J. Vis. Commun. Image R. 25 (2014) 1018–1030

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Model-based approach to spatial–temporal sampling of video clips for video object detection by classification ☆

Chi-Han Chuang [a,1], Shyi-Chyi Cheng [a,*], Chin-Chun Chang [a,1], Yi-Ping Phoebe Chen [b,2]

[a] Department of Computer Science and Engineering, National Taiwan Ocean University, 2 Pei-Ning Road, Keelung 202, Taiwan
[b] Department of Computer Science and Computer Engineering, La Trobe University, Victoria 3086, Australia

ARTICLE INFO

Article history:
Received 16 April 2013
Accepted 20 February 2014
Available online 13 March 2014

Keywords:
Semantic video objects
Spatial–temporal sampling
Human action detection
Video object model
Dynamic programming
Multiple alignment
Model-based tracking
Video object detetcion

ABSTRACT

For a variety of applications such as video surveillance and event annotation, the spatial–temporal boundaries between video objects are required for annotating visual content with high-level semantics. In this paper, we define spatial–temporal sampling as a unified process of extracting video objects and computing their spatial–temporal boundaries using a learnt video object model. We first provide a computational approach for learning an optimal key-object codebook sequence from a set of training video clips to characterize the semantics of the detected video objects. Then, dynamic programming with the learnt codebook sequence is used to locate the video objects with spatial–temporal boundaries in a test video clip. To verify the performance of the proposed method, a human action detection and recognition system is constructed. Experimental results show that the proposed method gives good performance on several publicly available datasets in terms of detection accuracy and recognition rate.

1. Introduction

Delineating the spatial–temporal boundaries of video objects in a video clip is one of the most important problems in computer vision due to its potential for many vision-based applications, such as video surveillance, man–machine interfaces, video indexing and retrieval, recognition of postures, analysis of sports events, and authoring of video games [1]. Recently, semantic-based video analysis tended to model a video clip as a graph whose nodes are high-level video objects performing a specific action individually [2]. Techniques of graph matching are then applied to annotate the event type of the input video clip [3]. Detection and classification of video objects from video clips help for bridging the semantic gap between high-level features and low-level features.

Conventional video object detection (VOD) algorithms, which characterize spatially cohesive objects with locally smooth trajectories, use the techniques for tracking or body pose estimation to extract spatial–temporal tubes from the input video clip [4–6]. However, when applied to real world videos, traditional techniques of tracking or body pose estimation are generally not reliable due to object occlusion, distortion and changes in lighting. Instead, we formulate the tracking process for VOD as a classification problem because objects across consecutive frames are, in general, spatially and temporally cohesive. Also, by assuming relatively slow camera motions, the shape and location of objects vary slowly from frame to frame. Thus, the size of the search space to track an object across many frames is reduced significantly by exploiting this coherence. By considering a parameter set in the feasible search space as a class, the object tracking for VOD is cast into a classification framework [7].

The image of an object often consists of several parts arranged in a deformable configuration [8]. The use of visual patterns of local patches in shape modeling is related to several ideas, including the approach of local appearance codebooks [9] and the generalized Hough transform (GHT) [10] for object detection. At training time, these methods learn a model of the spatial occurrence distributions of local patches with respect to object centers. At testing time, based on the trained object models, the appearances of points of interest in images or videos are matched against visual codebooks to detect a specific object using the Hough voting framework. The effectiveness of visual pattern grouping by Hough voting is heavily dependent on the quality of the learnt visual model, and thus the ability to precisely locate the target objects and extract typical features from training samples is very important.

☆ This work was supported in part by National Science Council Taiwan under Grant Nos. NSC 100-2221-E-019-054-MY3 and 101-2918-1-019-003.

* Corresponding author. Fax: +886 2 24623249.
E-mail addresses: csc@mail.ntou.edu.tw (S.-C. Cheng), phoebe.chen@latrobe.edu.au (Yi-Ping Phoebe Chen).
[1] Fax: +886 2 24623249.
[2] Fax: +61 394793060.

http://dx.doi.org/10.1016/j.jvcir.2014.02.014
1047-3203/

In this paper, we formulate a video object as a sequence of key-objects, where each is modeled by an effective visual dictionary [11]. A key-object is defined as the image object inside a key-frame when we use key-frames to present a video clip [12,13,53]. Given an activity class, we model the video object that performs the activity as a sequence of key-object visual dictionaries. At testing time, the system applies a well-known dynamic programming approach to detect the target video objects by optimally aligning the frames of the input video clip with the key-object dictionary sequence. More specifically, in this paper, the visual dictionary sequence of a class plays the role as an object template and the template matching approach with dynamic programming is used to locate the video objects in the video clips. The semantics of the detected video objects is further verified by a trained class-specific classifier.

