Master in Artificial Intelligence (UPC-URV-UB)

Master of Science Thesis

Tri-modal Human Body Segmentation

Cristina Palmero Cantariño

Advisor/s: Sergio Escalera Guerrero

January 2014

Abstract

Human body segmentation in visual data is a challenging problem in Computer Vision. This problem has been treated for decades, and it still remains an open issue. The main difficulties of human body segmentation in visual data come from the different points of view, changes in clothes and illumination, background artifacts, presence of occlusions and the articulated nature of the human body. An accurate segmentation of the human body will benefit the development of a new generation of potential applications in health, leisure, and security.

Recently, with the presence in the marked of new cheap sensors that provide complementary visual information to classical RGB, the generation of new multi-modal descriptors and fusion strategies have received special attention by the community. Examples of these sensors include Thermal imaging and depth data coming from infrared sensors, which can provide complementary and discriminative information to RGB descriptors in order to improve segmentation of humans in multi-modal visual data. In this master thesis project it has been designed a novel registered multi-modal RGB-Depth-Thermal data set of continuous image sequences. The data set has been collected and registered in collaboration with an expert team from the Aarlborg University. This data set has been manually annotated at pixel level at the regions containing subjects in all three modalities and will be available for the scientific community.

From the novel data set, a multi-modal adaptive background subtraction approach has been proposed in order to automatically detect the regions of interest that can contain a subject in the image. In addition, several descriptors from the state-of-the art have been tested and adapted to extract information from the different modalities. The different feature spaces are modeled via Gaussian Mixture Models for both subject and non-subject categories, providing a confidence score for each grid region of interest. The feature modalities have been tested independently to evaluate their performance for subject segmentation, and two approaches for multi-modal segmentation have been proposed. The first naïve approach just computes a threshold value about the combined confidences for all modalities given the score provided by the Gaussian classifiers. In the second approach, a SVM is trained combining the output of previous classifiers confidences as features, also extending the feature vector with previous classifier predictions, in a stacked learning fashion. The results show variable performance for the different modalities when segmenting people in multi-modal data, and improved segmentation accuracy of the multi-modal GMM-SVM stacked learning method.

Resumen

La segmentación de personas en datos visuales es uno de los problemas actuales más difíciles en el área de la Visión por Computador. Este problema se ha estudiado durante décadas por la comunidad, y aún en día sigue siendo tratado. Las principales dificultades de este problema vienen dadas por los diferentes puntos de vista, los cambios en ropa e iluminación, los artefactos presentes en el fondo así como por el alto nivel articulado del cuerpo humano, el cual presenta multitud de cambios en apariencia de la pose. Obtener una segmentación robusta del cuerpo humano beneficiaría en el desarrollo de una nueva generación de aplicaciones con alta impacto social en los campos de la salud, ocio y la seguridad.

Recientemente, con la aparición en el mercado de nuevos sensores económicos que proporcionan información visual complementaria al clásico RGB, se ha generado un nuevo interés por el estudio de descriptores multi-modales así como de nuevas técnicas de fusión y aprendizaje de datos. Ejemplos de estos sensores incluyen imágenes termales y mapas de profundidad por infrarrojos, los cuales pueden proporcionar información complementaria y altamente discriminativa a los descriptores estándar RGB, y como consecuencia, conseguir mejorar el rendimiento integrado de los modelos de segmentación de personas en datos visuales multi-modales.

En este proyecto de tesis de máster se ha desarrollado una nueva base de datos multi-modal RGB-Termal-Profundidad registrada de secuencia continua de datos visuales. La base de datos ha sido filmada y registrada en colaboración con un equipo experto del área de la Universidad de Aarlborg en Dinamarca. Los datos han sido manualmente anotados a nivel de píxel en las regiones donde aparecen sujetos realizando diferentes actividades e interaccionando con objetos presentes en la escena. Para analizar la base de datos se ha desarrollado un método de extracción de fondo adaptativo multi-modal que extrae las regiones de interés que potencialmente pueden contener un sujeto. Adicionalmente, un conjunto de descriptores del estado del arte han sido testeados y adaptados para extraer información de las diferentes modalidades. Los diferentes vectores de características han sido modelados mediante mixturas de Gausianas, proporcionando una métrica de confidencia de usuario u objeto dentro de una rejilla de las regiones detectadas. Finalmente se han propuesto dos metodologías de segmentación multi-modal. La primera y más “naïve” estima un sesgo de corte sobre las confidencias combinadas de todas las modalidades dada la puntuación obtenida por los clasificadores Gausianos. En el segundo método, los SVM son entrenados combinando las confidencias de salida de los primeros clasificadores junto con la predicción de clasificación de los mismos como nuevas características, siguiendo la filosofía de los métodos enlazados de clasificadores. Los resultados obtenidos muestran variabilidad en el rendimiento de cada modalidad para segmentar personas, y una mejora significativa de los métodos de fusión multi-modales, y en especial del método enlazado basado en GMM-SVM.

Resum

La segmentació de persones en dades visuals és un dels problemes actuals més difícils de l'àrea de la Visió por Computador. Aquest problema s'ha estudiat durant dècades per la comunitat, i encara avui dia segueix sent un cas d'estudi. Les principals dificultats d'aquest problema venen donades pels diferents punts de vista, els canvis en roba i il·luminació, els artefactes presents al fons de les imatges, així com per l'alt nivell articulat del propi cos humà. Obtenir una segmentació acurada de cos humà beneficiaria el desenvolupament d'una nova generació d'aplicacions amb alt impacte social en els camps de la salut, l'oci i la seguretat.

Recentment, amb l'aparició al mercat de nous sensors econòmics que proporcionen informació visual complementària a la clàssica RGB, s'ha generat un nou interès per l'estudi de descriptors multi-modals, així com en noves tècniques de fusió i aprenentatge de dades. Exemples d'aquests sensors inclouen imatges termals i mapes de profunditat per infrarojos, els quals poden proporcionar informació complementària i altament discriminativa als descriptors clàssics RGB, y com a conseqüència, aconseguir un millor rendiment integrat dels models de segmentació de persones a dades visuals multi-modals.

En aquest projecte de tesis de màster s'ha desenvolupat una nova base de dades muti-modal RGB-Termal-Profunditat registrada de seqüència contínua de dades visuals. La base de dades ha estat filmada i registrada en col·laboració amb un equip expert de l'àrea de la Universitat Aarolborg a Dinamarca. Les dades han estat manualment etiquetades a nivell de píxel a les regions on apareixen subjectes realitzant diferent activitats i interaccionant amb objectes presents a l'escena. Per analitzar la base de dades s'ha desenvolupat un nou mètode d'extracció de fons adaptatiu muti-modal que detecta les regions d'interès que potencialment poden contenir un subjecte. Addicionalment, un conjunt de descriptors de l'estat de l'art han estat testejats i adaptats per extreure informació de les diferents modalitats. Els diferents vectors de característiques han estat modelats mitjançant mixtures de Gaussianes, proporcionant una mesura de confidència de usuari i/o objecte dins d'una graella sobre les regions detectades. Finalment s'han proposat dos metodologies de segmentació multi-modal. La primera i més "naïve", estima una cota de separació de confidències combinades de totes les modalitats donada la puntuació dels classificadors Gaussians. En el segon mètode proposat, els SVM són entrenats combinant les confidències de sortida dels primers classificadores junt amb la predicció de classificació dels mateixos com a noves característiques, tot seguint la filosofia dels mètodes d'enllaçament de classificadors. Els resultats obtinguts mostren variabilitat en el rendiment de segmentació de cada modalitat, i una millora significativa dels mètodes proposats de fusió i segmentació multi-modal, i en especial del mètode per enllaçat basat en GMM-SVM.

Acknowledgements

I would like to thank my supervisor, Sergio Escalera, for his exemplary guidance, generous support, constructive comments and encouragement throughout the elaboration of this work. I would also like to thank Albert Clapés, who has been actively collaborating in this project. Without their persistent help this dissertation would not have been possible.

A special thanks to all the classmates I have met during this year and a half. This Master would not have been the same without them. Finally, I thank all my family for their constant and unconditional support.

Contents

Chapter 1

Introduction

Segmentation of people in images is still nowadays a very challenging and tough problem for the computer vision community due to the great diversity of poses that they can adopt and their variable appearance. Difficulties also arise from changes in lighting conditions and complex and cluttered backgrounds. The general idea of human body segmentation is to assign a label to every pixel or group of pixels in an image such that pixels with the same label share certain visual characteristics which entitles them to be considered as part of a human. These type of problems are commonly referred to as labeling problems. Despite extensive research done so far, some constraints have still to be taken into account and one often has to make assumptions about the scenario where the segmentation task is to be applied, such as static versus moving camera, indoor versus outdoor, and so on. Ideally, it should be tackled in an automatic fashion rather than relying on user intervention, which makes such task even more challenging.

There exist a wide range of possible applications for people segmentation such as surveillance, content-based image retrieval, activity recognition, patient caregiving or human-computer interaction among others. Such task is also often related to pose estimation problems, since it can be carried out efficiently once the person is detected and segmented in an image. Furthermore, it can facilitate the task of photo edition, chroma-keying or video compression. Hence, human body segmentation can be considered as an important preprocessing step for other tasks.

State of the art methods that tackle the human segmentation task mostly use color images recorded by RGB cameras as the main cue for further analysis, although they present several widely known intrinsic problems such as intensity similarities between background and foreground. More recently, the release of RGB-Depth devices such as Microsoft® Kinect^{TM}, a cheap multi-sensor device based on structured light technology, has allowed the community to use RGB images along with per-pixel depth information, often called depth maps, thus increasing the robustness of the methods. Besides, this device has helped boost research in human pose and behavior analysis.

1.1 Proposal

In this context, we propose adding a third modality: thermal imagery got from thermal infrared cameras, thus complementing other information sources and making easier the segmentation task. Although thermal cameras are relatively expensive devices, their market price is lowering substantially every year –as it happens with other sensory devices. Besides, they can capture data similar to standard color cameras but without having illumination problems, that is why infrared cameras are becoming popular in surveillance scenarios and other similar applications. To do so, we introduce a novel tri-modal database provided by researchers from Aalborg University in Denmark and Universitat de Barcelona. Such database contains people acting in three different video sequences, consisting of more than 2,000 frames each one, in which three different subjects appear and interact with objects performing diverse actions such as reading, working with a laptop, speaking on the phone, etc. There may be more than one subject appearing in scene. The dataset comes along with an algorithm that performs the registration among modalities.

In addition, we present a baseline methodology to automatically segment people in video sequences in indoor scenarios with a fixed camera. With all the available modalities, important features will be extracted using different state of the art descriptors, which are used to learn a probabilistic model so as to predict the image regions belonging to people. We will compare results from applying segmentation to the different modalities separately to results obtained by fusing features from all modalities.

To the best of our knowledge, this is the first dataset and work that combines color, depth and thermal modalities to perform the people segmentation task in videos, aiming to bring further benefits towards developing more robust solutions.

1.2 Outline

The remainder of this dissertation is organized as follows. Section 2 reviews the different approaches for human body segmentation that appear in the recent literature. Section 3 introduces and exhaustively explains the proposed baseline methodology, which will be experimentally evaluated in Section 4. Finally, Section 5 concludes this dissertation.

Chapter 2

Related work

Image segmentation is one of the oldest and most widely studied problems in computer vision [1–5]. First approaches had a tendency to use region splitting or merging, which correspond to divisive and agglomerative clustering respectively. Later, research focused on methods that try to optimize some criteria, such as inter-region boundary lengths, intra-region consistency or dissimilarity [6]. Due to the vast work available on image segmentation, in this section we are going to focus in the most recent and relevant works, techniques and methods applied specifically to human body segmentation that determine the state of the art.

2.1 Methods

Within the last decade a great number of novel approaches have emerged to respond to different requirements in the human segmentation context, such as trying to overcome changing illumination conditions, dealing with variable human poses or developing quasi-automatic systems that progressively lose the need for user intervention.

When dealing with indoor scenarios recorded by a stationary camera, the pixel-based background subtraction approach can be applied successfully. We can model the background distribution of the scene and detect moving objects by comparing each pixel to the model, which are considered as foreground. The result is a silhouette of the moving object, which can be further used for other tasks. Pixel intensity is the most commonly used feature in background modeling, though there are many approaches that use other type of information such as edge, motion, stereo or texture features. The parametric model that Stauffer and Grimson proposed in [7], which models the background using a mixture of gaussians (MoG), has been widely used and many variations have been suggested based on it. In [8], more advanced statistical background modeling techniques are deeply reviewed.

