Multi-modal RGB–Depth–Thermal Human Body Segmentation

Cristina Palmero[1,2] · Albert Clapés[1,2] · Chris Bahnsen[3] · Andreas Møgelmose[3] · Thomas B. Moeslund[3] · Sergio Escalera[1,2]

Received: 11 November 2014 / Accepted: 14 March 2016 / Published online: 13 April 2016

Abstract This work addresses the problem of human body segmentation from multi-modal visual cues as a first stage of automatic human behavior analysis. We propose a novel RGB–depth–thermal dataset along with a multi-modal segmentation baseline. The several modalities are registered using a calibration device and a registration algorithm. Our baseline extracts regions of interest using background subtraction, defines a partitioning of the foreground regions into cells, computes a set of image features on those cells using different state-of-the-art feature extractions, and models the distribution of the descriptors per cell using probabilistic models. A supervised learning algorithm then fuses the output likelihoods over cells in a stacked feature vector representation. The baseline, using Gaussian mixture models for the probabilistic modeling and Random Forest for the stacked learning, is superior to other state-of-the-art methods, obtaining an overlap above 75 % on the novel dataset when compared to the manually annotated ground-truth of human segmentations.

Keywords Human body segmentation · RGB · Depth · Thermal

1 Introduction

Human body segmentation is the first step used by most human activity recognition systems (Poppe 2010). Indeed, an accurate segmentation of the human body and correct person identification are key to successful posture recovery and behavior analysis tasks, and they benefit the development of a new generation of potential applications in health, leisure, and security.

Despite these advantages, segmentation of people in images poses a challenge to computer vision. The main difficulties arise from the articulated nature of the human body, changes in appearance, lighting conditions, partial occlusions, and the presence of background clutter. Although extensive research has been done on the subject, some constraints must be considered. The researcher must often make assumptions about the scenario where the segmentation task is to be applied, such as static versus moving camera and indoor versus outdoor location, among other factors. Ideally, it should be tackled in an automatic fashion rather than rely on user intervention, which makes such tasks even more challenging.

Communicated by Junsong Yuan, Wanqing Li, Zhengyou Zhang, David Fleet, Jamie Shotton.

Electronic supplementary material The online version of this article (doi:10.1007/s11263-016-0901-x) contains supplementary material, which is available to authorized users.

✉ Cristina Palmero
c.palmero.cantarino@gmail.com

Albert Clapés
aclapes@cvc.uab.cat

Chris Bahnsen
cb@create.aau.dk

Andreas Møgelmose
am@create.aau.dk

Thomas B. Moeslund
tbmg@create.aau.dk

Sergio Escalera
sergio@maia.ub.es

[1] Dept. Matemàtica Aplicada i Anàlisi, UB, Gran Via de les Corts Catalanes 585, 08007 Barcelona, Spain

[2] Computer Vision Center, Campus UAB, Edifici O, 08193 Cerdanyola del Vallès, Spain

[3] Aalborg University, Sofiendalsvej 11, 9200 Aalborg SV, Denmark

Most state-of-the-art methods that deal with such task use color images recorded by RGB cameras as the main cue for further analysis, although they present several widely known intrinsic problems, such as similarities in the intensity of background and foreground. More recently, the release of RGB–depth devices such as Microsoft Kinect® and the new Kinect 2 for Windows® has allowed the community to use RGB images along with per-pixel depth information. Furthermore, thermal imagery is becoming a complementary and affordable visual modality. Indeed, having different modalities and descriptions allow us to fuse them to have a more informative and richer representation of the scene. In particular, color modality adds contour and texture information and depth data provides the geometry of the scene, while thermal imaging adds temperature information.

