Neurocomputing 171 (2016) 982–990

Contents lists available at ScienceDirect

Neurocomputing

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

A comparative study of video-based object recognition from an egocentric viewpoint

Mang Shao*, Danhang Tang, Yang Liu, Tae-Kyun Kim

Department of Electrical and Electronic Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom

ARTICLE INFO

Article history:
Received 8 December 2014
Received in revised form
2 July 2015
Accepted 13 July 2015
Communicated by Liang Lin.
Available online 31 July 2015

Keywords:
Object instance recognition
Egocentric video
Comparative study

ABSTRACT

Videos tend to yield a more complete description of their content than individual images. And egocentric vision often provides a more controllable and practical perspective for capturing useful information. In this study, we presented new insights into different object recognition methods for video-based rigid object instance recognition. In order to better exploit egocentric videos as training and query sources, diverse state-of-the-art techniques were categorised, extended and evaluated empirically using a newly collected video dataset, which consists of complex sculptures in clutter scenes. In particular, we investigated how to utilise the geometric and temporal cues provided by egocentric video sequences to improve the performance of object recognition. Based on the experimental results, we analysed the pros and cons of these methods and reached the following conclusions. For geometric cues, the 3D object structure learnt from a training video dataset improves the average video classification performance dramatically. By contrast, for temporal cues, tracking visual fixation among video sequences has little impact on the accuracy, but significantly reduces the memory consumption by obtaining a better signal-to-noise ratio for the feature points detected in the query frames. Furthermore, we proposed a method that integrated these two important cues to exploit the advantages of both.

1. Introduction

Video-based object recognition (VbOR) methods have emerged during the last decade, but attracted less attention than image-based methods. Its motivation, however, has been further established by a recent research [1], by revealing that how human brain can effortlessly interpret a multitude of objects with different identity-preserving transformations. After exposing a monkey's visual system to an artificial visual world without temporal contiguity, neuroscientists observed that inferior temporal cortex neurons began to lose their capacity for being transformation invariant. This strongly encourages the exploitation of temporal information in object recognition tasks. For instance, some recent studies attempted this by using learned trajectory descriptors [2,3] or viewpoint invariant features [4,5] during visual fixation in video clips from different aspects.

On the other hand, the exploitation of spatial cues, either in 2D image layouts [6] or 3D object structures [7], is a flourishing branch of object recognition. The viewpoint-invariant theorem [8] states that the essential component of object recognition, regardless of viewing conditions, is structural information. Encoding object structural information requires only a small amount of memory, yet it is capable of producing a multitude of object representations via their interrelations and mental rotations. In the field of computer vision, stereo vision is often utilised to obtain precise depth perception, and hence 3D structure. On top of that, some recent studies have obtained impressive performance by using multi-view images to reconstruct 3D information to support object recognition [9], semantic segmentation [10] and pose estimation tasks [11,7]. Recently, due to the growing use of wearable vision devices, e.g., *Google Glass*, research into egocentric videos has attracted more and more attention. As one of the useful source of spatial information, egocentric vision has the advantages of being controllable during capturing informative viewpoints and being more practical than turntable settings.

In this comparative study, our goal is to explore the potential usage of egocentric videos for training and as query sources for the recognition of rigid 3D objects in realistic scenes. In particular, we aim to exploit the temporal and spatial cues provided by egocentric videos and to answer the following questions. Are they helpful? If so, are they helpful in terms of accuracy or efficiency? Can they be combined? It is worth noting that there have been

* Corresponding author. Tel.: +447703729830.
E-mail address: ms2308@ic.ac.uk (Shao).

http://dx.doi.org/10.1016/j.neucom.2015.07.023
0925-2312/

recent advances in object *category* recognition [12–14], but only a small number of studies have investigated the problems of *instance* object recognition [15], particularly in egocentric videos [16,17]. Therefore we highlight our contributions as below:

- We captured a *Sculptures in Victoria and Albert (V&A) Museum* dataset from an egocentric viewpoint.
- We categorised and compared diverse state-of-the-art object recognition frameworks and their video-based extensions.
- We proposed a hybrid solution that combines the advantages of both temporal and spatial cues.