Nonetheless, after obtaining the moving object contours we still need a way to assess whether they belong to a human or not. Human detection methods are strongly related to the task of human body segmentation since they allow to discriminate better between other objects. They usually produce a bounding box indicating where the person is, which in turn may be also useful as a prior for pixel-based or bottom-up approaches to refine the final human body silhouette. In the category of holistic body detectors, one of the most successful representations is the Histogram of Oriented Gradients (HOG) [9], still being the basis of many current detectors. Used along with a discriminative classifier –e.g. Support Vector Machines (SVM) [10] –it is able to accurately predict the presence of human subjects. Example-based methods [11] have been also proposed to address human detection, utilizing templates to compare the incoming image and locate the person, limiting the pose variability though.

Regarding descriptors, other possible representations apart from the already commented HOG are those that try to fit the human body into silhouettes [12], those that model color or texture such as Haar-like wavelets [13], optical flow quantized in Histrograms of Optical Flow (HOOF) [14], depth maps [15] and, more recently, descriptors including logical relations, e.g. Grouplets [16], which enables to recognize human-object interactions.

Instead of whole body detection, some approaches have been built under the assumption that the human body consists of an ensemble of body parts [17, 18]. Some of them are based on pictorial structures [19, 20]. In particular, [20, 21] from Yang and Ramanan along with [22] from Felzenszwalb have outperformed other existing methods using a Deformable Part-based Model (DPM) that consists on a root HOG-like filter and different part-filters that define a score map of an object hyphotesis, using latent SVM as a classifier. Another well-known part-based detector is Poselets [23, 24], which trains different homonymous parts to fire at a given part of the object at a given pose and viewpoint. Grammar models [25] and AND-OR graphs [26] have been also used in this context.

By the same token, other approaches model objects as an ensemble of local features. In this category there are included methods such as Implicit Shape Models (ISM) [27], consisting of visual words combined with location information. In addition, they are used in works that estimate the class-specific segmentation based on the detection result after a training stage [28].

Conversely, generative classifiers directly deal with the person segmentation problem. They function in a bottom-up manner, learning a model from a initial prior in the form of bounding boxes or seeds, and using it to yield an estimate for the background and target distributions, normally applying Expectation Maximization (EM) [29, 30]. One of the most popular is GrabCut [31], an interactive segmentation method based on graph cuts [32] and Conditional Random Fields (CRF) that, using a bounding box as an initialization region, combines pixel appearance information with neighborhood relations to refine silhouettes up to a very accurate level. Graph cuts method has been further applied to part-based approaches [33].

Having considered the properties of each one of the aforementioned segmentation categories, it is reasonable that several approaches have been proposed towards their combination, that is, top-down and bottom-up segmentation [34–37]. Just to name a few, ObjCut [38] combines pictorial structures and Markov Random Fields (MRF) to obtain the final segmentation. PoseCut [39] is also based on MRF and graph cuts but it has the added ability to deal with 3-D pose estimation from multiple views.

According to the nature of our proposal, we find appropriate to dedicate a few lines regarding thermal imagery and associated descriptors. In contrast to color or depth cues, thermal infrared imagery has not been used that widely for human detection and segmentation purposes, although it is experiencing a growing interest by the research community. Several specific descriptors have been proposed so far. For example, in [40], the authors extended the combination of edgelets and HOG features with AdaBoost and SVM cascade to infrared images. Background subtraction has been also adapted to deal with this kind of imagery in [41], which presented a statistical contour-based technique that eliminates typical halo artifacts produced by infrared sensors by combine foreground and background gradient information into a contour saliency map so as to find the strongest salient contours. More recently, [42] presented a person re-identification method that for the first time combined RGB, depth, and thermal features. An extensive survey of thermal cameras and their applications can be found in [43], including technological aspects and the nature of thermal radiation.

2.2 Benchmark datasets

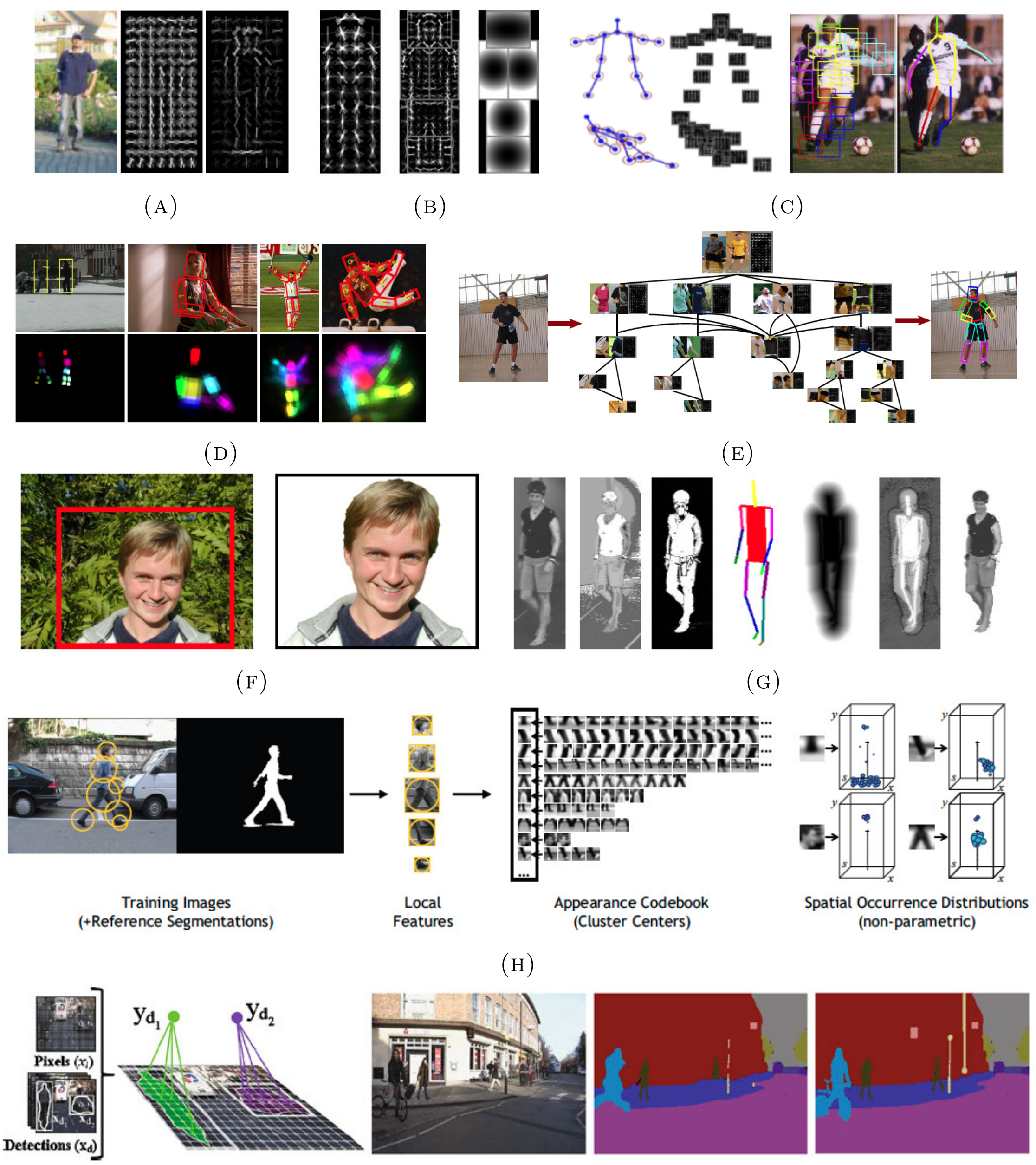
To advance research in this area, it is a must to have the right means to compare existent methods so as to allow improvements to be measured. It is not easy to find large image segmentation databases due to the tedious task that manual labeling implies. In this context, the appearance of crowdsourcing-based frameworks such as Amazon Mechanical Turk [44] or LabelMe [45] has encouraged users to participate in with little easy tasks such as image segmentation or annotation, thus gamificating somehow the laborious task of human-labeling and helping the computer vision community to obtain ground truth information at a lower cost.

Either way, there exist several static image-based human-labeled databases, which allow us to compare the great deal of available literature of image segmentation. The best known of these is the Berkeley Segmentation Dataset and Benchmark[1] [46], which consists of 12,000 segmentations of 1,000 Corel dataset color images, containing people or different objects. It also includes figure-ground labelings for a subset of the images. Authors of [47] made also available a database[2] containing 200 gray level images along with ground truth segmentations, which was specially designed to avoid potential ambiguities by only incorporating images that clearly depict one or two objects in the foreground that differ from its surroundings by either texture, intensity, or other low level cues, but it does not represent uncontrolled scenarios. The well known PASCAL Visual Object Classes Challenge [48] tended to include a subset of the color images annotated in a pixel-wise fashion for the segmentation competition. Although not considered to be benchmarks, Kinect-based datasets are also available, since this device is being widely used in human pose related works. In [49] a novel dataset[3] was presented, which contains 3,386 images of

[1] http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/

[2] http://www.wisdom.weizmann.ac.il/%7Evision/Seg_Evaluation_DB/index.html

[3] http://www.robots.ox.ac.uk/~vgg/data/humanSeg/



(I)

FIGURE 2.1: Examples of recent methods and descriptors: (A) HOG [9]: Person and his computed HOG descriptor, and the descriptor weighted by positive SVM weights; (B) Deformable Part-based Model [22]: coarse root filter, spatial model for the location of each part and cost of placing the center of a part at different locations relative to the root, respectively; (C) Mixture of parts from [21]: different trees obtained from the mixture of parts and estimation of parts and pose; (D) Pictorial structures [19]: pedestrian detection and upper-body and full body pose estimations, respectively; (E) Poselets [24]: each part shows its inferred poselet and the SVM HOG template; (F) GrabCut [31]: rectangle defined by the user that acts as a bounding box prior and object extracted based on it; (G) PoseCut [39] in order of appearance: original image, pixel likelihood for being labeled as foreground or background, segmentation after using the GMM models, optimal estimated pose, shape prior corresponding to the optimal pose, likelihood after fusing the previous information, and final segmentation; (H) ISM for segmentation [28]: training procedure, where local features are extracted from interest points and clustered to create an appearance codebook, which allows to learn a spatial occurrence distribution for each entry; (I) bottom-up top-down segmentation [36]: CRF structure, original image, and results before and after applying GrabCut to each detected bounding box, respectively.

segmented humans and ground truth automatically created by Kinect, and consisting of different human subjects across 4 different locations. Unfortunately, depth map images are not included in the public dataset.

However, there is a lack of a standard database of videos that can be used for evaluation purposes, such as visual surveillance approaches. There exist some popular ones which try to provide realistic settings and environmental conditions [50]. Among all of them, we underline the collective datasets of Project ETISEO[4] [51], owing to the fact that for some of the scenes the authors include, apart from color images, an additional imaging modality such as infrared footage. It consists of indoor and outdoor scenes of public places such as an airport apron or a subway station, and also includes a frame-based annotated ground truth. Depth modality is used in some works such as the RGB-D People Dataset[5] [52], which presented a dataset containing more than 3,000 RGB-Depth frames using Kinect. The sequences show mostly upright walking and standing persons from a range of orientations and different levels of occlusions, although the annotation is done in a bounding-box fashion, that is, only detecting people.

[4] http://www-sop.inria.fr/orion/ETISEO/download.htm

[5] http://www.informatik.uni-freiburg.de/~spinello/RGBD-dataset.html

Chapter 3

Proposed baseline

Let us write $\mathbf{F}_i = \{\mathbf{C}_i, \mathbf{D}_i, \mathbf{T}_i\}$ for a determined tri-modal frame, and $\mathbf{p}$ a pixel at an arbitrary location (x, y) in an image.

