Electronic Letters on Computer Vision and Image Analysis 14(3):26-28, 2015

From pixels to gestures: learning visual representations for human analysis in color and depth data sequences

Antonio Hernández-Vela

Dept. of Applied Mathematics and Analysis, Universitat de Barcelona & Computer Vision Center, Barcelona, Spain
Advisor/s: Sergio Escalera and Stan Sclaroff (Boston University)
Date and location of PhD thesis defense: 9 March 2015, Universitat de Barcelona

Received 27th February 2015; accepted 14th May 2015

Abstract

The visual analysis of humans from images is an important topic of interest due to its relevance to many computer vision applications like pedestrian detection, monitoring and surveillance, human-computer interaction, e-health or content-based image retrieval, among others.

In this dissertation we are interested in learning different visual representations of the human body that are helpful for the visual analysis of humans in images and video sequences. To that end, we analyze both RGB and depth image modalities and address the problem from three different research lines, at different levels of abstraction; from pixels to gestures: human segmentation, human pose estimation and gesture recognition (see Fig. 1).

First, we show how binary segmentation (object vs. background) of the human body in image sequences is helpful to remove all the background clutter present in the scene. The presented method, based on Graph cuts optimization, enforces spatio-temporal consistency of the produced segmentation masks among consecutive frames. Secondly, we present a framework for multi-label segmentation for obtaining much more detailed segmentation masks: instead of just obtaining a binary representation separating the human body from the background, finer segmentation masks can be obtained separating the different body parts.

At a higher level of abstraction, we aim for a simpler yet descriptive representation of the human body. Human pose estimation methods usually rely on skeletal models of the human body, formed by segments (or rectangles) that represent the body limbs, appropriately connected following the kinematic constraints of the human body. In practice, such skeletal models must fulfill some constraints in order to allow for efficient inference, while actually limiting the expressiveness of the model. In order to cope with this, we introduce a top-down approach for predicting the position of the body parts in the model, using a mid-level part representation based on Poselets.

Finally, we propose a framework for gesture recognition based on the bag of visual words framework. We leverage the benefits of RGB and depth image modalities by combining modality-specific visual vocabularies in a late fusion fashion. A new rotation-variant depth descriptor is presented, yielding better results than other state-of-the-art descriptors. Moreover, spatio-temporal pyramids are used to encode rough spatial and temporal

Correspondence to: <ahernandez@cvc.uab.cat>

Recommended for acceptance by Jorge Bernal
DOI http://dx.doi.org/10.5565/rev/elcvia.723
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

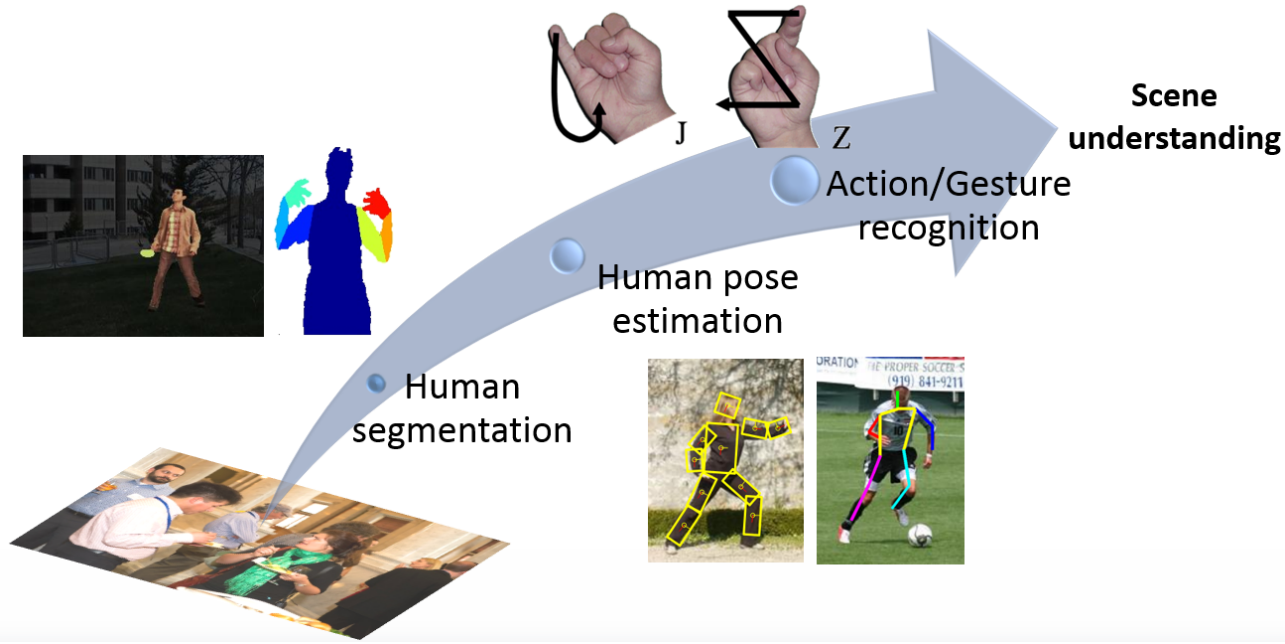

Figure 1: Different levels of abstraction for human analysis: Human segmentation, human pose estimation and gesture recognition.

structure. In addition, we present a probabilistic reformulation of Dynamic Time Warping for gesture segmentation in video sequences. A Gaussian-based probabilistic model of a gesture is learnt, implicitly encoding possible deformations in both spatial and time domains.

References

[1] A. Hernández-Vela, S. Sclaroff and S. Escalera. Poselet-based Contextual Rescoring for Human Pose Estimation via Pictorial Structures. *International Journal of Computer Vision*, 2014, Under review.

[2] A. Hernández-Vela, S. Sclaroff and S. Escalera. Contextual Rescoring for Human Pose Estimation. In *British Machine Vision Conference*, 2014.

[3] A. Hernández-Vela, M.A. Bautista, X. Pérez-Sala, V. Ponce-López, S. Escalera, X. Baró, O. Pujol and C. Angulo. Probability-based Dynamic Time Warping and Bag-of-Visual-and-Depth-Words for Human Gesture Recognition in RGB-D. *Pattern Recognition Letters*, 2013.

[4] A. Hernández-Vela, M. Reyes, V.Ponce and S. Escalera. GrabCut-Based Human Segmentation in Video Sequences. *Sensors*, 2012.

[5] A. Hernández-Vela, C. Gatta, S. Escalera, L. Igual, V. Martín-Yuste, M. Sabaté and P. Radeva. Accurate coronary centerline extraction, caliber estimation and catheter detection in angiographies, *IEEE Transactions on Information Technology in Biomedicine*, 2012.

[6] A. Hernández-Vela, N. Zlateva, A. Marinov, M. Reyes, P. Radeva, D. Dimov and S. Escalera. Human Limb Segmentation in Depth Maps based on Spatio-Temporal Graph Cuts Optimization. *Journal of Ambient Intelligence and Smart Environments (JAISE)*, 2012.

[7] A. Hernández-Vela, N. Zlateva, A. Marinov, M. Reyes, P. Radeva, D. Dimov and S. Escalera. Graph Cuts Optimization for Multi-Limb Human Segmentation in Depth Maps. In *IEEE Computer Vision and Patern Recognition*, 2012.