2. Methods

Given exemplar videos of target objects, the purpose of VbOR is to identify them in query videos. Due to the egocentric setting in our study, each video captured multiple views of only one target object that appeared roughly in the centre. Therefore, the whole video was assigned and recognised with one label. In this comparative study, we focused on the methods represented by the taxonomy shown in Fig. 1. In terms of utilising spatial information, these methods can be categorised mainly into 2D and 3D approaches. Among the 2D approaches, there are three different ways to represent videos: image-based, set-based and video-based. In image-based methods, each video is treated as independent images, where a straightforward combination of individual results is applied to obtain the final output of the video. In set-based methods, each video is treated as a set of unordered images with underlying mathematical structure, such as a manifold. In video-based approaches, each video is represented as a set of ordered images, i.e., with temporal information. By contrast, 3D-based VbOR utilises reconstructed 3D information from multi-view images. This is a relatively new area with only a small set of methods. Thus we consider these methods as a separate category. In the following subsections, we analyse the pros and cons of each framework. Comparative evaluation can be found in Section 5.

2.1. Image-based methods

To select representative image-based methods, we adopted three baselines from state-of-the-art object recognition frameworks based on their image classification techniques, i.e., (a) point-to-point (P2P), (b) image-to-image (I2I) and (c) point-to-class (P2C), as illustrated in Fig. 2. The image classification results are combined later via voting.

Point-to-point methods measure the similarity between two images based on their corresponding local image appearance, which is usually encoded by a feature descriptor.

In the seminal paper by Lowe [18], image classification was performed by matching a set of keypoints detected in image regions. Using robust fitting algorithms, e.g., *RANSAC* [19], the correspondences can be constrained further by dominant transformation between the matched pairs. This technique can improve the recognition precision significantly. But it may fail when there is no similar viewpoint in the database to a query image. Recent advances in graph matching [20,21] have relaxed the geometric constraint between point correspondences for articulated or deformable object recognition. However, these methods are generally computationally expensive and infeasible for large-scale problems.

Image-to-image methods compute the vector of visual word frequencies in images to facilitate similarity measurement. In general, I2I methods are efficient and suitable for large-scale problems because of the compactness of their image representations. The Euclidean distance in a feature space reflects the similarity between features. Thus we can also apply learning-based classifiers, e.g., linear support vector machine (SVM) and Random Forests, to facilitate a better generalisability and efficient recognition. I2I methods have been applied widely to various image classification tasks, e.g., scene recognition [6], image categorisation [22,23], object recognition [24] and video image retrieval [25]. These methods have achieved state-of-the-art performance on most publicly available benchmark datasets of image classification. [12,13]. However, despite the success of these methods, the vector quantisation process may degrade the discriminatory power of individual image features, which is crucial for instance recognition problems.

Point-to-class (P2C) methods have also achieved impressive results on several benchmark datasets in recent years. The concept was emphasised in [26] to sidestep the negative effects of vector quantisation in I2I methods, and later improved and extended in [27,28]. The basic idea is to directly measure the similarity between query features and training features in every object class without vector quantisation. Compared with P2P methods, P2C has better generalisability, because images are decomposed into image features that can be matched simultaneously across all training images. This approach is also suitable for large-scale problems because the feature-matching procedure can be accelerated to real-time using approximated nearest neighbour algorithms. The main drawback is that P2C methods are based on non-parametric classifiers and consequently consume more memory because all the features are retained.

2.2. Set-based methods

Set-based methods aim to capture the inherent characteristics of a set based on the assumption that the members of the set follow a particular statistical distribution, as shown in Fig. 3. For

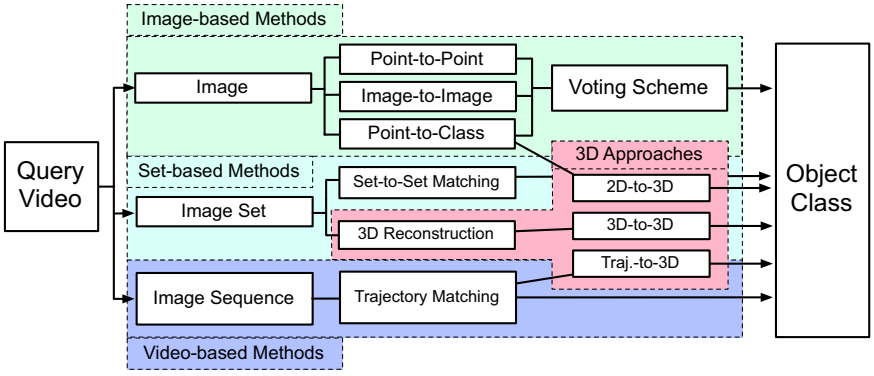

Fig. 1. Method categorisation and experimental setup.