3.1 Extraction of masks and regions of interest

3.1.1 Background subtraction

The first step of our baseline is to attempt to reduce the search space. A static videocamera with fixed orientation observing an indoor scene is a common practice which enables to detect and isolate new objects entering the scene assuming that the images of the scene without new objects, known as background, exhibit some regular behavior that can be described by a statistical model. Thus, in order to perform background subtraction one has first to learn a model of the background. Once learned, this model is compared against the new incoming images and parts that do not fit are considered foreground. A widely used approach for background modeling in this context is Mixture of Gaussians MOG [53], which assigns a GMM per pixel with a fixed number of components. Sometimes background presents periodically moving parts such as noise or sudden and gradual illumination changes. Such problems are often tackled with adaptive algorithms that keep learning the pixel's intensity distribution after the learning stage with a decreased learning rate. However, this also causes that intruding objects that stand still for a period of time vanish, so in our case a non-adaptive approach is more convenient.

Although this background subtraction technique performs fairly well, it has to deal with the intrinsic problems of the different image modalities. For instance, color-based algorithms may fail due to shadows, similarities in color between foreground and background, highlighted regions, and sudden lighting changes. Thermal imagery may also have this kind of problems, plus the inconvenience of temperature changes in objects. A halo effect is also observed around warm items. Regarding to depth-based approaches, they may produce misdetections due to the presence of foreground objects with similar depth to the background. However, they are more robust to lighting artifacts and shadows. Depth data is quite noisy and many pixels in the image may have no depth due to multiple reflections, transparent objects, or scattering in certain surfaces such as human tissue and hair. Furthermore, a halo effect around humans or objects is usually perceived due to parallax issues. A comparison is shown in Fig. 3.1, where the actual foreground objects are the human and the objects on the table. As we can observe, RGB fails at extracting the human legs due to the similarities in color with the chair at the back. The thermal cue segments the human body more accurately, but includes some undesired reflections and the jar and mugs with a surrounding halo. The pipe tube is also extracted as foreground due to temperature changes along time. Despite its drawbacks, depth-based background subtraction is the one that seems to give the most accurate result.

Therefore, the binary foreground masks of our proposed baseline are computed applying background subtraction to the depth modality previously registered to the RGB one, thus allowing us to use the same masks for both modalities. Let us consider the depth value of a pixel at frame i as $\mathbf{d}_i$. The background model $p(\mathbf{d}_i|BG)$ –where BG represents the background – is estimated from a training set of depth images represented by $\mathcal{D}$ using the T first frames of a sequence such that $\mathcal{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_T\}$. This way, the estimated model is denoted by $\hat{p}(\mathbf{d}_i|\mathcal{D}, BG)\}$, modeled as a mixture of gaussians. In particular, we use the available implementation in OpenCV of the method presented in [54], which uses an on-line clustering algorithm that constantly adapts the number of components of the mixture for each pixel during the learning stage. GMMs are further explained in section 3.4.1.

Once the binary foreground masks are obtained, a 2-D connected component analysis is performed using basic mathematical morphological operators and setting a minimum value for each connected component area (except in left and rightmost sides of the image which may be caused by a new incoming item) to clean the noisy output mask. On another front, foreground masks for the thermal modality are computed using the provided registration algorithm with the depth/color foreground masks as input. From now on, we will use $FG = \{FG^{\text{color}}, FG^{\text{depth}}, FG^{\text{thermal}}\}$ to refer to them.

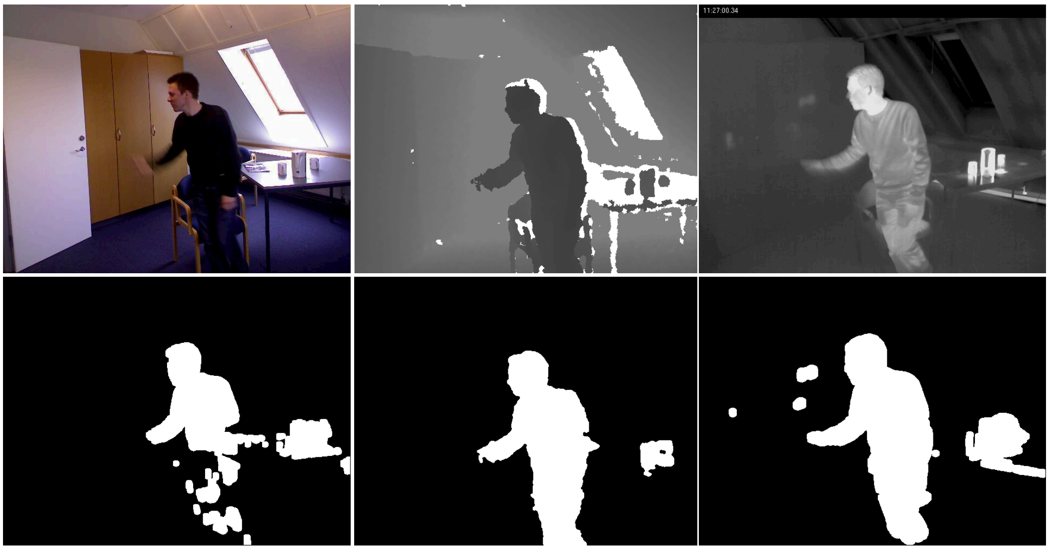


Figure 3.1: Background subtraction for different visual modalities of the same scene (RGB, depth and thermal respectively).

3.1.2 Bounding box generation from regions of interest

To further process the information of each connected component of the previously extracted depth-based foreground masks, rectangular bounding boxes are to be generated encapsulating such components individually over time, whose function is to denote the regions of interest of a foreground mask. This way, we define the set of bounding boxes of the i-th frame generated from the depth-based masks as $B_i^{\text{depth}} = \{b_{ij} \mid \forall j = \{1, \ldots, n\}\}$, being b_{ij} the j-th bounding box and n the number of bounding boxes generated in that frame, which is equal to the number of connected components. Similarly, bounding boxes generated from the resulting thermal masks are denoted by B_i^{thermal}. Since color and depth modalities share the same foreground masks, $B_i^{\text{depth}} = B_i^{\text{color}}$, having also the same number of bounding boxes, condition that B_i^{thermal} does not currently fulfill. At the end, each frame will have to contain the same number of bounding boxes in each one of the modalities, which in turn have to correspond to the same regions of interest among them in order to allow a proper comparison. This issue will be tackled in section 3.1.3.

A region of interest $r \in R$ should contain a separated person or object. However, different subjects or objects may overlap in space, resulting in a bigger component containing more than one item, for this reason each component has to be analyzed to find the correct bounding boxes that surround each region of interest. One of the advantages of the depth cue is that we can use the depth value in each range pixel to know whether an item is farther or not than other. We can assume that a given connected component belongs to just one item if its disparity distribution has a low standard deviation, that is, there is no rapid change in disparity. For those that have a greater standard deviation, Otsu's method [55] can be used to split the blob by automatically finding a threshold assuming a bimodal distribution. It calculates the optimal threshold separating the two classes such that their intra-class variance is minimal. We will define π as the function that applies this method to a set of bounding boxes.

For such purpose, we define $\mathbf{c}$ as a vector containing the depth range values that correspond to a given connected component, with mean μ_c and standard deviation σ_c, and σ_{otsu} as a parameter that defines the maximum σ_c allowed to not apply π. Note that erroneous white or black pixels do not have to be taken into account in $\mathbf{c}$ when finding the Otsu's threshold because they would change the disparity distribution, thus leading to incorrect divisions. Hence, if $\sigma_c > \sigma_{\text{otsu}}$, π is applied. However, the assumption of bimodal distribution may not hold, so to take into account the possibility of more than two overlapping items the process is applied recursively to the divided regions in order to extract all of them.

Once the different items are found, the regions belonging to them are labeled using a different id per item. Besides, they are again surrounded by new bounding boxes denoted by:

$$O_i^{\text{depth}} = \{o_{im} | \forall m = \{1, \ldots, M_n\}_j\} \tag{3.1}$$

where $\{o_{im}\}_j$ is the set of new M_n bounding boxes generated by the bounding box b_{ij}.

3.1.3 Bounding box transformation and correspondence to other modalities

As stated previously, depth and color cues use the same foreground masks, so we can take advantage of the same bounding boxes for both modalities. However, since the thermal cue uses a transformation of these masks by applying the registration algorithm frame by frame, new bounding boxes with different coordinates are computed for that modality, which must correspond to the same regions of interest of the depth and color cues. In case of overlapping items, it would suffice by finding the registered labeled connected components to generate the new bounding boxes. Unfortunately, the algorithm cannot register connected components up to a pixel level, meaning that those that have more than one id in depth or color masks would have just one id in thermal ones, thus being surrounded by just one big bounding box. The problem here is twofold: (1) find the correspondence between B_i^{depth} and B_i^{thermal} such that a bounding box of B_i^{depth} and the matched one of B_i^{thermal} contain the same region of interest r; and (2), if a bounding box from B_i^{depth} was modified after π, compute the corresponding sub bounding boxes that belong to the matching bounding box in B_i^{thermal}.

The correspondence task (1) is achieved using an iterative algorithm that, taking into account the deviation among depth/color and thermal modalities, searches for the bounding boxes that match the best, both in terms of bounding box coordinates and area similarity. The correspondence function is denoted as:

$$b_{iq}^{\text{thermal}} = \beta(b_{ij}^{\text{depth}}) \tag{3.2}$$

Bounding boxes of both sets are ordered beforehand in a row-major order using the top-left corner of the bounding box as a reference, thereby easing the search task. Bounding boxes which appear in thermal but do not have a correspondence in depth are omitted, whereas those in the depth modality that do not have a correspondence either are copied as they are, that is, with the same coordinates, following one of the main aforesaid constraints which states that the number of bounding boxes must be the same among modalities.

Once we have found the correspondence between both sets, we can use the information stored in O_i^{depth} in order to address the task of bounding box splitting (2). Let us define $h_{b_{iq}^{\text{thermal}}}$ and $v_{b_{iq}^{\text{thermal}}}$ as the height and width of a thermal bounding box respectively, and similarly for depth bounding boxes $h_{b_{ij}^{\text{depth}}}$ and $v_{b_{ij}^{\text{depth}}}$, such that $b_{iq}^{\text{thermal}} = \beta(b_{ij}^{\text{depth}})$. Given the deviation among modalities, we assume that the dimensions of two matched bounding boxes are proportional. Therefore, $h_{b_{iq}^{\text{thermal}}} \propto h_{b_{ij}^{\text{depth}}}$ and $v_{b_{iq}^{\text{thermal}}} \propto v_{b_{ij}^{\text{depth}}}$. Thus, the ratio between both bounding boxes is:

$$k_{\text{h}} = \frac{h_{b_{ij}^{\text{depth}}}}{h_{b_{iq}^{\text{thermal}}}} \tag{3.3}$$

$$k_{\text{v}} = \frac{v_{b_{ij}^{\text{depth}}}}{v_{b_{iq}^{\text{thermal}}}} \tag{3.4}$$

Such ratio can be utilize to create a new bounding box $o_{ik}^{\text{thermal}} \in O_i^{\text{thermal}}$ in such a vay that its dimensions are:

$$h_{o_{ik}^{\text{thermal}}} = k_{\text{h}} h_{o_{ij}^{\text{depth}}} \tag{3.5}$$

$$v_{o_{ik}^{\text{thermal}}} = k_{\text{v}} v_{o_{ij}^{\text{depth}}} \tag{3.6}$$

The expansion or shrinking of o_{ik}^{thermal} is produced taking as reference the center of o_{ij}^{depth}, considering the boundaries of b_{iq}^{thermal} as the growth limit, meaning that if the new bounding box has to expand vertically but it reaches the bottom of b_{iq}^{thermal}, it will stop expanding downwards but will continue doing so upwards, and similarly for left and right sides, until reaching the desired measures if possible.

As a result, we obtain the final set of bounding boxes $O^{\text{thermal}} \equiv O^{\text{depth}} = O^{\text{color}}$, which although not having the same coordinates denote the same regions of interest R. Henceforth, we will simply use R to refer to such regions.

3.2 Grid partitioning

Given the precision got in the registration, particularly because of the depth-to-thermal transformation, we are not able to make a pixel-to-pixel correspondence among all the modalities. Instead, the association is made among greater information units: grid cells.

Each region of interest $r \in R$ associated to either a segmented subject or object is partitioned in a grid of cells. We write G_{rij} to refer to the position (i, j) in the r-th region, such that $i \in \{1, ..., v_{\text{grid}}\}$ and $j \in \{1, ..., h_{\text{grid}}\}$. Regarding to the whole set of (i, j)-cells $\{G_{rij}\}_{\forall r \in R}$, it is denoted by G_{ij}.

