Multi-scale Deep Learning for Gesture Detection and Localization

Natalia Neverova[1,2(✉)], Christian Wolf[1,2], Graham W. Taylor[3], and Florian Nebout[4]

[1] Université de Lyon, CNRS, Lyon, France
{natalia.neverova,christian.wolf}@liris.cnrs.fr
[2] INSA-Lyon, LIRIS, UMR5205, 69621 Villeurbanne cedex, France
[3] University of Guelph, Guelph, Canada
gwtaylor@uoguelph.ca
[4] Awabot, Lyon, France
florian.nebout@awabot.com

Abstract. We present a method for gesture detection and localization based on multi-scale and multi-modal deep learning. Each visual modality captures spatial information at a particular spatial scale (such as motion of the upper body or a hand), and the whole system operates at two temporal scales. Key to our technique is a training strategy which exploits i) careful initialization of individual modalities; and ii) gradual fusion of modalities from strongest to weakest cross-modality structure. We present experiments on the *ChaLearn 2014 Looking at People Challenge* gesture recognition track, in which we placed first out of 17 teams.

Keywords: Gesture recognition · Multi-modal systems · Deep learning

1 Introduction

Visual gesture recognition is one of the central problems in the rapidly growing fields of human-computer and human-robot interaction. Effective gesture detection and classification is challenging due to several factors: cultural and individual differences in tempos and styles of articulation, variable observation conditions, the small size of fingers in images taken in typical acquisition conditions, noise in camera channels, infinitely many kinds of out-of-vocabulary motion, and real-time performance constraints.

Recently, the field of deep learning has matured and made a tremendous impact in computer vision, demonstrating previously unattainable performance on the tasks of object detection, localization [1,2], recognition [3] and image segmentation [4,5]. Convolutional neural networks (ConvNets) [6] have excelled on several scientific competitions such as ILSVRC [3], Emotion Recognition in the Wild (EmotiW 2013) [7], Kaggle Dogs vs. Cats [2] and Galaxy Zoo. Taigman et al. [8] recently claimed to have reached human-level performance using ConvNets for face recognition. On the other hand, extending these models to problems involving the understanding of *video* content is still in its infancy, this idea

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part I, LNCS 8925, pp. 474–490, 2015.
DOI: 10.1007/978-3-319-16178-5_33