Image and Vision Computing 52 (2016) 206–217

Contents lists available at ScienceDirect

Image and Vision Computing

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

Discovering object aspects from video ☆

Anestis Papazoglou*, Luca Del Pero, Vittorio Ferrari

University of Edinburgh, UK

ARTICLE INFO

Article history:
Received 13 September 2015
Received in revised form 19 March 2016
Accepted 21 April 2016
Available online 4 May 2016

Keywords:
Visual aspects
Object aspect discovery

ABSTRACT

We investigate the problem of automatically discovering the visual aspects of an object class. Existing methods discover aspects from still images under strong supervision, as they require time-consuming manual annotation of the objects' location (*e.g.* bounding boxes). Instead, we explore using video, which enables automatic localisation by motion segmentation. We introduce a new video dataset containing over 10,000 frames annotated with aspect labels for two classes: cars and tigers. We evaluate several strategies for aspect discovery using state-of-the-art descriptors (*e.g.* CNN), and assess the benefits of using automatic video segmentation. For this, we introduce a new protocol to evaluate aspect discovery directly, in contrast to the general trend of evaluating it indirectly (*e.g.* its impact on a recognition pipeline). Our results consistently show that leveraging the nature of video to discover visual aspects yields significantly more accuracy. Finally, we discuss two new applications to showcase the potential of aspect discovery: image retrieval of aspects, and learning aspect transitions from video.

1. Introduction

Traditionally, visual aspects have been defined as distinct viewpoints of rigid 3-D objects [1,2,3,4]. However, viewpoint alone cannot capture the appearance variations of complex, articulated objects in natural images. For example, tigers seen from a similar viewpoint can look very different due to articulated pose (*e.g.* a tiger lying and a tiger standing, Fig. 1). We use a broader notion of aspect that considers four factors of variation: viewpoint, articulated pose, occlusions and cropping by the image border. We explore the problem of automatically discovering such aspects from natural images of an object class. This task requires finding different object instances showing the same aspect (*e.g.* tigers running to the right, face close-ups, Fig. 1).

While some recent methods discover aspects from *still images* [5,6,7,8,9,10], they all require manual annotations of the object's location (*e.g.* bounding boxes). Location annotations allow focussing on the appearance of the object rather than the background, but they are expensive and time-consuming to create. In this paper instead we discover aspects from *video*, where we can segment the foreground objects from the background automatically, by exploiting motion [11,12,13]. Hence, it is possible to discover aspects under weak supervision, *i.e.* only one label per video shot is required. As an additional advantage, we can easily obtain video data for a large number of classes from several sources (*e.g.* DVDs, YouTube).

We present an extensive exploration of weakly-supervised aspect discovery in video, which we pose as an image clustering problem (Section 5). We measure the quality of the discovered aspects in terms of the compactness and diversity of the clustering (Section 6.1). We experiment with several modern appearance descriptors (SIFT [14], shape contexts [15], CNN features [16]), and various levels of spatial support (*e.g.* whole image, foreground segmentation). This enables to carefully evaluate the benefits of automatically segmenting objects (Section 6).

Our exploration relies on a new protocol for evaluating aspect discovery directly. In contrast, previous works evaluate aspect discovery indirectly, typically by measuring its impact on object detection performance [5,6,7,8]. For this, we collected a large dataset sourced from videos of two different classes, car and tiger (for a total of 2664 video shots, Section 4). The choice of the car and tiger classes allows us to explore two very different scenarios. Cars are rigid objects, and the major factors of aspect variations are different viewpoint, occlusions and croppings. Tigers display a broader range of different poses due to their complex articulation (Fig. 1). As an additional difference, cars exhibit higher intra-class variability in color and shape than tigers (*e.g.* different makes).

We annotated a few frames per shot with ground-truth aspect labels using an efficient labelling scheme (totalling over 10,000 frames, Section 3). This scheme captures the four factors of aspect variation by labelling simple, discrete properties of the object's physical parts. For example, we can distinguish between the top two

☆ This paper has been recommended for acceptance by Sinisa Todorovic.
* Corresponding author.
E-mail addresses: a.papazoglou@sms.ed.ac.uk (A. Papazoglou), ldelper@staffmail.ed.ac.uk (L. Pero), vferrari@staffmail.ed.ac.uk (V. Ferrari).

http://dx.doi.org/10.1016/j.imavis.2016.04.014
0262-8856/© 2016 Elsevier B.V. All rights reserved.