In this paper we present a novel dataset of RGB–depth–thermal video sequences that contains up to three individuals who appear concurrently in three indoor scenarios, performing diverse actions that involve interaction with objects. Sample imagery of the three recorded scenes is depicted in Fig. 1. The dataset is presented along with an algorithm that performs the calibration and registration among modalities. In addition, we propose a baseline methodology to automatically segment human subjects appearing in multi-modal video sequences. We start reducing the search space by learning a model of the scene to subsequently perform background subtraction, thus segmenting subject candidate regions in all available and registered modalities. Such regions are then described using simple but reliable uni-modal feature descriptors. These descriptors are used to learn probabilistic models so as to predict the candidate region that actually belongs to people. In particular, likelihoods obtained from a set of Gaussian mixture models (GMMs) are fused in a higher level representation and modeled using a Random Forest classifier. We compare results from applying segmentation to the different modalities separately to results obtained by fusing features from all modalities. In our experiments, we demonstrate the effectiveness of the proposed algorithms to perform registration among modalities and to segment human subjects. To the best of our knowledge, this is the first publicly available 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 new—and more robust—solutions.

The remainder of this paper is organized as follows: Sect. 2 reviews the different approaches for human body segmentation that appear in the recent literature. Section 3 presents the new dataset, including acquisition details, the calibration device, the registration algorithm, and the ground-truth annotation. Section 4 presents the proposed baseline methodology for multi-modal human body segmentation, which is experimentally evaluated in Sect. 5 along with the registration algorithm. We present our conclusions in Sect. 6.

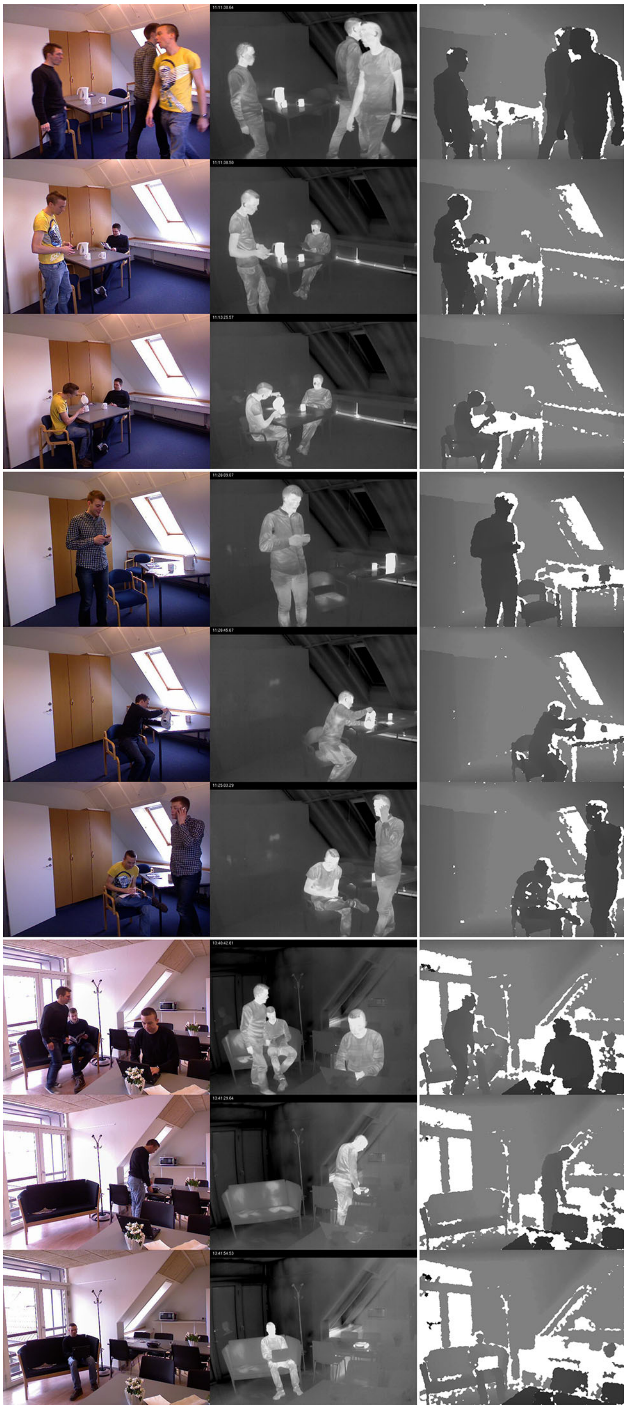

Fig. 1 Three views of each of the three scenes shown in the RGB, thermal, and depth modalities, respectively

2 Related Work

Multi-modal fusion strategies have gained attention lately due to the decreasing price of sensors. They are usually based on existing modality-specific methods that, once combined, enrich the representation of the scene in such a way that the weaknesses of one modality are offset by the strengths of another. Such strategies have been successfully applied to the human body segmentation task, which is one of the most widely studied problems in computer vision.

