Int J Comput Vis (2013) 101:437–458
DOI 10.1007/s11263-012-0549-0

# Random Forests for Real Time 3D Face Analysis

**Gabriele Fanelli · Matthias Dantone · Juergen Gall · Andrea Fossati · Luc Van Gool**

Received: 5 December 2011 / Accepted: 16 July 2012 / Published online: 1 August 2012


**Abstract** We present a random forest-based framework for real time head pose estimation from depth images and extend it to localize a set of facial features in 3D. Our algorithm takes a voting approach, where each patch extracted from the depth image can directly cast a vote for the head pose or each of the facial features. Our system proves capable of handling large rotations, partial occlusions, and the noisy depth data acquired using commercial sensors. Moreover, the algorithm works on each frame independently and achieves real time performance without resorting to parallel computations on a GPU. We present extensive experiments on publicly available, challenging datasets and present a new annotated head pose database recorded using a Microsoft Kinect.

G. Fanelli (✉) · M. Dantone · A. Fossati · L. Van Gool
Computer Vision Laboratory, ETH Zurich, Sternwartstrasse 7, 8092 Zurich, Switzerland
e-mail: fanelli@vision.ee.ethz.ch

M. Dantone
e-mail: dantone@vision.ee.ethz.ch

A. Fossati
e-mail: fossati@vision.ee.ethz.ch

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

J. Gall
Perceiving Systems Department, Max Planck Institute for Intelligent Systems, Spemannstrasse 41, 72076 Tübingen, Germany
e-mail: juergen.gall@tue.mpg.de

L. Van Gool
Department of Electrical Engineering/IBBT, K.U. Leuven, Kasteelpark Arenberg 10, 3001 Heverlee, Belgium

**Keywords** Random forests · Head pose estimation · 3D facial features detection · Real time

## 1 Introduction

Despite recent advances, people still interact with machines through devices like keyboards and mice, which are not part of natural human-human communication. As people interact by means of many channels, including body posture and facial expressions, an important step towards more natural interfaces is the visual analysis of the user's movements by the machine. Besides the interpretation of full body movements, as done by systems like the Kinect for gaming, new interfaces would highly benefit from automatic analysis of facial movements, as addressed in this paper.

Recent work has mainly focused on the analysis of standard images or videos; see the survey of Murphy-Chutorian and Trivedi (2009) for an overview of head pose estimation from video. The use of 2D imagery is very challenging though, not least because of the lack of texture in some facial regions. On the other hand, depth-sensing devices have recently become affordable (e.g., Microsoft Kinect or Asus Xtion) and in some cases also accurate (e.g., Weise et al. 2007).

The newly available depth cue is key for solving many of the problems inherent to 2D video data. Yet, 3D imagery has mainly been leveraged for face tracking (Breidt et al. 2011; Cai et al. 2010; Weise et al. 2009a, 2011), often leaving open issues of drift and (re-)initialization. Tracking-by-detection, on the other hand, detects the face or its features in each frame, thereby providing increased robustness.

A typical approach to 3D head pose estimation involves localizing a specific facial feature point (e.g., one not affected by facial deformations like the nose) and determining