Tri-modal Human Body Segmentation

Cristina Palmero[1], Albert Clapés[1,2], Chris Bahnsen[3], Andreas Møgelmose[3]
Advisors: Sergio Escalera[1,2], Tomas B. Moeslund[3]
E-mail: c.palmero.cantarino@gmail.com, aclapes@cvc.uab.cat, sergio@maia.ub.es, {am,cb,tbm}@create.aau.dk
[1] Dept. Matemàtica Aplicada i Anàlisi, UB, Gran Via de les Corts Catalanes 585, 08007, Barcelona
[2] Computer Vision Center, Campus UAB, Edifici O, 08193, Bellaterra, Barcelona
[3] Aalborg University, Sofiendalsvej 11, 9200 Aalborg SV, Denmark

Keywords: Body segmentation, Multi-modal features

1 Motivation

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, human-computer interaction, and so on and so forth. In the state-of-the-art, we mostly find the usage of color images recorded by cameras or, more recently, with the apparition of RGB-Depth devices – such as Microsoft® Kinect™ –, the usage of depth maps in combination with the 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 and, 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 (as it happens with other sensory devices).

The main contribution of this paper is a novel tri-modal database of people acting in three different scenes, consisting of more than 2,000 frames each one, in which three different subjects appear and interact with objects performing different actions such as reading, working on a laptop, speaking by phone, etc. In addition, a human segmentation baseline methodology is also proposed, consisting in segmenting first the people in each of the modalities separately and, finally, fusing the results in an optimization graph-cuts framework.

2 Method

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, the subject regions are described at pixel-level using particular descriptors in each of the modalities. Once the subject pixels have been described in all the modalities, Gaussian Mixture Models (GMMs) 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 combined in the Graph Cuts optimization step. A graphical representation of the proposed methodology is shown in Figure 1, which is explained in more detail below.

Background subtraction is performed separately in the different cues. In each modality, a model of the background is modeled given an initial set of F frames, learning a gaussian distribution for each pixel. Since alignment among dataset modalities is not accurate, foreground regions are fused and aligned at near pixel level by re-scaling of nearest detected regions. Thus, the segmented regions in the three different scenes can be merged to later describe them.

Then, the descriptors can be computed for all the pixels in all the modalities. Each modality involves its own specific descriptors: in the color cue, we have computed the histogram of oriented gradients (HOG) [2] in a $h \times w$ window centered at the pixel; the color cue also allows us to obtain motion information by computing dense optical flow and describing the distribution of the resultant vectors, known as histogram of oriented optical flow (HOOF); in the depth cue, histograms of oriented surface normals (HOSF) have been used; and, in the thermal cue, histograms of thermal intensities concatenated with histograms of oriented gradients (HI-HOG). This implies having 4 different descriptions for each pixel.

At this point, the distributions of the previously computed descriptors are modeled by GMMs. Instead of learning only one GMM for the pixels in each kind of description, we divide the subject regions in a grid of cells and learn a GMM in each cell, thus having $4 \times \#cells$ GMMs.

Eventually, in the testing step, the new extracted regions are also divided in cells, their pixels described and the probabilities predicted from the corresponding GMM. Finally, these obtained probabilities are combined together with the extracted human body probabilities from Ramanan et. al. method [3], weighting each one, and used as the data-term in Graph Cuts optimization algorithm [4], so as to obtain a final segmentation of the human body.

3 Results

The results obtained so far are mainly qualitative but allow us to develop approaches to the problems that we face before starting to retrieve quantitative conclusions. Descriptive examples of the different stages are represented in Figure 2 and 3.

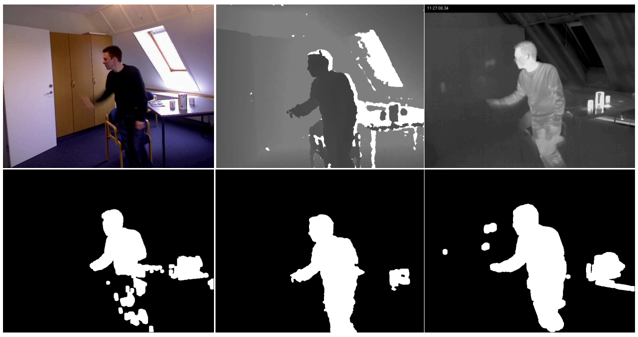


Figure 2: Background subtraction for different visual modalities of the same scene.

References

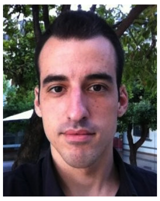
[1] Andreas Møgelmose, Albert Clapés, Chris Bahnsen, Thomas B. Moeslund and Sergio Escalera. Tri-modal Person Re-identification with RGB, Depth and Thermal Features. *9th IEEE Workshop on Perception Beyond the Visible Spectrum*, 2013

[2] Dalal Navneet and Bill Triggs. Histogram of oriented gradients for human detection. In *CVPR 2005. IEEE Computer Society Conference on.*, Vol. 1, p. 886-893, 2005.

[3] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-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, *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on.*, Vol. 1, p. 105-112, 2001.

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 interuniversity 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 in research are computer vision and machine learning applied to human pose recovery and behavior analysis, and also the human-machine natural interaction technologies.

