Int J Comput Vis
DOI 10.1007/s11263-013-0646-8

Object and Action Classification with Latent Window Parameters

Hakan Bilen · Vinay P. Namboodiri · Luc J. Van Gool

Received: 1 October 2012 / Accepted: 18 July 2013

Abstract In this paper we propose a generic framework to incorporate unobserved auxiliary information for classifying objects and actions. This framework allows us to automatically select a bounding box and its quadrants from which best to extract features. These spatial subdivisions are learnt as latent variables. The paper is an extended version of our earlier work Bilen et al. (Proceedings of The British Machine Vision Conference, 2011), complemented with additional ideas, experiments and analysis. We approach the classification problem in a discriminative setting, as learning a max-margin classifier that infers the class label along with the latent variables. Through this paper we make the following contributions: (a) we provide a method for incorporating latent variables into object and action classification; (b) these variables determine the relative focus on foreground versus background information that is taken account of; (c) we design an objective function to more effectively learn in unbalanced data sets; (d) we learn a better classifier by iterative expansion of the latent parameter space. We demonstrate the performance of our approach through experimental evaluation on a number of standard object and action recognition data sets.

Keywords Object classification · Action classification · Latent SVM

This work was supported by the EU Project FP7 AXES ICT-269980.

H. Bilen (✉) · L. J. Van Gool
ESAT-PSI/iMinds, Ku Leuven, Kasteelpark Arenberg 10,
3001 Heverlee, Belgium
e-mail: hakan.bilen@esat.kuleuven.be

L. J. Van Gool
e-mail: luc.vangool@esat.kuleuven.be

V. P. Namboodiri
Alcatel-Lucent Bell Labs, Copernicuslaan 50, 2018
Antwerp, Belgium
e-mail: vinay.namboodiri@alcatel-lucent.com

L. J. Van Gool
Computer Vision Laboratory, ETH Zürich, Sternwartstrasse 7,
8092 Zurich, Switzerland

1 Introduction

In object detection, which includes the localization of object classes, people have trained their systems by giving bounding boxes around exemplars of a given class label. Here we show that the classification of object classes, i.e. the flagging of their presence without their localization, also benefits from the estimation of bounding boxes, even when these are not supplied as part of the training. The approach can also be interpreted as exploiting non-uniform pyramidal schemes. As a matter of fact, we demonstrate that similar schemes are also helpful for action class recognition.

In this paper we address the *classification* of objects (e.g. person or car) and actions (e.g. hugging or eating) (Pinz 2005) in the sense of PASCAL VOC (Everingham et al. 2007), i.e. indicating their presence but not their spatial/temporal localization (the latter is referred to as detection in VOC parlance). The more successful methods are based on a uniform pyramidal representation built on a visual word vocabulary (Boureau et al. 2010; Lazebnik et al. 2006; Wang et al. 2010). The focus then is often on the best features to use. In this paper, we augment the classification through an orthogonal idea, i.e. by adding more flexible spatial information. This will be formulated more generally as inferring additional unobserved or 'latent' dependent parameters. In particular, we focus on two such types of parameters:

Published online: 03 August 2013