In turn, a grid cell G_{rij} consists in a set of image subregions $\{\mathbf{G}^c_{rij}\}_{\forall c \in \mathcal{C}}$, provided by the set of visual cues $\mathcal{C} = \{\text{color}, \text{depth}, \text{thermal}\}$. Accordingly, $\{\mathbf{G}^c_{rij}\}_{\forall r \in R}$, the set of (i, j)-cells in the modality c, is indicated by G^c_{ij}.

The grid cell is the unit of information processed in the different modalities' description and classification procedures. The next section provides the details about the feature extraction computed from different visual cues at cell level.

3.3 Feature extraction

Each modality involves its own specific descriptors. In the case of the color modality two kind of descriptors are extracted for each cell, Histogram of Oriented Gradients (HOG) and Histogram of Oriented Optical Flows (HOOF), whereas in the depth and thermal modality the Histogram of Oriented Normals (HON) and Histogram of Intensities and Oriented Gradients (HIOG) are respectively computed. Eventually, from this feature extraction process, a set of descriptions is obtained $D_{ij} = \{\mathbf{D}^d_{ij}\}_{\forall d \in \mathcal{D}}$ for each grid position (i, j), being $\mathcal{D}$ the set of considered descriptors $\{\text{HOG}, \text{HOOF}, \text{HON}, \text{HIOG}\}$.

The color modality is also used to compute a sequence of probability-like maps at pixel-level (SM). Such descriptor is also included in this section but for the moment is not included in the set of descriptions $\mathcal{D}$, owing to the fact that its intrinsic characteristics differ from the others and will be treated in a distinct fashion.

3.3.1 Color

The color cue is nowadays the most popular imagery modality and has been extensively used to extract a range of different feature descriptions. It is usually represented by the RGB color space, which expresses the color as a triplet $(\text{red}, \text{green}, \text{blue})$, but other models are also available. Notwithstanding its simplicity and properties, it suffer from some drawbacks such as illumination changes, shadows, camouflage, among others, which may inconvenience some tasks.

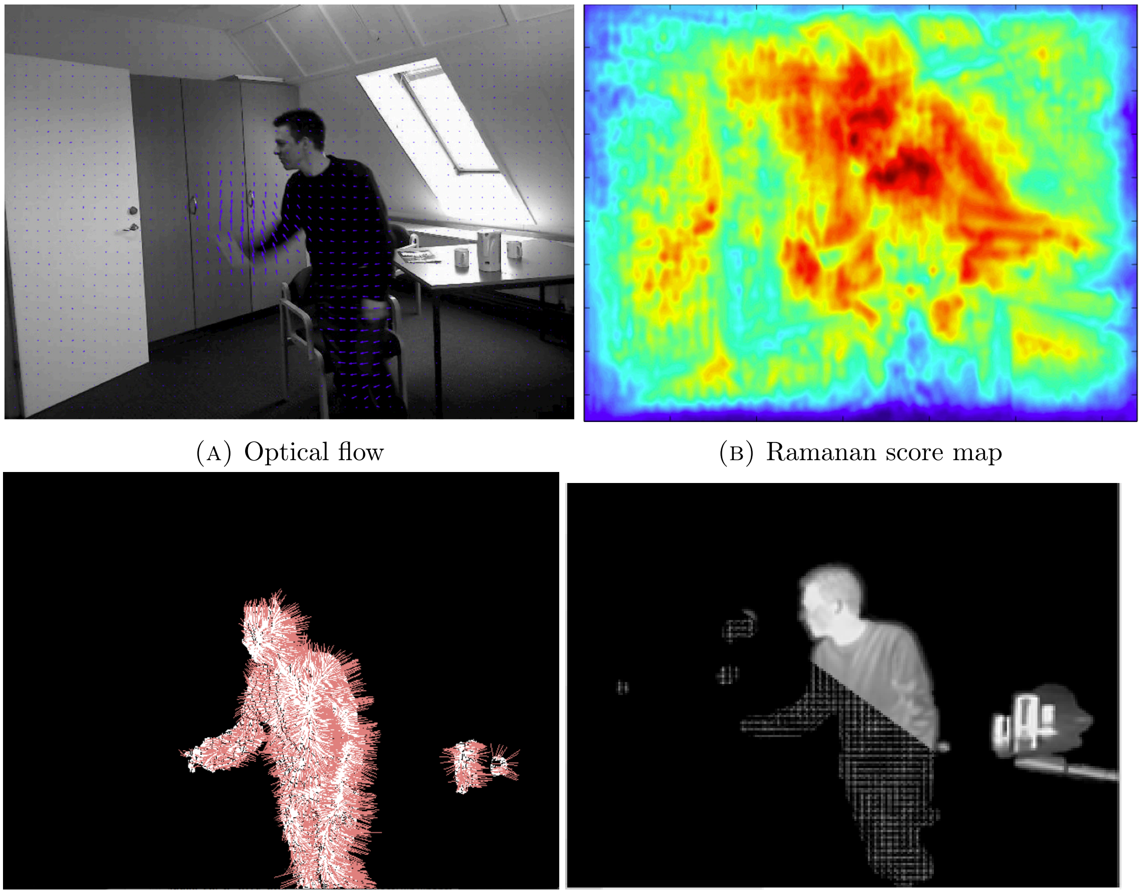
3.3.1.1 Histogram of oriented gradients (HOG)

For RGB cue, a simple implementation of HOG [9] is to be computed for each grid cell, known as detection window in the HOG context. Each window is resized to a $h_{\mathrm{w}}^{\mathrm{HOG}} \times v_{\mathrm{w}}^{\mathrm{HOG}}$ pixel area and divided in rectangular blocks of $h_{\mathrm{b}}^{\mathrm{HOG}} \times v_{\mathrm{b}}^{\mathrm{HOG}}$ pixels, which are in turn divided into rectangular local spatial regions called cells of size $h_{\mathrm{c}}^{\mathrm{HOG}} \times v_{\mathrm{c}}^{\mathrm{HOG}}$ pixels, thus having 4 cells per block and 8 blocks per window. We use RGB color space with no gamma correction. In order to compute the gradients, two kernels in the x and y directions with no smoothing are used for each channel so as to find and take the channel with the greatest gradient magnitude for each pixel. The gradient at point $\mathbf{p}$ of detection window is:

$$\mathcal{G}_{\mathbf{p}}^{x} = [\text{-}1\ 0\ 1] * \mathbf{C}_{\mathbf{p}} \tag{3.7}$$

$$\mathcal{G}_{\mathbf{p}}^{y} = [\text{-}1\ 0\ 1]^{\mathrm{T}} * \mathbf{C}_{\mathbf{p}} \tag{3.8}$$

The gradient magnitude $\mathbf{M}$ and orientation $\boldsymbol{\Theta}$ of the gradient at point $\mathbf{p}$ are:



(A) Optical flow

(B) Ramanan score map

(C) Depth normals

(D) Thermal intensities and oriented gradients

Figure 3.2: Example of descriptors from RGB modality: (A) represents the motion vectors using a forward scheme, that is, the optical flow orientation gives insight into where the person is moving to in the next frame; (B) score maps representing the hypothesis of a pixel belonging to a person; (C) the computed surface normal vectors; (D) the intensity and the thermal gradients orientations.

$$\mathbf{M}_{\mathbf{p}} = \sqrt{(\mathcal{G}^{\mathrm{x}}_{\mathbf{p}})^2 + (\mathcal{G}^{\mathrm{y}}_{\mathbf{p}})^2} \tag{3.9}$$

$$\mathbf{\Theta}_{\mathbf{p}} = \tan^{-1}\left(\frac{\mathcal{G}^{\mathrm{y}}_{\mathbf{p}}}{\mathcal{G}^{\mathrm{x}}_{\mathbf{p}}}\right) \tag{3.10}$$

Gradient orientation is also computed for each pixel in the dominant channel and assigned into a κ-bin histogram over each cell using unsigned gradients such that bins are evenly spaced over 0° and 180°. As stated in [9], signed contrast is uninformative for humans due to the wide range of clothing and background colors. For each gradient vector, its contribution to the histogram is given by the vector magnitude, that is, stronger magnitudes have a bigger impact on the histogram. Owing to local variations in illumination and foreground-background contrast gradient strengths vary over a wide range so cells are grouped into larger, spatially connected blocks. Hence, the information of each cell is concatenated. Then, the gradient strengths are locally normalized applying the L2-norm over each block. Block overlap is not applied in this case so as to lower the final descriptor dimensions.

3.3.1.2 Histogram of oriented optical flow (HOOF)

Since we are working with video sequences, 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 [14]. The optical-flow vectors of the whole image are computed using the luminance information of the image with the Gunnar Farnebäck's algorithm [56] available in OpenCV[1] [57], which is based on modeling the neighborhoods of each pixel of two consecutive frames by quadratic polynomials. It represents the image signal in the neighborhood of each pixel by a 3-D surface and finds where the surface has moved in the next frame. As a result, a set of 2-D vectors denoting the movement of each pixel for the horizontal $\mathbf{u}$ and vertical $\mathbf{v}$ directions in the compared frames is found. It allows a wide range of parameterizations, which will be specified in section 4.

The resulting motion vectors, whose example is shown in Fig. 3.2a, are masked and quantized to produce weighted votes for local motion based on their magnitude which are locally accumulated into a ν-bin histogram over each grid cell according to the signed (0° - 360°) vector orientations. In contrast to HOG, HOOF uses signed optical flow since the orientation provides more discriminative power. Magnitude and orientation of a motion vector at pixel $\mathbf{p}$ are calculated as in Eq. 3.9 and Eq. 3.10 respectively.

[1] `http://code.opencv.org.`

3.3.1.3 Score maps (SM)

In [20, 21] a method for detecting articulated people and estimating their pose from static images is described based on a new representation of deformable part models using a mixture of small, non-oriented parts in such a way that jointly captures spatial relations between part locations and co-occurrence relations between part mixtures. We take advantage of part of their model and the basic available implementation[2] so as to obtain a pixel-level measure, named score, that gives intuition into the presence of a person at a given location. It includes pre-trained full body models from the PARSE image dataset [17].

Briefly explained, their method uses a set of linear filters $\mathcal{F}$, which are rectangular templates that specify weight vectors for a particular human body part, having $M = 26$ body parts. Each part includes $C = 6$ different component filters. Therefore, we define $f_c^m \in \mathcal{F}$ to represent the c-th component filter that corresponds to the m-th body part, being $c = \{1, \dots, C\}$ and $m = \{1, \dots, M\}$. These filters are applied to dense feature maps computed by using a variation of HOG. The method also defines a feature pyramid H to obtain scores from placing the filters at different positions and scales of the pyramid, which is computed based on an initial image pyramid, in such a way that each feature map is computed from each level $l = \{1, \dots, L\}$ of the image pyramid, where L is the number of levels and depends on the original image size.

Let p_l denote the position (x, y) in the l-th level of the pyramid, and $\phi(H, p_l)$ a feature vector contained in the sub window of H with top-left corner at p_l whose dimensions are the defined by the filter. A part filter score can be considered as the response of the dot product between a filter and a subwindow of the feature map. Thus, the score of a point p_l for a given filter c of part m is:

$$\text{score}(p_l)_c^m = f_c^m \cdot \phi(H, p_l) \tag{3.11}$$

Since our aim is not to find separated human body parts but full-body detections, we computed a combination of the different parts scores for each level such that:

[2] http://www.ics.uci.edu/~dramanan/software/pose/

$$\text{score}(p_l) = \frac{1}{C}\frac{1}{M}\sum_{c\in C}\sum_{m\in M}\text{score}(p_l)_c^m \tag{3.12}$$

For obtaining the final score score(p), we proceed in a similar way with the scores obtained from the different levels. Note that the output score maps of each scale have different size so in order to compute the mean of the L different scaled score maps they are all resized to the original image size. Representing score$(p_l)'$ as the resized version of score(p_l):

$$\text{score}(p) = \frac{1}{L}\sum_{l\in L}\text{score}(p_l)' \tag{3.13}$$

The final score map for each color image $\mathbf{C}_i \in \mathbf{F}_i$ is an array:

$$\text{score}(\mathbf{C}_i) = \{\text{score}(p_1), \ldots, \text{score}(p_n)\}$$

with n equal to the size of the original frame. An example of score map is depicted in Fig. 3.2b.

