Computer Vision and Image Understanding xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Viewpoint refinement and estimation with adapted synthetic data

Pau Panareda Busto*,[a],[b], Juergen Gall[b]

[a] Airbus Group Innovations, TX4-ID, Munich, Germany
[b] University of Bonn, Computer Vision Group, Bonn, Germany

ARTICLE INFO

Keywords:
Domain adaptation
Pose estimation
Synthetic data

ABSTRACT

Estimating the viewpoint of objects in images is an important task for scene understanding. The viewpoint estimation accuracy, however, depends highly on the amount of training data and the quality of the annotation. While humans excel at labelling images with coarse viewpoint annotations like front, back, left or right, the process becomes tedious and the quality of the annotations decreases when finer viewpoint discretisations are required. To solve this problem, we propose a refinement of coarse viewpoint annotations, which are provided by humans, with synthetic data automatically generated from 3D models. To compensate between the difference between synthetic and real images, we introduce a domain adaptation approach that aligns the domain of the synthesized images with the domain of the real images. Experiments show that the proposed approach significantly improves viewpoint estimation on several state-of-the-art datasets.

1. Introduction

In order to estimate the viewpoint of objects in images precisely, an accurate annotation of the training data is required. Humans, however, perform poorly for estimating the viewpoint of an object accurately as illustrated in Fig. 1. Instead of annotating real images, synthetic data can be generated using 3D models (Marín et al., 2010; Mottaghi et al., 2015; Pishchulin et al., 2011; Sun and Saenko, 2014; Vázquez et al., 2014; 2011). While synthetic data provides accurate viewpoints, it either lacks the realism of real images or it is very expensive to generate. In particular, collecting a large variation of textured 3D shapes and combining them with coherent background scenes and illumination conditions is time-consuming.

We address this issue by leveraging human annotators and synthetic data, as depicted in Fig. 2, to avoid manual annotation by humans of fine viewpoints, which is time-consuming and erroneous, and to avoid the synthesis of a realistic dataset that captures the variations of real images, which is time and memory consuming. To this end, we ask humans to annotate only four coarse views, sketched in Fig. 3(a), and introduce an approach that refines the labels using synthetic data. Since synthetic data and real images belong to different domains as illustrated in Fig. 3(b), a domain adaptation approach is used for the refinement. General domain adaptation approaches like (Gong et al., 2012; Hoffman et al., 2013), however, are not sufficient for label refinement since they fail to distinguish viewpoint rotations by 180°. We therefore present a task-specific approach that takes advantage of the coarse labels of the real training samples.

A preliminary version of this work appeared in Busto et al. (2015). While the approach in Busto et al. (2015) was limited to cars, we extend the method to other categories and provide a thorough experimental evaluation. We also evaluate our approach with state-of-the-art features extracted from convolutional neural networks (CNN) (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014) and study the effect of truncated and occluded object instances. In addition, we also show how the refined datasets are able to obtain in some cases comparable or even better results than annotated training data with full human supervision. The evaluation, which is performed on six datasets for viewpoint estimation, reveals that our approach outperforms state-of-the-art domain adaptation methods.

2. Related work

2.1. Viewpoint estimation

Methods for viewpoint estimation are often based on popular object class detectors (Dalal and Triggs, 2005; Felzenszwalb et al., 2010; Girshick et al., 2014; Leibe et al., 2004) and learn a discrete set of pose classifiers. In Liebelt and Schmid (2010), Fidler et al. (2012), Pepik et al. (2012) and Hejrati and Ramanan (2014), annotations from 2D images are enhanced with 3D metadata to formulate 3D geometric models. On the contrary, Gu and Ren (2010) learns a mixture-of-templates that inherently captures the characteristics of projected views and Ozuysal et al. (2009) refines the hypothesis of 16 viewpoint detectors from 2D images with additional view specific Naïve Bayes classifiers.

