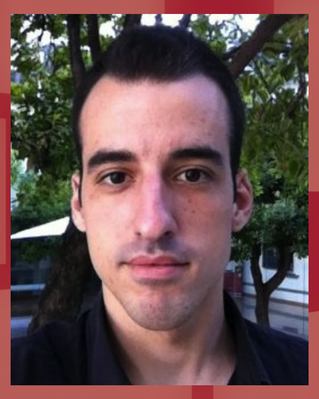
Cristina Palmero, Albert Clapés, Chris Bahnsen, Andreas Møgelmose
Advisors: Sergio Escalera, Tomas B. Moeslund

Centre de Visió per Computador

Universitat Autònoma de Barcelona



Tri-modal Human Body Segmentation

Abstract

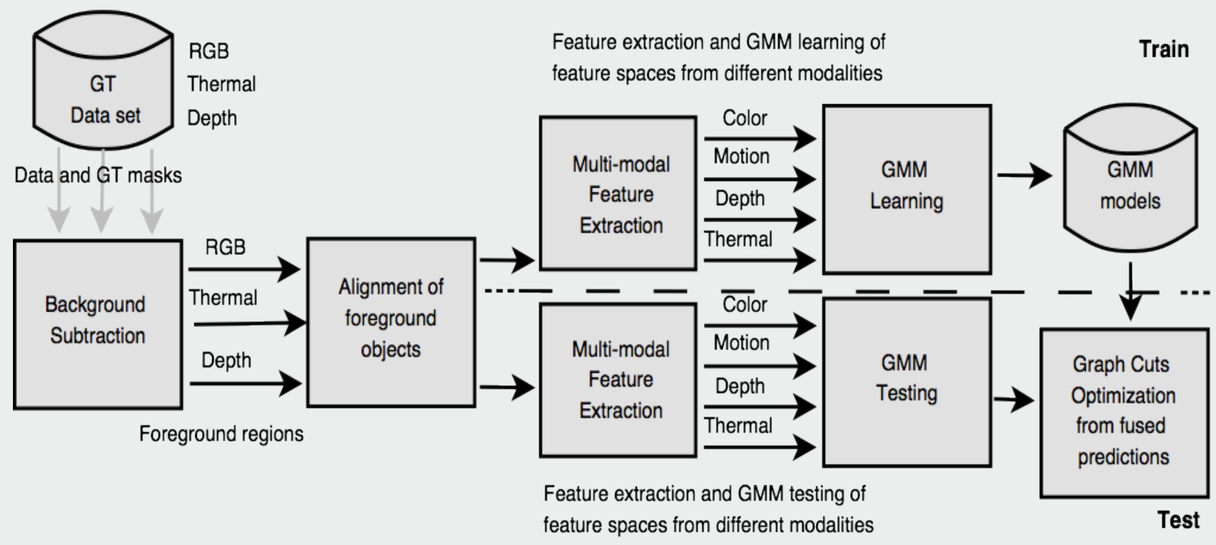
Segmentation of people in images is still nowadays a very challenging and difficult problem in computer vision. There exist lots of possible applications for people segmentation such as surveillance, patient caregiving or human-computer interaction. State-of-the-art approaches mainly consist in the usage of color images recorded by cameras or, more recently, with the launch of RGB-Depth devices - such as Microsoft® Kinect™ -, the usage of detph maps in combination with information provided by the color cue. In this context, we propose adding a third modality that is the thermal imagery got from thermal infrared cameras, thus complementing other information sources and making easier the segmentation task [1]. Although thermal cameras are relatively expensive devices, their market price is lowering substantially every year. The main contribution of this paper is a novel tri-modal database of people acting in three different scenes, consisting of more than 2000 frames each one, in which three different subjects appear and interact with objects performing different actions such as reading, working on a laptop or speaking by phone. In addition, a human segmentation baseline methodology is proposed. Having the modalities already registered (from a previous work), background subtraction is initially performed in each of the modalities in order to extract candidate subject and object regions. Then, in the training phase, subject regions are described at pixel-level using particular descriptors in each of the modalities [2,3]. Once the subject pixels have been described in all the modalities, Gaussian Mixture Models (GMM) are learnt. These GMMs are the ones used in the testing phase to compute the probabilities of being a subject pixel in the different modalities. Finally, the probability maps are fused in an optimization graph-cuts framework [4].