3.3.2 Depth

The grid cells in the depth modality $\mathbf{G}_{ij}^{\text{depth}}$ are depth dense maps represented as planar images of pixels (in projective coordinates) that take depth values in millimeters. From the depth representation in projective coordinates it is possible to obtain the "real world" ones by using the intrinsic parameters of the depth sensor. This conversion generates 3-D point cloud structures $\mathcal{P}_{ij}$ in which the distances among points are actual distances – those that can be measured in the real world. Finally, in each point cloud $\mathcal{P}_{rij} \in \mathcal{P}_{ij}$ the surface normals are computed and their angles' distribution summarized in a δ-bin histogram, eventually describing the cell from the depth modality point of view.

3.3.2.1 Histogram of oriented depth normals (HON)

In order to describe an arbitrary point cloud $\mathcal{P}$ the surface normals vectors have to be computed first. A surface normal of a 3-D point is a perpendicular vector to a 3-D plane which is tangent to the surface in that point. Thus, the normal 3-D vector at a given point $\mathbf{p} = (p_x, p_y, p_z) \in \mathcal{P}$ can be seen as the problem of determining the normal of a 3-D plane tangent to $\mathbf{p}$. A plane is represented by the origin point $\mathbf{q}$ and the normal vector $\mathbf{n}$. From the neighboring points $\mathcal{K}$ of $\mathbf{p} \in \mathcal{P}$, we first set $\mathbf{q}$ to be the average of those points:

$$\mathbf{q} \equiv \bar{\mathbf{p}} = \frac{1}{|\mathcal{K}|} \sum_{\mathbf{p} \in \mathcal{P}^{\mathcal{K}}} \mathbf{p} \tag{3.14}$$

Then, the solution of $\mathbf{n}$ can be approximated using the covariance matrix $C \in \mathbb{R}^{3\times3}$ of the points in $\mathcal{P}^{\mathcal{K}}_{\mathbf{p}}$. The covariance matrix C is computed as follows:

$$C = \frac{1}{|\mathcal{K}|} \sum_{i=1}^{|\mathcal{K}|} (\mathbf{p}_i - \bar{\mathbf{p}})(\mathbf{p}_i - \bar{\mathbf{p}})^{\mathrm{T}} \tag{3.15}$$

being C a symmetric positive semi-definite matrix. Solving the next equation by means of eigenvalue decomposition:

$$C\mathbf{u}_j = \lambda_j \mathbf{u}_j,\ j \in \{0, 1, 2\} \tag{3.16}$$

where $\lambda_j \in \mathbb{R}$ and $\mathbf{u}_j \in \mathbb{R}^3$ represent the j-th eigenvalue and eigenvector of C respectively, a solution for $\mathbf{n}$ is found to be the eigenvector $\mathbf{u}_j$ with the associated smaller λ_j. Formally,

$$\mathbf{n} = \mathbf{u}_z, \quad \text{where } z =_j \lambda_j,\ j \in \{0, 1, 2\} \tag{3.17}$$

The sign of $\mathbf{n}$ can be either positive or negative, and it can not be disambiguated from the calculations. We adopt the convention of re-orienting consistently all the computed normal vectors towards the depth sensor's viewpoint $\mathbf{z}$. The computed normal vectors over a human body region is shown in Figure 3.2c. Points are illustrated in white, whereas normal vectors are red lines (instead of arrows for the sake of the visual comprehension). The next step is to build the histogram describing the distribution of the normal vectors' orientations.

A 3-D normal vector got from the previous calculations is expressed in cartesian coordinates (n_x, n_y, n_z). Nonetheless, a normal vector can be also expressed in spherical coordinates using three parameters: the radius s, the inclination θ, and the azimuth φ. In our case, s is a constant value, so this parameter can be omitted. Regarding θ and φ, the cartesian to spherical coordinates transformation is calculated as:

$$\theta = \left(\frac{n_z}{n_y}\right), \ \varphi = \frac{\sqrt{(n_y^2 + n_z^2)}}{n_x} \tag{3.18}$$

Therefore, a 3-D normal vector can be characterized by a pair (θ, φ) and the depth description of a cell consists of a pair of concatenated δ_θ-bin and δ_φ-bin histograms, describing the two angular distributions of the body surface normals within the cell. Moreover, each of the two histograms is normalized before the concatenation, dividing by the number of elements, to end up with a relative angles frequency count.

3.3.3 Thermal

The thermal cue is a very informative feature for the task of people detection/segmentation. A pixel part of a human region gives off heat and hence a relatively large value in terms of thermal intensity.

3.3.3.1 Histogram of thermal intensities and oriented gradients (HIOG)

The descriptor got from a cell in the thermal cue $\mathbf{G}_{rij}^{\text{thermal}}$ is the concatenation of two histograms. The first one is an histogram summarizing the thermal intensities, which range in the interval $[0, 255]$. The intensities are the ones in the masked region of the cell, i.e. not taking into account the background pixels. Instead, the second histogram is summarizing the orientations of thermal gradients. These gradients are computed convolving a first derivative kernel in both directions (as in Eq. 3.7-3.8). Then, their orientation is calculated and binned in the histogram weighted by their magnitude. Finally, the two histograms are normalized dividing by the sum of the accumulations in the bins and concatenated. We used τ_{i} bins for the intensities part and τ_{g} bins for the gradients' orientations.

3.4 Cell classification

Since we are intended to segment human body regions, we need to distinguish those from the other foreground regions segmented by the background subtraction algorithm. These other foreground regions, apart from subjects, are the objects – they could be other living beings, e.g. cats or dogs, though these are not considered in this work.

From the previous step, each grid cell has been described using the different descriptors $\mathcal{D}$. For the purpose of classification, we train several Gaussian Mixture Models for each grid position (i, j), kind of description $d \in \mathcal{D}$, and foreground class either subject or orbject. Concretely, the set of GMMs modeled from the set of grid cells positioned in (i, j) is $\mathcal{M}_{ij} = \{\mathcal{M}_{ij}^{d,\text{sub}}, \mathcal{M}_{ij}^{d,\text{obj}}\}_{\forall d \in \mathcal{D}}$. In Fig. 3.3, the different steps of the baseline up to this point are illustrated.

Then, in testing time, an unseen cell can be predicted to be a subject or an object depending on the likelihood got in the probability density function (PDF) of the different mixtures. We will denote the likelihood values got from the GMMs $\mathcal{M}_{ij}^{d,\text{sub}}$ and $\mathcal{M}_{ij}^{d,\text{obj}}$ by $\mathcal{L}_{ij}^{d,\text{sub}}$ and $\mathcal{L}_{ij}^{d,\text{obj}}$ respectively. The final classification of a G_{rij} is performed by combining somehow and comparing $\{\mathcal{L}_{rij}^{d,\text{sub}}\}_{\forall d \in \mathcal{D}}$ and $\{\mathcal{L}_{rij}^{d,\text{obj}}\}_{\forall d \in \mathcal{D}}$. How to intelligently combine the previous classification results is explained in Section 3.5.

3.4.1 Gaussian Mixture Models

Gaussian Mixture Models (GMMs) is an unsupervised learning method for fitting multiple Gaussians to a set of multi-dimensional data points[3]. It is often used as a probabilistic

[3] This technique uses properties of gaussians, thus its generalization to fit other functions is not straightforward.

clustering and an alternative to the k-means algorithm. As in the case of k-means, the number of components K (or gaussians) in the mixture is a parameter that needs to be specified to the algorithm. The GMMs are trained using the very general Expectation-Maximization algorithm [58]. The goal is to end up maximizing the overall likelihood $\mathcal{L}$ of the model:

$$\mathcal{L} = \prod_{\mathbf{x} \in \mathbf{X}}^{N} p(\mathbf{x}) \tag{3.19}$$

where $\mathbf{x}$ is a multi-dimensional data point (in this case representing the descriptor of an arbitrary grid cell $\mathbf{G}^{d}_{rij}$) and $p(\mathbf{x})$ is the probability of $\mathbf{x}$ being drawn from the model. This probability is the value assigned by the mixture PDF to that point, which is in fact a linear combination of K gaussian PDFs:

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(\mathbf{x}|k)P(k) \tag{3.20}$$

being $p(\mathbf{x}|k)$ the value assigned by the k-th gaussian PDF to $\mathbf{x}$ (the height of the PDF function at that point), whereas $P(k)$ is the importance, or weight, of the k-th component in the mixture. In fact, since the model is a mixture of gaussians, $p(\mathbf{x})$ can be expressed as the mixture of parametrized gaussian functions:

$$p(\mathbf{x}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)P(k), \tag{3.21}$$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi^{\rho/2}|\boldsymbol{\Sigma}|^{1/2}} \exp^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{3.22}$$

In order to be able to predict at some point new given examples, a training procedure is needed to model the parameters of the mixture, i.e. the means $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K\}$ and covariances matrices $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K\}$. This is done by the two-step procedure called Expectaction-Maximization.

3.4.2 Expectation-Maximization: modeling a GMM

Let be $\mathbf{X}$ the set of ρ-dimensional points and have initialized the parameters for the K components $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and the contribution of the components $\{P(k_1), ..., P(k_K)\}$. The first step to perform is the expectation calculation, or *E-Step*, that consists on computing the K posteriors for all the points $\mathbf{x} \in \mathbf{X}$. The posterior $P(k|\mathbf{x})$ is the probability the point $\mathbf{x}$ belongs to the component k, and it is exactly:

$$p_{k\mathbf{x}} = P(k|\mathbf{x}) = \frac{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)P(k)}{p(\mathbf{x})} \tag{3.23}$$

Next, it is the turn of the maximization step, or *M-step.* In this step, it is supposed the assignments of individual points are known but not the model. The parameters of the components and their weights are re-estimated – because of the previous calculations – as:

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{\mathbf{x}\in\mathbf{X}} p_{k\mathbf{x}}\,\mathbf{x}}{\sum_{\mathbf{x}\in\mathbf{X}} p_{k\mathbf{x}}} \tag{3.24}$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{\sum_{\mathbf{x}\in\mathbf{X}} p_{k\mathbf{x}}\,(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x} - \hat{\boldsymbol{\mu}}_k)^{\mathrm{T}}}{\sum_{\mathbf{x}\in\mathbf{X}} p_{k\mathbf{x}}} \tag{3.25}$$

$$\hat{P}(k) = \frac{1}{|\mathbf{X}|}\sum_{\mathbf{x}\in\mathbf{X}} p_{k\mathbf{x}} \tag{3.26}$$

It can be proven that alternating E and M steps, the algorithm converges to at least a local maximum of overall likelihoods. A typical initialization is to start with K randomly chosen data points as starting means, and equal covariance matrices. Nonetheless, convergence is sometimes slow, because of having many points laying in "plateaus". Another possibility, as it has been done in this work, is to use *k-means* to have a better initialization because of a more robust estimate of the initial parameters, increasing the convergence speed and the chances of finding a better solution.

Moreover, though not explained, dealing with likelihoods may cause underflow problems in the computations. The approach to cope with this problem is to apply logarithms, that is dealing with log-likelihoods instead of likelihoods. Despite this changes some calculations re-formulated using the what is called "log-sum-exp" trick, the EM algorithm is still a valid approach to maximize the log-likelihood of the model given $\mathbf{X}$.

3.5 Multi-modal fusion

Having different modalities and descriptions allow us to fuse them to have a more informative and richer representation of the scenario which in turn can improve the final segmentation result. Such fusion can be achieved using several approaches, which are detailed below.

Before fusing the results got in the GMMs from different modalities and classes, a normalization step is required. The more simple possible strategy that normally yields good results is to perform a min-max normalization. This is done simply by subtracting the min value of a set of values and dividing by the difference between the max and the min. In our case, the normalization is done within the set of log-likelihoods $\Delta_{ij}^{d} = \{\mathcal{L}_{ij}^{d,\text{sub}}, \mathcal{L}_{ij}^{d,\text{obj}}\}$ obtained subject and object GMMs using a kind of description $d \in \mathcal{D}$. Concretely, a log-likelihood $\xi \in \Delta_{ij}^{d}$ is min-max normalized:

$$\hat{\xi} = \frac{\xi - \min(\Delta_{ij}^{d})}{\max(\Delta_{ij}^{d}) - \min(\Delta_{ij}^{d})} \tag{3.27}$$

The normalized log-likelihoods $\hat{\mathcal{L}}_{ij}^{d,\text{sub}}, \hat{\mathcal{L}}_{ij}^{d,\text{obj}}$ range in the interval $[0,1]$, which will be further used to obtain the prediction $\hat{t}_{r}^{d} = \{0,1\}$ of a region r, which denotes if such region belongs to an object or a subject respectively, and the predicted set of binary masks $\hat{S}$, which defines the human body segmentation.