In this section we focus on the most recent and relevant studies, techniques and methods of individual and multi-modal human body segmentation. We also review the existing multi-modal datasets devoted to such task.

Color methods Background subtraction is one of the most applied techniques when dealing with image segmentation in videos. The parametric model that Stauffer and Grimson (1999) proposed, which models the background using a mixture of gaussians (MoG), has been widely used, and many variations based on it have been suggested. Bouwmans (2011) thoroughly reviewed more advanced statistical background modeling techniques. Nonetheless, after obtaining the moving object contours one still needs a way to assess whether they belong to a human entity. Human detection methods are strongly related to the task of human body segmentation because they allow us to discriminate better among other objects. They usually produce a bounding box that indicates where the person is, which in turn may be 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) (Dalal and Triggs 2005), which is the basis of many current detectors. Used along with a discriminative classifier—e.g. support vector machines (SVM)—it is able to accurately predict the presence of human subjects. Example-based methods (Andriluka et al. 2010) have also been proposed to address human detection, utilizing templates to compare the incoming image and locate the person but limiting the pose variability.

In terms of descriptors, other possible representations, apart from the already commented HOG, are those that try to fit the human body into silhouettes (Mittal et al. 2003), those that model color or texture such as Haar-like wavelets (Viola et al. 2005), optical flow quantized in histograms of optical flow (HOF) (Dalal et al. 2006), and, more recently, descriptors including logical relations, e.g. *Grouplets* (Yao and Fei-Fei 2010), which enable observers to recognize human-object interactions.

Instead of whole body detection, some approaches have been built on the assumption that the human body consists of an ensemble of body parts (Ramanan 2006; Pirsiavash and Ramanan 2012). Some of these are based on pictorial structures (Andriluka et al. 2009; Yang and Ramanan 2011). In particular, Yang and Ramanan (2011), Yang and Ramanan (2013), and Felzenszwalb et al. (2010) outperform other existing methods using a deformable part-based model (DPM). This model consists of a root HOG-like filter and different part-filters that define a score map of an object hypothesis, using latent SVM as a classifier. Another well-known part-based detector is *Poselets* (Bourdev and Malik 2009; Wang et al. 2011), which trains different homonymous parts to fire at a given part of the object at a given pose and viewpoint. More recently, Wang et al. (2013) have proposed *Motionlets* for human motion recognition. Grammar models (Girshick et al. 2011) and AND–OR graphs (Zhu et al. 2008) have been also used in this context.

Other approaches model objects as an ensemble of local features. This category includes methods such as implicit shape models (ISM) (Leibe et al. 2004), which consist of visual words combined with location information. These are also used in works that estimate the class-specific segmentation based on the detection result after a training stage (Leibe et al. 2008).

Conversely, generative classifiers deal directly with the person segmentation problem. They function in a bottom-up manner, learning a model from an 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) (Shi and Malik 2000; Carson et al. 2002). One of the most popular is GrabCut (Rother et al. 2004; Gulshan et al. 2011), an interactive segmentation method based on Graph Cuts (Boykov and Jolly 2001) and conditional random fields (CRF) that combines pixel appearance information with neighborhood relations to refine silhouettes, using a bounding box as an initialization region.

Having considered the properties of each of the aforementioned segmentation categories, it is understandable that a combination of several approaches would be proposed, namely top-down and bottom-up segmentation (Lin et al. 2007; Mori et al. 2004; Ladický et al. 2010; Levin and Weiss 2006; Fidler et al. 2013). To name just a few, ObjCut (Kumar et al. 2005) combines pictorial structures and Markov random fields (MRF) to obtain the final segmentation. PoseCut (Bray et al. 2006) is also based on MRF and Graph Cuts but has the added ability to deal with 3D pose estimation from multiple views.