The first challenge of the approach comes from the computational complexity at worst $O(n^3)$ for video objects of n frames, which leads to the learning algorithm of high time complexity. Thus, it is crucial to find a more efficient key-object representation for video objects. Another challenge in key-object representation of the video object is the computation of temporal boundaries among key-objects. To tackle the above problems, this paper presents the following contributions to video object detection and classification. Firstly, a learning algorithm used to generate a codebook sequence is proposed to detect and classify video objects from a video clip simultaneously. The computational issue of the learning approach is also discussed. Secondly, the use of dynamic programming together with the learnt key-object codebooks allows the optima alignment between the frames of a video sequence and the codebook sequence of a specific object class. Thirdly, based on the alignment result, every key-object codebook locates potential objects in the corresponding frames using the Hough voting. A simple object correspondence follows to track the detected objects across frames and locate multiple video objects in the input video. Next, a class-specific SVM classifier is used to filter the detected video objects that do not belong to the activity class. Thus, the model-based object tracking approach provides the ability to detect multiple video objects from the input video clip. Finally, the user interaction to manually delineate the target video object to learn a specific model sequence is reduced to a minimum. Experimental results show that the proposed method gives good performance on several publicly available datasets in terms of detection accuracy and recognition rate.

The remainder of this paper is organized as follows. Section 2 presents the related work for semantic video object detection and recognition. Section 3 define the problem of video object detection by classification and related background knowledge. Section 4 deals with the computational issues to create the video object model for an activity class in the learning phase and to detect video objects with the learnt codebooks in the testing phase. In this section, we also present the proposed codebook matching algorithm using dynamic programming to optimally select proper codebooks for interpreting the target objects of frames in a test video sequence. Section 5 describes the experimental tests to illustrate the effectiveness of the proposed method. Finally, conclusions are drawn in Section 6.

2. Related work

A common element of all previous VOD methods in the literature is that they estimate the object boundaries to delineate all objects in video frames by tracking [4,11]. A primary motivation for the work presented here is to question the benefits of tracking object boundaries across frames for video-based applications, such as activity analysis. In practice, the accuracy of any boundary estimate is limited by a number of systemic factors such as image resolution, noise, motion skew and the accuracy of the model. For example, formulating VOD as motion segmentation using optical flow rests on the assumption of brightness constancy, which is violated at moving boundaries, resulting in poor estimates of object contours [14]. For some applications, the object detection at each frame only needs to be known up to a limited precision, as long as good shape and trajectories are maintained.

In addition to segmentation, conventional algorithms of VOD also try to detect and segment the observed motions into semantic meaningful instances of particular activities from videos [15,16]. To reach this goal, recent approaches consider the detection and recognition of the video object as an extension of 2D object detection [8,9] with higher dimensionality. Some well-known approaches include space–time interest-point detectors [17] and bag-of-words models [18]. These techniques aim at employing a combination of local space–time features and global 3D shape features to estimate the space–time boundaries of a given video object. Two issues which are therefore of particular importance are dealing with local patch sampling and exploring the rich relationships among spatial–temporal "words" inherited from objects [19].

A video clip often consists of multiple video objects, each performing a certain activity. In most video event retrieval methods, computing the content differences between video frames is usually required and is a significant part of overall computation [20]. This heavy computational issue can be tackled by the following steps: detecting the video objects from the input video clip; annotating these objects with correct class labels to transform the video clip into a high-level graph; and performing graph matching [21] to compute the semantic difference between video clips [22]. Thus, video object classification is the key step in high-level video-based applications. Conventional machine learning techniques are applied to train state-of-the-art methods using a large, diverse set of manually annotated images. The typical level of annotation needed is a bounding-box for each object instance [8,23]. To ensure the performance of a detector, a large amount of annotated instances is generally needed [24]. Recently, object classification approaches borrowed from unlabeled or weakly annotated data have attracted much attention to reduce tedious manual annotation to a minimum [25–27]. However, training a detector without location annotation is very difficult and performance is still below fully supervised methods [28,29].

Another challenge for the investigated methods for video-based applications is the preservation of logical coherence between object detection and recognition. Current machine learning algorithms for video object classification apply the principles of similarity and classify the detected objects by maximizing the contrast between the classes [30]. The learnt object models are often incapable of reconstructing the modeled objects and therefore cannot be used to detect objects from video clips. To solve this problem, recent approaches consider the detection and recognition of the video objects simultaneously [15,16]. These techniques aim at employing a combination of local space–time features and global 3D shape features to estimate the space–time boundaries of a given video object. While Yao et al. concluded in [15] that dense sampling would be superior to sparse sampling based on interest-point detectors [31], they used fixed-size space–time patches which are not robust to temporal scaling in action detection and recognition.

Recent approaches for video object detection follow the following steps: a target object is initialized by human annotation or with a preexisting detector in one frame, and then a classifier is trained online to redetect the object in each frame [32]. A significant limitation to these approaches is that the trained classifier is a video-specific detector but not a generic class detector. In contrast, Ali et al. [33] proposed a semi-supervised boosting variant that exploits temporal consistency of video frames to learn a complex