown in Fig. 1.

3.1. Local feature detection

As is pointed in (Mikolajczyk and Schmid, 2001), the repeatability of the Harris corner detector fails when image resolution changes significantly. In contrast, the SURF detector is more robust to variations. Ta et al. (2009) propose an incremental SURF detection scheme to detect matching candidates of each keypoint in a local neighborhood, aiming to make establishing feature correspondences easier. However, the neighborhood has to be three dimensional (including the scale space), which will take time to search. Moreover, it will be a waste of memory since there are often overlaps between the neighborhoods of different keypoints within the object.

In this subsection, we efficiently detect keypoints in each frame by predicting the object region. As feature matching is performed within the object region in our tracking scheme, predicting the target object means telling the possible range matching candidates are located in. Suppose we are observing a binary variable describing whether on a particular day it rains or not. If we consider the