SM description has not been taken into account thus far owing to its uniqueness. However, it only can be combined with the other descriptions if they are represented in an equal manner. For this purpose, a set of object score maps is computed from the original subject score maps, by taking the maximum score $\text{score}_{\max}$ of the whole set and computing the inverse, such that:

$$\text{score}(p)^{obj} = \text{score}_{\text{max}} - \text{score}(p)^{sub} \tag{3.28}$$

Afterwards, to modify the description to a cell level, the scores are first normalized applying, again, min-max normalization. Then, the mean is computed for each cell to obtain a value per cell, as happens in the rest of the descriptions.

For the sake of simplicity, from this point we will consider the normalized version of SM to be part of the set of descriptors $\mathcal{D}$, and the cell probability-like values to be also part of Δ_{ij}^{d}, even though they do not represent exactly the same.

3.5.1 Individual prediction

The first step is to compute the individual prediction of each of the descriptions separately. Since all the descriptions are cell-based except for the non-normalized SM, which is pixel-based, we differentiate two different approaches of individual prediction.

3.5.1.1 Cell-based descriptions

A grid cell voting is performed using the normalized log-likelihoods of subject and object of each of them, comparing both to decide if that cell contains a person, such that

$$v = \sum_{i,j} \mathbb{1}\{\hat{\mathcal{L}}_{ij}^{d,\text{sub}} > \hat{\mathcal{L}}_{ij}^{d,\text{obj}}\} \tag{3.29}$$

We also define a threshold v_{thr} that refers to the minimum number of positive votes needed to assign the subject label to the given region. Such threshold is given by

$$v_{thr} = \frac{v_{\text{grid}} h_{\text{grid}}}{2} \tag{3.30}$$

However, the decision of a cell can be given by a little difference between both log-likelihoods. For this reason, in order to decide whether the region belongs to a person, such differences are taken into account if an agreement is not reached by the cells. The final decision is thus described as:

$$\hat{t}_r^d = \mathbb{1}\Big\{v > v_{thr}\Big\} \bigvee \Big\{\mathbb{1}\{v = v_{thr}\} \cdot \mathbb{1}\{\big(\sum_{i,j} \hat{\mathcal{L}}_{ij}^{d,\text{sub}} - \hat{\mathcal{L}}_{ij}^{d,\text{obj}}\big) > 0\Big\} \tag{3.31}$$

That is, in the event of a tie, the sum of the difference in log-likelihoods among cells would decide the final label for that region.

In order to create the predicted segmentation masks, we use the extracted foreground masks FG from the background subtraction step. For each frame, regions that belong to that frame are masked to 0 if they were predicted to be a subject, and left as they are otherwise. A prediction conflict could arise if there is an overlap between the bounding boxes that denote the regions and their labels differ. In that case, the overlapped region would be masked to 0 –thereby being considered as a subject –if, as applied before, the sum of the difference between subject and object likelihoods is negative, and kept unchanged in any other way.

3.5.1.2 Pixel-based descriptions

The color modality is also described using the score maps obtained by the Ramanan method, in such a way that each pixel has its own score score(p) of hypothesis of being part of a person. Therefore, we need a different way to obtain the prediction of a given region. To that end, two parameters are introduced: α, defining the minimum score of a pixel to be considered of a person; and η, denoting the minimum percentage of pixels inside a region that are considered as person that are needed to label the whole region as a person. Thus, being N_r the number of pixels of a region r the decision, is defined as:

$$\hat{t}_r^d = \mathbb{1}\Big\{\frac{1}{N_r}\sum_{i=1}^{N_r}\mathbb{1}\{\text{score}(p_i) > \alpha\} > \eta\Big\} \tag{3.32}$$

The predicted segmentation masks are created in a similar way to the cell-based descriptions case. The decision in case of prediction conflict between bounding boxes, however, is tackled in an specific manner. As we will explain in Section , α and η may differ depending on the cross-validation settings. Being N_b the number of conflicting bounding boxes, and α_b and η_b the specific parameters for bounding box b, the conflicting region r_{overlap} will be marked as person if the following expression holds:

$$\frac{1}{N_b}\sum_{j=1}^{N_b}\Bigg(\frac{1}{N_r}\sum_{i=1}^{N_r}\mathbb{1}\{\text{score}(r_{\text{overlap}}) > \alpha_b\}\Bigg) > \frac{1}{N_b}\sum_{j=1}^{N_b}\eta_b \tag{3.33}$$

3.5.2 Naive approach

A basic fusion approach is to combine the descriptors in such a way that all contribute with equal weight. We propose a cell-level fusion using the normalized subject and object log-likelihoods and modified score maps. Consequently, a voting stage is first performed among all descriptors, whose individual predictions $\hat{t}_r^d$ are the votes, such that:

$$v = \sum_{d \in \mathcal{D}} \hat{t}_r^d \tag{3.34}$$

Note that some of the predictions may be wrong, thereby affecting negatively the voting. If the majority of the votes consider the region to be a person, meaning that there is a strong agreement between descriptions, those descriptions that consider the given region to be a subject will not be taken into account in the cell-level fusion stage. Defining the selected descriptions as $\mathcal{D}' \subset \mathcal{D}$, their combination $\bar{\mathcal{L}}_{ij}^{d,\text{sub}}$ and $\bar{\mathcal{L}}_{ij}^{d,\text{obj}}$ is given by:

$$\bar{\mathcal{L}}_{ij}^{d,\text{sub}} = \sum_{d \in \mathcal{D}'} \hat{\mathcal{L}}_{ij}^{d,\text{sub}} \tag{3.35}$$

$$\bar{\mathcal{L}}_{ij}^{d,\text{obj}} = \sum_{d \in \mathcal{D}'} \hat{\mathcal{L}}_{ij}^{d,\text{obj}} \tag{3.36}$$

Once the combined log-likelihoods are obtained, the approach follows the same procedure as the individual prediction for cell-based descriptions to obtain a final prediction $\hat{t}_r$ for each region, and similarly to obtain the segmentation masks. Note that this time, since a fused result is obtained , there will only be one set of segmented masks for all the descriptions –except for the thermal description whose FG masks were different owing to the non-accurate pixel-wise registration among modalities.

3.5.3 SVM-based approach

Support vector machines is a non-probabilistic supervised binary classifier that learns a model which represents the instances as points in space, mapped in such a way that instances of different classes are separated by a hyperplane in a high dimensional space. However, if the dataset is not linearly separable in that space the hyperplane will fail in classifying properly. This can be solved by mapping the dataset instances into a higher dimensional space using a kernel function, thus making easier the dataset division [10].

SVM is often used in the literature as a discriminative classifier for object recognition, and particularly in human detection approaches, usually yielding successful results. We therefore propose a fusion using SVM with different types of kernel. In particular, our baseline includes linear kernel and radial basis function (RBF). Linear SVM just requires one parameter, the penalty parameter ζ, which specifies a trade-off between model complexity and misclassification of training examples and can take values in the interval $[0, inf]$. Higher values of ζ causes closer fitting to the training set, which may tend to overfitting. The performance of the RBF kernel is also influenced by the γ parameter, which controls the shape of the separating hyperplane. Higher values usually increase the number of support vectors. Weights $\mathbf{w}$ obtained from linear SVM represent the hyperplane used to separate between classes, but they can also give us an insight of the level of importance or level of influence that each feature has when classifying the instance [59]. The SVM models has been trained using the available implementation of the LibSVM[4] library [60].

Let $\mathcal{R}$ be the data set of regions and features that will be used in the SVM approach. Each region will be described by the different probabilities of subject and object of the different descriptions $\{\hat{\mathcal{L}}_{ij}^{d,\text{sub}}, \hat{\mathcal{L}}_{ij}^{d,\text{obj}}\}$ such that each region r will be described by $2 \times |\mathcal{D}| \times v_{\text{grid}} \times h_{\text{grid}}$ feature values. The computed ground truth labels $t_r^m \in T$ will be used in the training stage to indicate the class of r. Note that some regions have an unknown label $\{-1\}$; such regions are not used for training. Region prediction will be again denoted by $\hat{t}_r$.

However, we are not taken into account the previous predictions obtained for each description individually. In order to take into account this information, which may help in the classification stage, the set of individual predictions $\hat{t}_r^d$ can be included in the data set $\mathcal{R}$, such that each region r is described by $2 \times |\mathcal{D}| \times v_{\text{grid}} \times h_{\text{grid}} + |\mathcal{D}|$. This approach follows the Stacked Learning scheme [61, 62], which involves training a new learning algorithm combining previous predictions obtained with other learning algorithms.

Accordingly, we use four SVM classifiers: (1) simple linear SVM, (2) simple RBF SVM, (3) stacked linear SVM, and (4) stacked RBF SVM. Segmented masks are created following, again, the aforementioned procedure.

[4] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

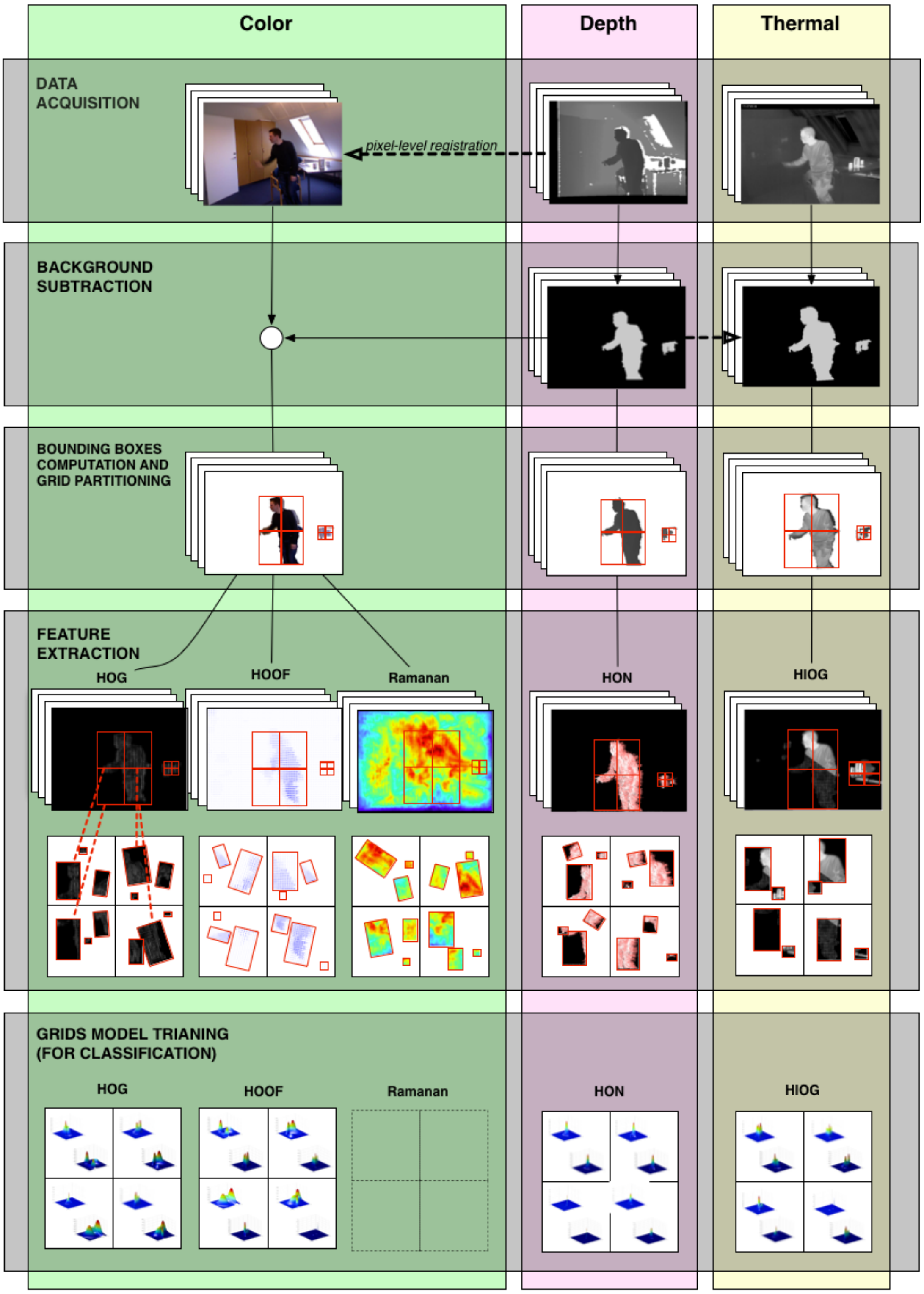


FIGURE 3.3: The main steps of the proposed baseline method, before reaching the fusion step.