Depth methods Most of the aforementioned contributions use RGB as the principal cue to extract the corresponding descriptors. The recent release of affordable RGB–depth devices such as Microsoft®Kinect® has encouraged the community to start using depth maps as a new source of information. Shotton et al. (2011) was one of the first contributions, which used depth images to extract the human body pose, an approach that is also the core of the Kinect® human recognition framework.

A number of standard computer vision methods already mentioned for color cues have been applied to depth maps. For example, a combination of Graph Cuts and Random Forest has been applied to part-based human segmentation (Hernández-Vela et al. 2012b). Holt et al. (2011) proposed the use of *Poselets* as a representation that combines part-based and example-based estimation aspects for human pose estimation. Generative models have also been considered, such as in Charles and Everingham (2011), where they are used to learn limb shape models from depth, silhouette and 3D pose data. Active shape models (ASM), Gabor filters (Pugeault and Bowden 2011), template matching, geodesic distances (Schwarz et al. 2011), and linear programming (Windheuser et al. 2011) have also been employed in this context.

Notwithstanding the former, the emergence of the depth modality has lead to the design of novel descriptors. Plagemann et al. (2010), for example, proposed a key-point detector based on the saliency of depth maps for identifying body parts. Point feature histograms, based on the orientations of surface normal vectors and taking advantage of a 3D point cloud representation, have also been proposed for local body shapes representation (Hernández-Vela et al. 2012a). Xia et al. (2011) applied a 2D Chamfer match over silhouettes for human detection and segmentation based on contouring depth images. A more recent contribution is the Histogram of Oriented 4D Normals (HON4D) (Oreifej and Liu 2013), which proposes a histogram that captures the distribution of the surface normal orientations in the 4D space of depth, time, and spatial coordinates. Recently, Lopes et al. (2014) presented a method that describes hand poses by a 3D spherical descriptor of cloud density distributions.

Thermal methods In contrast to color or depth cues, thermal infrared imagery has not been used widely for segmentation purposes, although it is attracting growing interest by the research community. Several specific descriptors have been proposed. For example, HOG and SVM are used in Suard et al. (2006), while Zhang et al. (2007) extended such combination with *Edgelets* and AdaBoost. Other examples include joint shape and appearance cues (Dai et al. 2007), probabilistic models (Bertozzi et al. 2007), shape context descriptor (SCD) with AdaBoost (Wang et al. 2010), and descriptors invariant to scale, brightness and contrast (Olmeda et al. 2012). Background subtraction has also been adapted to deal with this kind of imagery (Davis and Sharma 2004). In that study, the authors presented a statistical contour-based technique that eliminates typical halo artifacts produced by infrared sensors by combining foreground and background gradient information into a contour saliency map in order to find the strongest salient contours. An example of human segmentation is found in Fernández-Caballero et al. (2011), which applies thresholding and shape analysis methods to perform such task.

Most of the cited contributions focus on pedestrian detection applications. Indeed, thermal imaging has attracted the most attention for occupancy analysis (Gade et al. 2013) and pedestrian detection applications, due to the cameras' ability to see without visible illumination and the fact that people cannot be identified in thermal images, which eliminates privacy issues. In addition to these, a key advantage of thermal imaging for detecting people is its discriminative power, due to the big difference in heat intensity where a human is present.

For more, we refer the reader to Gade and Moeslund (2014), an extensive survey of thermal cameras and more applications, including technological aspects and the nature of thermal radiation.