* Corresponding author at: Airbus Group Innovations, Department of Data-Driven Technologies, Munich, Germany.
E-mail addresses: s6papana@uni-bonn.de, pau.panareda-busto@airbus.com (P. Panareda Busto), gall@iai.uni-bonn.de (J. Gall).

https://doi.org/10.1016/j.cviu.2018.01.005
Received 17 January 2017; Received in revised form 20 August 2017; Accepted 14 January 2018

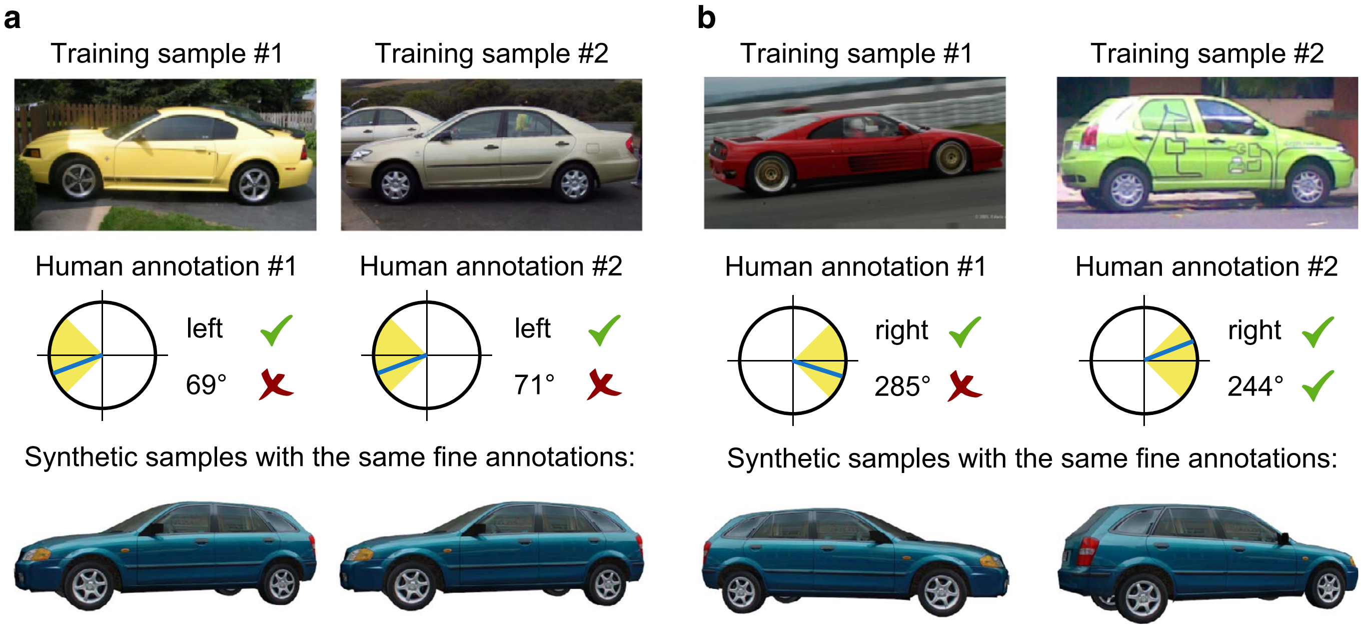


Fig. 1. Faulty annotations of fine viewpoints are introduced in human-annotated training datasets. While coarse labels like left or right are correct, the viewpoint annotations in degrees are not precise (a) and sometimes inconsistent (b) samples and fine annotations are taken from the Pascal3D+ dataset (Xiang et al., 2014).