Chapter 4

Evaluation

4.1 Parameters and settings

After some experiments regarding the use of Otsu's threshold in the background subtraction and generation of bounding boxes stage, we set $\sigma_{\text{otsu}} = 8.3$ for a connected component area of at least 0.1% of the image, or $\sigma_{\text{otsu}} = 12$ for other cases.

Since it is not possible to have a pixel-to-pixel correspondence among modalities, we define the correspondence at a grid cell level. The grids have been partitioned in $M \times N$ cells, being $M = 2$ and $N = 2$. The main idea of the grid partitioning is to reduce the variability of the regions in each GMM. At the same time, they are large enough to not condition the eventually computed overlap measure.

For the HOG descriptor, we defined: $H_{\text{w}} = 64 \times 128$, $H_{\text{b}} = 32 \times 32$, $H_{\text{c}} = 16 \times 16$ and $H_h = 9$. The information of each cell is concatenated resulting in a vector of 36 values per block. This brings the final vector size of a grid cell to 4 blocks vertically × 2 blocks horizontally × 4 cells per block × 9 bins per cell = 288 components.

In order to compute the optical flow, and based on the tests performed in [63], we set the parameters of the given implementation according to the values that gave the best performance. In particular, the averaging window size was set to 2, the size of the pixel neighborhood considered when finding polynomial expansion in each pixel was set to 5 and the standard deviation of the Gaussian that is used to smooth derivatives used as a basis for the polynomial expansion to 1.1. The remaining parameters were set to their default OpenCV values. For the HOOF descriptor, we defined $V_{\mathrm{b}} = 8$, to finally produce an 8-D feature vector.

For the depth descriptors, we defined $\theta_I = 8$ and $\phi_G = 8$.

For the thermal descriptors, we defined $T_I = 8$ and $T_G = 8$.

The parameter set up in the training of the GMMs is simply the number of mixtures, which have been set to a typical value of 3 mixture components.

4.2 Experimental methodology and validation measures

The proposed baseline has been validated by means of a K-fold cross-validation (CV). The R regions of interest have been divided in disjoint partitions, in which the cells' classifications and log-likelihoods' normalizations have been performed independently. In each iteration of the cross-validation, K-1 partitions are used to train the GMMs and the other one is used for testing, that is, each region of interest in the test set is predicted (at cell level) using models trained in an independent dataset (train set). Once all the regions in the K different test partitions have been predicted, all the regions throughout the sequence of frames have been also predicted, and a final performance measure can be computed at frame level comparing the results of the predictions with the groundtruth, e.g. an overlap measure, explained below.

Moreover, in order to select the α and η parameters for the individual prediction of the SM descriptor, a coarser-to-fine search strategy has been followed. The first coarse grid search is utilized to roughly estimate their value. In this search, a K-fold CV has tested 6×5 combinations; α took the middle 6 of the 8 equidistant values in the range $[\mathrm{score}_{\mathrm{min}}, \mathrm{score}_{\mathrm{max}}]$ and η the middle 5 of 7 equidistant values in the range $[0, 1]$. Posteriorly, the fine search around the best coarse combination in each fold has been performed to find the best fine combination. A second K-fold CV tested the fine combinations, which consisted of a 7×5 grid centered in the corresponding best coarse combination. In this case the criterion to guide the search of the parameters selection is simply the subject detection accuracy, got comparing the result of the prediction to the ground truth.

Another coarse-to-fine grid search has been applied in order to select the SVM parameters γ and ζ. The coarse search is first used to identify a better region on the grid. For linear SVM, ζ is searched in the range $[2^{-5}, 2^{15}]$ in steps of 2^2, that is, 11 values. RBF SVM uses in turn the same range of values for ζ, whereas γ is searched in the interval $[2^-13, 2^3]$ in steps of 2^2, thus testing 11×9 combinations. After finding the best combination, a finer grid search on that region has been conducted, varying in $2^{1.5}$ in each direction centered on the value that produced the highest classification accuracy. Both procedures have been validated with K-fold CV, using the computed ground truth labels $t_r^m \in T$ to train the models.

Table 4.1: Best cross-validation results for parameter selection of SVM models

	Linear			**RBF**		
SVM Type	γ	ζ	accuracy	γ	ζ	accuracy
Simple	-	45.2548	96.56 %	5.6569	22.6274	97.65 %
Stacked	-	1024	96.78 %	0.5	512	97.67 %

Lastly, we have used the Jaccard Index [64], also known as the Jaccard similarity coefficient, in order to compare the similarity between the groundtruth masks and the predicted masks in terms of overlap, thus assessing the performance of the proposed baseline. The degree of overlap between two binary sets A and B is computed as the ratio between the size of the intersection divided by the size of the union:

$$overlap(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{4.1}$$

This measure takes values in $[0, 1]$, 0 meaning no overlap and 1 meaning perfect agreement between sets. GT represents connected components of the ground truth binary masks, and $\hat{S}$ those of predicted binary masks from the different modalities individually or from the different fusion approaches. For each frame, the overlap is computed per person id and connected component, in such a way that connected components that have the same person id or are connected in the ground truth constitute a set A, and they are compared to the blobs that coincide in the predicted binary masks, which constitute a set B. The overlap of each frame is then averaged by the number of sets found. It is therefore a pessimistic measure because a very tiny blob misclassified as a person in the predicted binary masks will account for 0 overlap, thus decreasing the mean overlap of the frame, so it can be considered as a lower bound on how accurate the prediction is. The final overlap is computed as the mean overlap of all frames having at least one blob, whether they be in the ground truth or in the predicted binary mask.

As commented in Section 3.1.1, the depth cue suffers from a halo effect around people or objects, thus complicating an accurate pixel-level segmentation at blob contours when applying background subtraction. This lack of accuracy is also caused by possible distortions, noise or other problems, and decreases the final overlap. Hence, a *do not care region* (DCR) is often used. Such region is taken per frame by centering a morphology operator of different sizes at the ground truth binary masks blob contours and subtract it from those masks and from the predicted ones to compute the overlap. This way, we can compare the effect of using a growing DCR to the actual overlap.

4.3 Experimental results

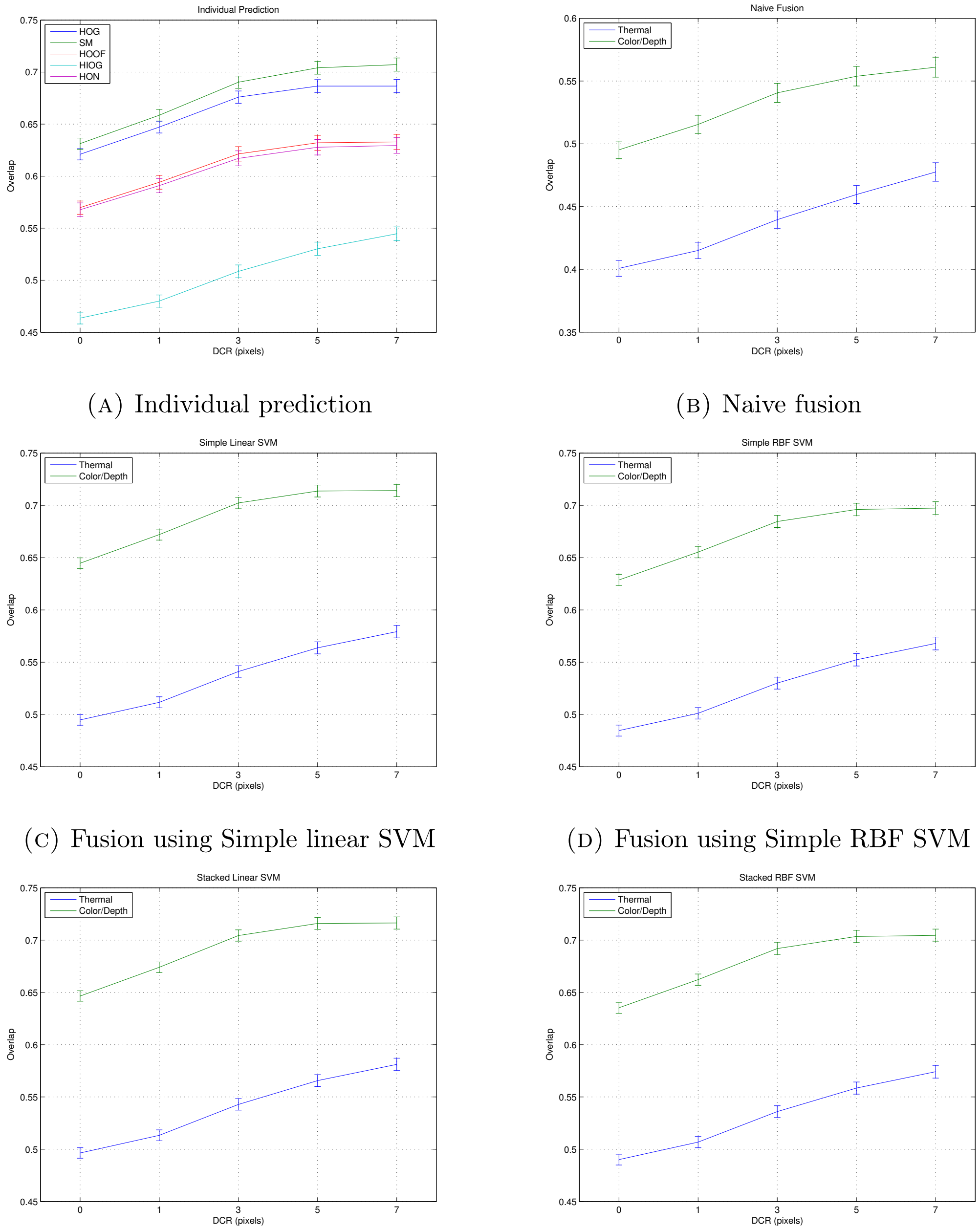
As explained in the last section, we assess the performance of the proposed baseline using the Jaccard overlap measure (Eq. 4.1). Figure 4.1 depicts the obtained overlap for individual predictions and fusion predictions with different fusion approaches. Tables 4.2 and 4.3 are included to compare the differences between using the descriptors separately and after fusing them. Notice that in plots showing fusion results, only two cases are considered, owing to color and depth modalities share the same original FG masks.

Table 4.2: Overlap results of the individual predictions for each description

	HOG	**SM**	**HOOF**	**HIOG**	**HON**
0	62.10 %	63.12 %	56.97 %	46.35 %	56.76 %
1	64.71 %	65.85 %	59.41 %	47.99 %	59.09 %
3	67.59 %	69.02 %	62.13 %	50.85 %	61.70 %
5	68.65 %	70.40 %	63.20 %	53.02 %	62.77 %
7	68.65 %	70.72 %	63.28 %	54.45 %	62.94 %

Table 4.3: Overlap results of fusion using Stacked Linear SVM for each modality

DCR	**Thermal**	**Color/Depth**
0	49.64 %	64.65 %
1	51.33 %	67.39 %
3	54.29 %	70.43 %
5	56.56 %	71.58 %
7	58.11 %	71.63 %



(A) Individual prediction

(B) Naive fusion

(C) Fusion using Simple linear SVM

(D) Fusion using Simple RBF SVM

(E) Fusion using Stacked Linear SVM

(F) Fusion using Stacked RBF SVM

FIGURE 4.1: Overlap results for the different individual and fusion prediction approaches.

