Comparing random forest approaches to segmenting and classifying gestures

Ajjen Joshi, Camille Monnier, Margrit Betke, Stan Sclaroff

PII: S0262-8856(16)30099-3
DOI: doi: 10.1016/j.imavis.2016.06.001
Reference: IMAVIS 3516

To appear in: *Image and Vision Computing*

Received date: 1 October 2015
Revised date: 30 May 2016
Accepted date: 5 June 2016

Please cite this article as: Ajjen Joshi, Camille Monnier, Margrit Betke, Stan Sclaroff, Comparing random forest approaches to segmenting and classifying gestures, *Image and Vision Computing* (2016), doi: 10.1016/j.imavis.2016.06.001

Comparing Random Forest Approaches to Segmenting and Classifying Gestures

Ajjen Joshi[1], Camille Monnier[2], Margrit Betke[1] and Stan Sclaroff[1]

[1] *Department of Computer Science, Boston Univeristy, Boston, MA 02215 USA*
[2] *Charles River Analytics, Cambridge, MA 02138 USA*
Corresponding Author E-mail: ajjendj@bu.edu

Abstract

A complete gesture recognition system should localize and classify each gesture from a given gesture vocabulary, within a continuous video stream. In this work, we compare two approaches: a method that performs the tasks of temporal segmentation and classification simultaneously with another that performs the tasks sequentially. The first method trains a single random forest model to recognize gestures from a given vocabulary, as presented in a training dataset of video plus 3D body joint locations, as well as out-of-vocabulary (non-gesture) instances. The second method employs a cascaded approach, training a binary random forest model to distinguish gestures from background and a multi-class random forest model to classify segmented gestures. Given a test input video stream, both frameworks are applied using sliding windows at multiple temporal scales. We evaluated our formulation in segmenting and recognizing gestures from two different benchmark datasets: the NATOPS dataset of 9,600 gesture instances from a vocabulary of 24 aircraft handling signals, and the CHALEARN dataset of 7754 gesture instances from a vocabulary of 20 Italian communication gestures. The performance of our method compares favorably with state-of-the-art methods that employ Hidden Markov Models or Hidden Conditional Random Fields on the NATOPS dataset. We conclude with a discussion of the advantages of using our model for the task of gesture recognition and segmentation, and outline weaknesses which need to be addressed in the future.