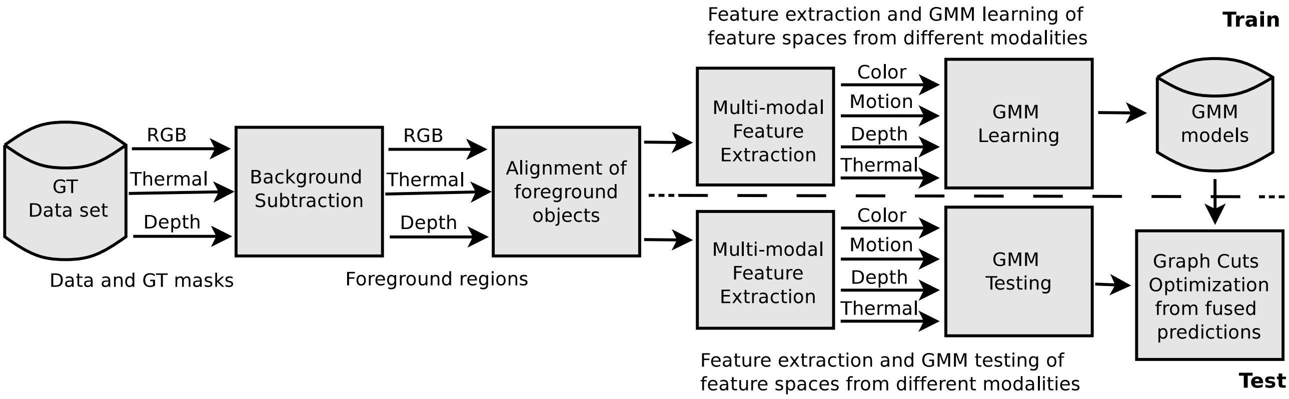
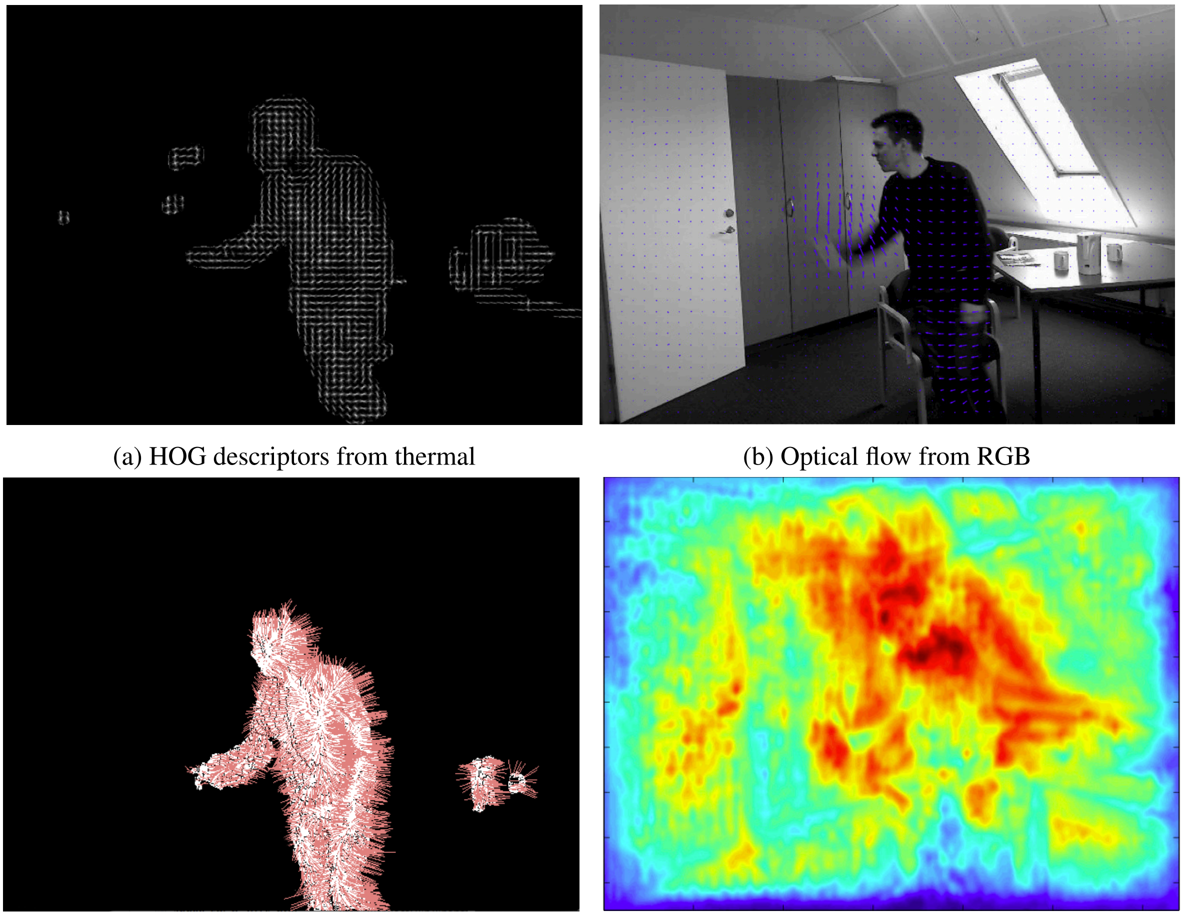

Feature extraction and GMM learning of feature spaces from different modalities
Train
GT Data set
RGB
Thermal
Depth
Data and GT masks
Background Subtraction
RGB
Thermal
Depth
Foreground regions
Alignment of foreground objects
Multi-modal Feature Extraction
Color
Motion
Depth
Thermal
GMM Learning
GMM models
Multi-modal Feature Extraction
Color
Motion
Depth
Thermal
GMM Testing
Graph Cuts Optimization from fused predictions
Feature extraction and GMM testing of feature spaces from different modalities
Test

Figure 1: Pipeline of the presented methodology.



(a) HOG descriptors from thermal

(b) Optical flow from RGB

(c) Normals from depth

(d) Ramanan probability map

Figure 3: Example of descriptors from different modalities.