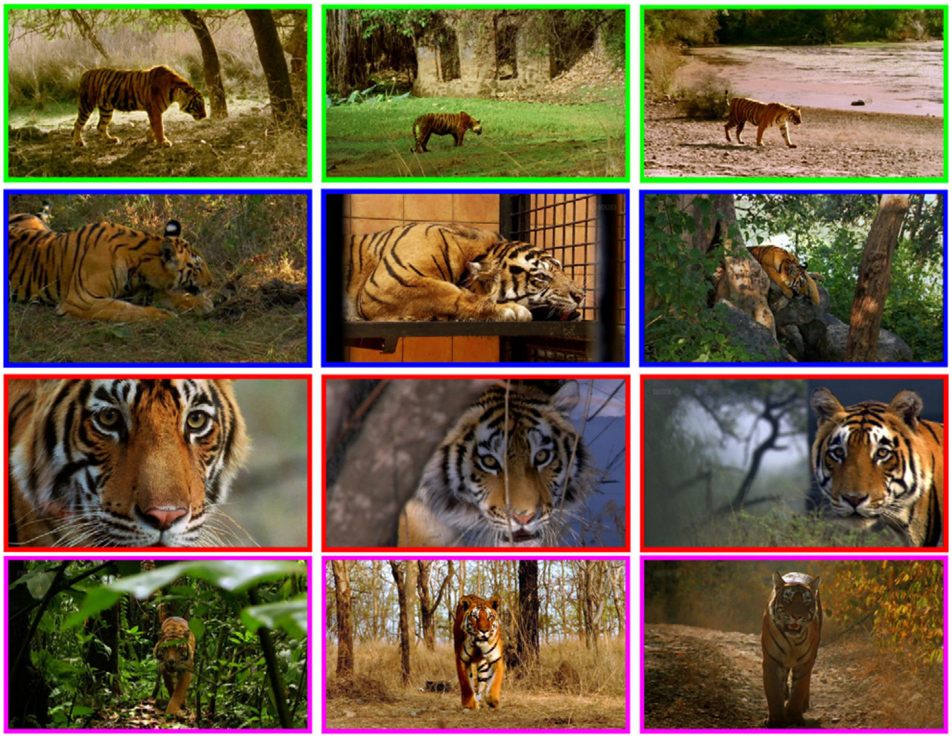

Fig. 1. Aspects discovered by our method (one per row). Despite showing tigers from the same viewpoint, the top two aspects look very different due to articulated pose and cropping. Our notion of aspect considers all these factors (Section 3).

aspects in Fig. 1 by considering that the hind legs are not visible in the second. We plan to release this dataset and the aspect labels.

Our experimental exploration demonstrates the great potential of using video for weakly supervised discovery (Section 6). In particular, the accuracy of the discovered aspects improves significantly if we use motion segmentation to get an estimate of the object location. After evaluating aspect discovery directly, we also show that it is useful for other applications. First, we use the aspects discovered by our system to enable a new kind of image retrieval based on aspects (Section 7.1). Second, we exploit the temporal nature of video to learn models of aspect transitions (*e.g.* from lying to standing, Section 7.2).

The rest of the paper is organized as follows. We start by discussing the two main components of our evaluation protocol: the labelling scheme (Section 3) and the dataset (Section 4). We then present several strategies for aspect discovery (from both videos and still images, Section 5) and present the results of our extensive exploration (Section 6). We conclude by introducing two applications that benefit from aspect discovery (Section 7).

2. Related work

2.1. Early work on aspects

Early work considered simple objects for which all possible aspects could be exhaustively enumerated [1,2,3]. More recently, Cyr and Kimia [4] tried to learn a manageable collection of representative views of an object instance. All these methods are limited to synthetic views of a single object instance.

2.2. Aspect discovery

Several methods [5,6,7,8,9,10,17,18,19] discover aspects implicitly, in order to train specialised classifiers for each of them (components of a mixture model). Some of these works [5,6,7,8] cluster HOG descriptors extracted from bounding boxes in the training images (manually annotated). Others [9,10] use exemplar SVMs [20] as a similarity measure between bounding boxes to drive the clustering. A few methods require additional time-consuming annotations, such as the location of object parts [17] or keypoints [18,19]. None of the methods above is weakly supervised. Moreover, while aspect discovery is a crucial intermediate step in their pipeline, it is evaluated only indirectly by measuring the performance improvement of the overall system.

2.3. Aspects in multi-view models

The works above use the discovered components in isolation. In contrast, other methods take the relationships between different aspects into account to build multi-view models [21,22,23,24,25]. They either require expensive bounding-box and viewpoint annotations for each training image [21,22,23] or very detailed 3-D CAD models [24,25]. Only the work of [26] uses video for this task. Their method is trained on a single short cellphone video per class, taken by walking around the object. While this procedure captures viewpoints well, it might fail to record other factors of variation, such as articulated pose. Moreover, it is not easily applicable for certain classes, such as wild animals. In practice, [26] only considers common rigid objects *i.e.* cars, motorbikes, wheelchairs, *etc.*

2.4. Modelling pose variations with parts

In the context of object detection and segmentation, some works [18,27,28] model variations in pose and articulation using poselets, *i.e.* parts that are tightly clustered in both appearance and configuration space (*e.g.* crossed hands, frontal face). This is somewhat related to our definition of aspects in terms of part properties (Section 3). However, learning poselets requires manual annotation