Method

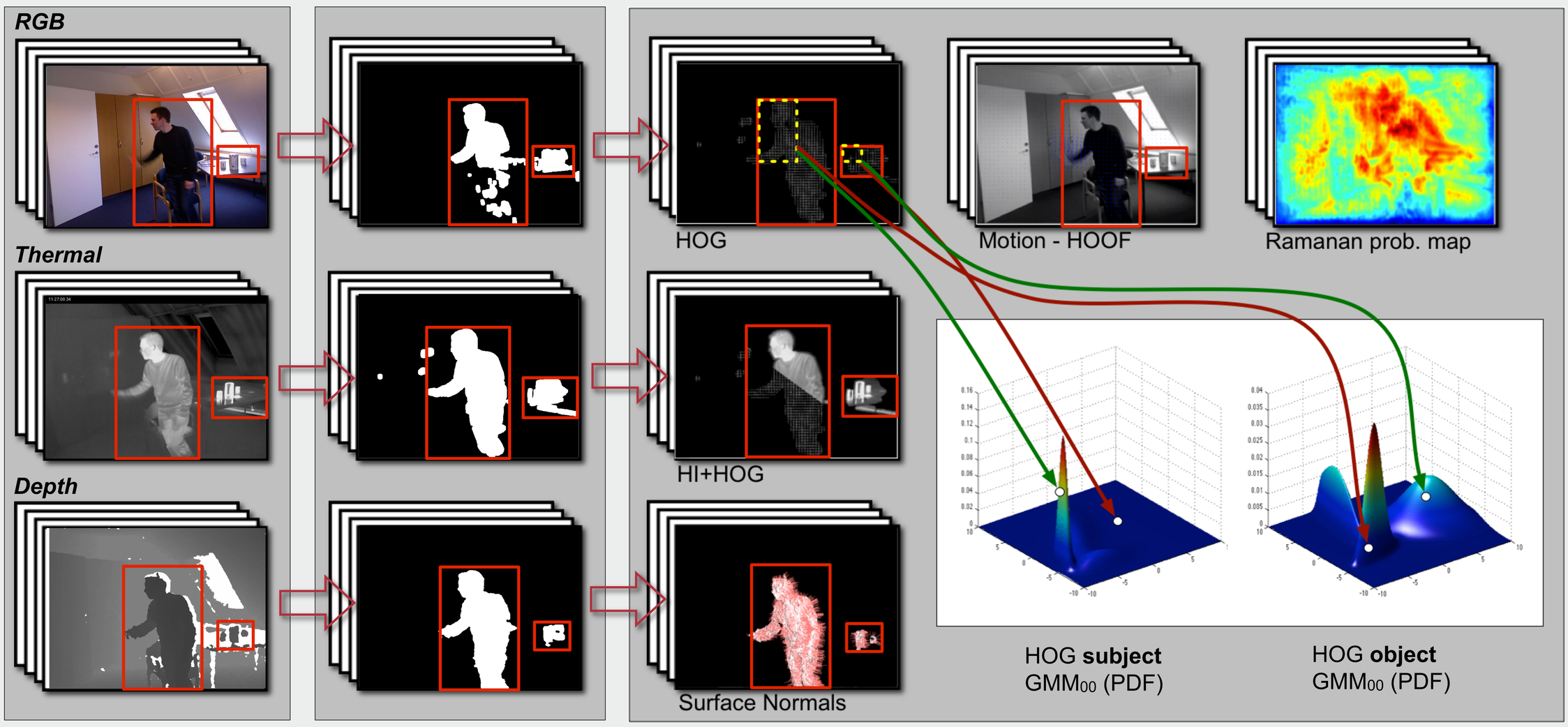
- **Background subraction** is performed in each modality separately. Then, foreground regions are fused and aligned by re-scaling of nearest detected regions.
- The **segmented foreground regions** are divided in a m-by-n **grid of cells**.
- **Pixels are described** using particular descriptors in the different modalities.
- A set of **GMMs are trained**; one for each of the cells and for each modality. Moreover, subjects and objects are trained separately, thus having:

$$2 * \#modalities * m * n \text{ GMMs.}$$

- At **testing time**, the process is repeated so as to obtain for each test image the **subject and object probability maps**.
- The **probability maps are fused** using the **Graph Cuts optimization** framework.



Qualitative Results



Biographies

Cristina Palmero received her Bachelor degree in Audiovisual Telecommunication Systems Engineering at Universitat Politècnica de Catalunya (UPC), Terrassa, Spain, in 2011. She is currently studying her Master degree in Artificial Intelligence at Universitat Politècnica de Catalunya (UPC) and Master degree in Computer Vision at Universitat Autònoma de Barcelona (UAB). She is mainly interested in signal and digital image processing and computer vision techniques applied to human behavior analysis, scene understanding and robotics.

Albert Clapés received his B.S. degree in Computer Science at Universitat de Barcelona in 2012. He is currently studying the interuniversitary M.S. degree in Artificial Intelligence at Universitat Politècnica de Catalunya. He is a research fellow at Department of Applied Mathematics and Analysis in Universitat de Barcelona and an eventual member of the Computer Vision Center (Universitat Autònoma de Barcelona). His main interests are computer vision and machine learning applied to human pose recovery and behaviour analysis, and also the human-machine natural interaction technologies.

References

[1] Andreas Møgelmose, Albert Clapés, Chris Bahnsen, Thomas B. Moeslund and Sergio Escalera, "Tri-modal Person Re- identification with RGB, Depth and Thermal features", *9th IEEE Workshop on Perception Beyond the Visible Spectrum*, 2013.
[2] Dalal Navneet and Bill Triggs, "Histogram of oriented gradients for human detection", CVPR 2005, IEEE Computer Society Conference on,. Vol. 1, p. 886-893, 2005.
[3] Yi Yang and Deva Ramanan, "Articulated pose estimation with flexible mixture-of-parts", CVPR 2011, IEEE Conference on., p. 1385-1392, 2011.
[4] Yuri Boykov and M-P. Jolly, "Interactive graph cuts for optimal boundary region segmentation of objects in ND images", ICCV 2001. Proceedings. 8th IEEE International Conference on., Vol. 1 p. 105-112, 2001.