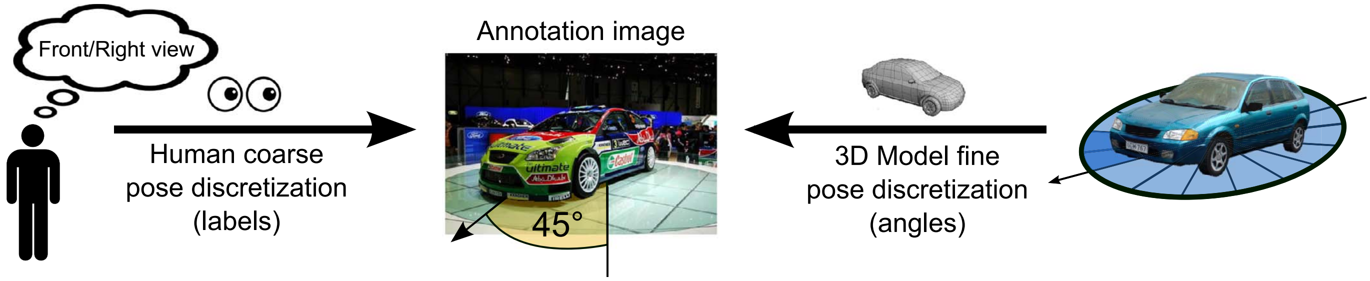
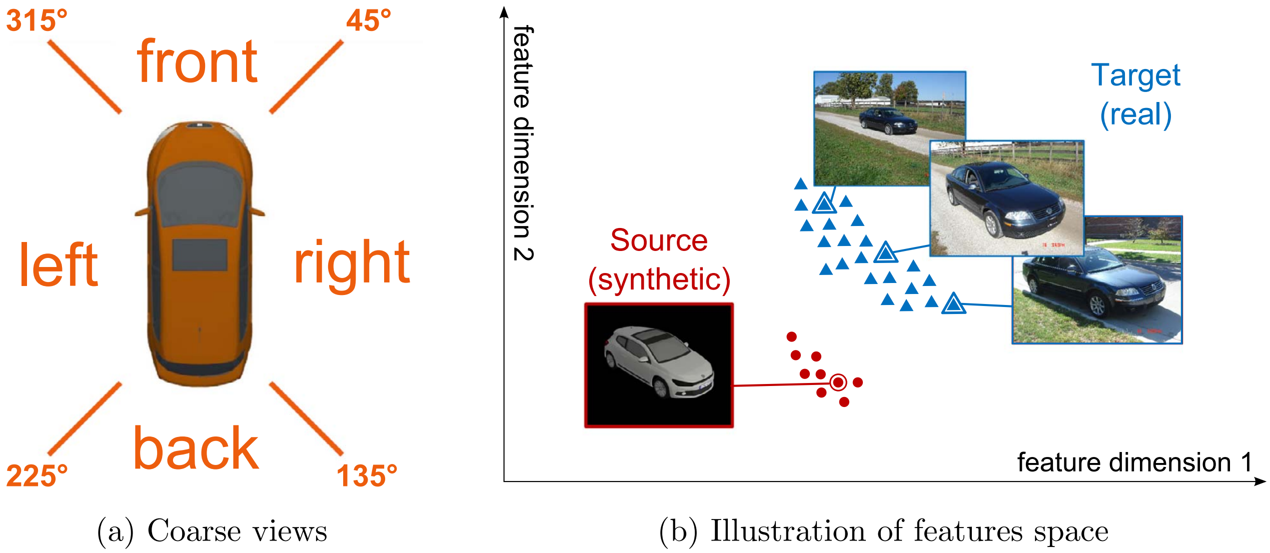


Fig. 2. Humans are perfect for annotating coarse viewpoints of objects in real images, but fail to estimate pose accurately at a fine level. 3D graphic models can be used to synthesize data at very accurate fine angles, but it is time-consuming to model all appearance variations present in real images. We therefore propose to leverage the abilities of humans of estimating coarse viewpoints and the pose accuracy of synthetic data.



(a) Coarse views

(b) Illustration of features space

Fig. 3. (a) The four views available for real images. (b) Synthetic and real images with the same annotated viewpoint lie in different domains within the feature space.

More recently, CNNs for object classification (Krizhevsky et al., 2012) have been retrained using 2D pose annotations in order to provide viewpoint probabilities as output channels coupled with the object class probability (Pepik et al., 2015; Tulsiani and Malik, 2015). In the study pursued in Ghodrati et al. (2014), simple frameworks that extract features from 2D bounding boxes with powerful encoders provided the same