4.4 Discussion

The obtained results show that, effectively, fusing different descriptions enhances the representation of the scene, thus increasing the final overlap when segmenting subjects and discriminating from other artifacts present in the scene. However, the selection of the fusion approach is crucial. Our proposed naive approach for fusing individual confidences of each modality decreases the overlap of the color modalities up to 8%. On the other hand, as observed by the SVM experiments and in particular the stacked SVM experiments including the prediction labels as new features, we obtain significant performance improvements regarding each individual modality and the naïve fusion strategy. More precisely, we achieve the best results using the stacked version of the linear SVM kernel, thereby increasing the overlap considerably. Surprisingly, linear kernel outperforms RBF by 2%, and stacked versions slightly improve the simple ones.

Figure 4.1a demonstrates that the most informative descriptions are HOG and SM. It is important to note that thermal descriptions cannot reach as good overlap values as the other modalities owing to the fact that the binary masks FG^{thermal} were created from FG^{depth} using the registration algorithm, which cannot be accurate up to pixel level, in such a way that the ground truth and registered masks will moderately differ, especially in left and right sides of the image. Therefore, we cannot state whether the proposed thermal description performs accurately.

Furthermore, an upward trend is observed as DCR grows, although at higher DCR levels it stabilizes. This is comprehensible because usually the contours of the predicted masks are not accurately defined. Indeed, an accurate pixel-level segmentation is a rather complex task in state of the art techniques.

Having analyzed the experimental results, it is worth investigating the causes of some misclassifications. One of the problems is originated in the beginning of the chain. Since background subtraction reduces the search space, it may reject some actual person parts. That mainly happens when a person is in the same depth than something which belongs to the background model. Another issue is that some regions considered as unknown – mostly those generated when one person overlaps other – considerably differ from those that are used to train the different models. Consequently, the classification of such regions is not a trivial task.

Chapter 5

Conclusions and future work

In this master thesis project it has been proposed a solution for human body segmentation in multi-modal data. A baseline method to segment people using different cues has been proposed. Furthermore, a novel registered and annotated multi-modal RGB-Depth-Thermal data set of several video sequences has been introduced, which contains several subjects interacting with everyday objects.

The first contribution of this work was an adaptive multi-modal background subtraction approach in order to extract the regions of interest that belong to a user or a moving object in the scene with high confidence. From the set of regions of interest coming from the different data modalities, different state of the art descriptors have been used and adapted to describe different feature vectors from each region. In particular, HOG, HOF, and gabor-based features have been computed from RGB still images and image sequences, histogram of intensity gradients from thermal data, and histograms of normal vectors from depth maps coming from infrared sensors. The set of descriptors have been selected as the most discriminative ones given the results previously reported in literature.

Given the proposed and adapted descriptions, we learnt a Gaussian mixture model for each distribution of feature vectors from both objects and users in a grid fashion, obtaining a set of confidence score from each region of interest part belonging to user or object. Those confidence scores were used independently to segment users in the different modalities. As shown in our results, the segmentation performance by each modality varies and they offer complementary information. In this sense, we proposed two fusion strategy methodologies to combine the scores of independent modalities. In our first naïve approach, a simple combination and threshold-based rule is proposed, which did not offer improved accuracy. Our second proposal was to combine the confidence scores and previous GMM predictions as new feature vectors for SVM classifiers, in a stacked learning fashion. As a result, we found significant performance improvements of the proposed fusion strategy in comparison to each isolate modality. More interestingly, the included predictions from previous classifiers enhanced final segmentation performance. Thus, in conclusion, the results have shown variable performance for the different modalities when segmenting people with multi-modal information, being the multi-modal GMM-SVM stacked learning method the one that has achieved the best results.

Despite the obtained results, this proposal clearly leaves further room of improvement. To begin with, the first background subtraction stage could combine the different modalities in order to learn the model. Furthermore, a clustering of poses at cell-level could be added before learning the GMMs. GrabCuts could also be applied to the predicted segmentation binary masks to refine and smooth the contours, which would also produce a rise in the segmentation accuracy. Finally, if all the modalities were aligned up to pixel level, local-based feature extraction and description could be carried out. In that sense, as future work we plan to use that local information to allow the method to discard from the final segmented binary masks those objects that are next to the segmented subjects because of user-object interactions.

Bibliography

[1] Claude R Brice and Claude L Fennema. Scene analysis using regions. *Artificial intelligence*, 1(3):205–226, 1970.

[2] Edward M Riseman and Michael A Arbib. Computational techniques in the visual segmentation of static scenes. *Computer Graphics and Image Processing*, 6(3):221–276, 1977.

[3] Ron Ohlander, Keith Price, and D Raj Reddy. Picture segmentation using a recursive region splitting method. *Computer Graphics and Image Processing*, 8(3): 313–333, 1978.

[4] Azriel Rosenfeld and Larry S Davis. Image segmentation and image models. *Proceedings of the IEEE*, 67(5):764–772, 1979.

[5] Robert M Haralick and Linda G Shapiro. Image segmentation techniques. *Computer vision, graphics, and image processing*, 29(1):100–132, 1985.

[6] Richard Szeliski. *Computer vision: algorithms and applications*. Springer, 2011.

[7] Chris Stauffer and W Eric L Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2. IEEE, 1999.

[8] Thierry Bouwmans. Recent advanced statistical background modeling for foreground detection: A systematic survey. *RPCS*, 4(3):147–176, 2011.

[9] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[10] Marti A. Hearst, ST Dumais, E Osman, John Platt, and Bernhard Scholkopf. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18–28, 1998.

[11] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Monocular 3d pose estimation and tracking by detection. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 623–630. IEEE, 2010.

[12] Anurag Mittal, Liang Zhao, and Larry S Davis. Human body pose estimation using silhouette shape analysis. In *Proceedings. IEEE Conference on Advanced Video and Signal Based Surveillance, 2003.*, pages 263–270. IEEE, 2003.

[13] Paul Viola, Michael J Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161, 2005.

[14] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *Computer Vision–ECCV 2006*, pages 428–441. Springer, 2006.

[15] Christian Plagemann, Varun Ganapathi, Daphne Koller, and Sebastian Thrun. Real-time identification and localization of body parts from depth images. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 3108–3113. IEEE, 2010.

[16] Bangpeng Yao and Li Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 9–16. IEEE, 2010.

[17] Deva Ramanan. Learning to parse images of articulated bodies. In *Advances in neural information processing systems*, pages 1129–1136, 2006.

[18] Hamed Pirsiavash and Deva Ramanan. Steerable part models. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3226–3233. IEEE, 2012.

[19] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1014–1021. IEEE, 2009.

[20] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1385–1392. IEEE, 2011.

[21] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures-of-parts. 2012.

[22] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[23] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1365–1372. IEEE, 2009.

[24] Yang Wang, Duan Tran, and Zicheng Liao. Learning hierarchical poselets for human parsing. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1705–1712. IEEE, 2011.

[25] Ross B Girshick, Pedro F Felzenszwalb, and David A Mcallester. Object detection with grammar models. In *Advances in Neural Information Processing Systems*, pages 442–450, 2011.

[26] Long Zhu, Yuanhao Chen, Yifei Lu, Chenxi Lin, and Alan Yuille. Max margin and/or graph learning for parsing the human body. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[27] Bastian Leibe, Ales Leonardis, and Bernt Schiele. Combined object categorization and segmentation with an implicit shape model. In *Workshop on Statistical Learning in Computer Vision, ECCV*, volume 2, page 7, 2004.

[28] Bastian Leibe, Aleš Leonardis, and Bernt Schiele. Robust object detection with interleaved categorization and segmentation. *International journal of computer vision*, 77(1-3):259–289, 2008.

[29] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[30] Chad Carson, Serge Belongie, Hayit Greenspan, and Jitendra Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(8): 1026–1038, 2002.

[31] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.

[32] Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary &amp; region segmentation of objects in nd images. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 105–112. IEEE, 2001.

[33] Antonio Hernández-Vela, Nadezhda Zlateva, Alexander Marinov, Miguel Reyes, Petia Radeva, Dimo Dimov, and Sergio Escalera. Graph cuts optimization for multi-limb human segmentation in depth maps. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 726–732. IEEE, 2012.

[34] Zhe Lin, Larry S Davis, David Doermann, and Daniel DeMenthon. An interactive approach to pose-assisted and appearance-based segmentation of humans. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[35] Greg Mori, Xiaofeng Ren, Alexei A Efros, and Jitendra Malik. Recovering human body configurations: Combining segmentation and recognition. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–326. IEEE, 2004.

[36] L'ubor Ladickỳ, Paul Sturgess, Karteek Alahari, Chris Russell, and Philip HS Torr. What, where and how many? combining object detectors and crfs. In *Computer Vision–ECCV 2010*, pages 424–437. Springer, 2010.

[37] Anat Levin and Yair Weiss. Learning to combine bottom-up and top-down segmentation. In *Computer Vision–ECCV 2006*, pages 581–594. Springer, 2006.

[38] M Pawan Kumar, PHS Ton, and Andrew Zisserman. Obj cut. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 18–25. IEEE, 2005.

[39] Matthieu Bray, Pushmeet Kohli, and Philip HS Torr. Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. In *Computer Vision–ECCV 2006*, pages 642–655. Springer, 2006.

[40] Li Zhang, Bo Wu, and Ram Nevatia. Pedestrian detection in infrared images based on local shape features. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[41] James W Davis and Vinay Sharma. Robust background-subtraction for person detection in thermal imagery. *IEEE Int. Wkshp. on Object Tracking and Classification Beyond the Visible Spectrum*, 2004.

[42] Andreas Møgelmose, Albert Clapés, Chris Bahnsen, Thomas B Moeslund, and Sergio Escalera. Tri-modal person re-identification with rgb, depth and thermal features. *Perception Beyond the Visible Spectrum.*

[43] Rikke Gade and Thomas B Moeslund. Thermal cameras and applications: a survey. *Machine Vision and Applications*, 25(1):245–262, 2014.

[44] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, pages 1–8, 2008. doi: 10.1109/CVPRW.2008.4562953.

[45] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008.

[46] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 416–423. IEEE, 2001.

[47] S. Alpert, M. Galun, R. Basri, and A. Brandt. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007. doi: 10.1109/CVPR.2007.383017.

[48] M Everingham, L Van Gool, CKI Williams, J Winn, and A Zisserman. The pascal visual object classes challenge 2012 results. See http://www. pascal-network. org/challenges/VOC/voc2012/workshop/index. html, 2012.

[49] Varun Gulshan, Victor Lempitsky, and Andrew Zisserman. Humanising grabcut: Learning to segment humans using the kinect. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1127–1133. IEEE, 2011.

[50] Thomas B Moeslund. *Visual analysis of humans: looking at people.* Springer, 2011.

[51] Anh T Nghiem, Francois Bremond, Monique Thonnat, and Valery Valentin. Etiseo, performance evaluation for video surveillance systems. In *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 476–481. IEEE, 2007.

[52] Luciano Spinello and Kai O Arras. People detection in rgb-d data. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 3838–3843. IEEE, 2011.

[53] Thierry Bouwmans, Fida El Baf, Bertrand Vachon, et al. Background modeling using mixture of gaussians for foreground detection-a survey. *Recent Patents on Computer Science*, 1(3):219–237, 2008.

[54] Zoran Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 28–31. IEEE, 2004.

[55] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[56] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis*, pages 363–370. Springer, 2003.

[57] Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer vision with the OpenCV library*. O'reilly, 2008.

[58] Todd K Moon. The expectation-maximization algorithm. *Signal processing magazine, IEEE*, 13(6):47–60, 1996.

[59] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[60] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[61] William W Cohen. Stacked sequential learning. Technical report, DTIC Document, 2005.

[62] Carlo Gatta, Eloi Puertas, and Oriol Pujol. Multi-scale stacked sequential learning. *Pattern Recognition*, 44(10):2414–2426, 2011.

[63] Karla Brkić, Srđan Rašić, Axel Pinz, Siniša Šegvić, and Zoran Kalafatić. Combining spatio-temporal appearance descriptors and optical flow for human action recognition in video data. *arXiv preprint arXiv:1310.0308*, 2013.

[64] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 32–41. ACM, 2002.

[65] Craig M Shakarji et al. Least-squares fitting algorithms of the nist algorithm testing system. *JOURNAL OF RESEARCH-NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY*, 103:633–641, 1998.

[66] Radu Bogdan Rusu. Semantic 3d object maps for everyday manipulation in human living environments. *KI-Künstliche Intelligenz*, 24(4):345–348, 2010.