Combining modalities Given the increasing popularity of depth imagery, it is not surprising that a number of algorithms that combine both depth and RGB cues have appeared to benefit from multi-modal data representation (Stefańczyk and Kasprzak 2012; Clapés et al. 2012; Sheasby et al. 2012; Hernández-Vela et al. 2012a; Teichman and Thrun 2013; Scharwächter et al. 2013; Sheasby et al. 2013; Alahari et al. 2013). A recent example is *PoseField* (Vineet et al. 2013), a filter-based mean-field inference method that jointly estimates human segmentation poses, per-pixel body parts, and depth, given stereo pairs of images. Indeed, disparity computation from stereo images is another widely-used approach for obtaining depth maps without range and outdoor limitations. Even background subtraction approaches can profit from such a fusion, since it is possible to reduce those misdetections that cannot be tackled by each modality individually (Gordon et al. 1999; Fernández-Sánchez et al. 2013; Camplani and Salgado 2014; Giordano et al. 2014).

Similar to the RGB–depth combination, thermal imaging has also been fused with color cues to enrich data representation. Such combinations have been applied to pedestrian tracking (Leykin and Hammoud 2006; Leykin et al. 2007), in which the authors apply a codeword-based background subtraction model and a Kalman filter to track pedestrian candidates. The pedestrian classification is handled by a symmetry analysis based on a Double Helical Signature. In Davis and Sharma (2007), Contour Saliency Maps are used to improve a single-Gaussian background subtraction. RGB–thermal human body segmentation is tackled by Zhao and Sen-ching (2012) and, unlike the previously described approaches, the authors' dataset contains objects in close range of the cameras. This means that one cannot rely on a fixed transformation to register the modalities. Instead, the geometric registration is performed at a blob level between visual objects corresponding to human subjects.

Only a few scholars have considered the fusion of RGB, depth, and thermal features (RGB–D–T) to improve detection and classification capabilities. The latest contributions include people following, human tracking, re-identification, and face recognition. Susperregi et al. (2013) used a laser scanner, along with the RGB–D–T sensors, for people detection and people following. The detection is performed separately on each modality and fused on a decision level. Chun and Lee (2013) performed RGB–D–T human motion tracking to determine the 2D position and orientation of people in a constrained, indoor scenario. In Møgelmose et al.

(2013), features extracted on the three modalities are combined to perform person re-identification. More recently, Nikisins et al. (2014) performed RGB–D–T face recognition based on Local Binary Patterns, HOG, and HAAR-features. Irani et al. (2015) provide an interesting approach by using spatiotemporal features and combining the three modalities to estimate pain level from facial images. However, little attention has been paid to human segmentation applications combining such cues.

Existing datasets Up to this point we have extensively reviewed methods related to multi-modal human body segmentation. Such task is often a first step towards further sophisticated pose and behavior analysis approaches. To advance research in this area, it is necessary to have the right means to compare methods so as to measure improvements. There are several static and continuous image-based human-labeled datasets that can be used for that purpose (Moeslund 2011), which try to provide realistic settings and environmental conditions. The best known of these is the Berkeley Segmentation Dataset and Benchmark (Martin et al. 2001), which consists of 12,000 segmented items of 1000 Corel dataset color images containing people or different objects. It also includes figure-ground labelings for a subset of the images. Alpert et al. (2007) also made available a database containing 200 gray level images along with ground-truth segmentations. This dataset was specially designed to avoid potential ambiguities by incorporating only those images that clearly depict one or two objects in the foreground that differ from their surroundings in terms of texture, intensity, or other low level cues. However, the dataset does not represent uncontrolled scenarios. The well known PASCAL Visual Object Classes Challenge (Everingham et al. 2012) 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, and this device is widely used in human pose related works. Gulshan et al. (2011) presented a novel dataset consisting of 3386 images of segmented humans and ground-truth automatically created by Kinect®, which consists of different human subjects across four different locations. Unfortunately, depth map images are not included in the public dataset.

Despite this large body of work, little attention has been given to multi-modal video datasets. We underline the collective datasets of Project ETISEO (Nghiem et al. 2007), owing to the fact that for some of the scenes the authors include an additional imaging modality, such as infrared footage, in addition to color images. It consists of indoor and outdoor scenes of public places such as an airport apron or a subway station, as well as a frame-based annotated ground-truth. Depth maps computed from stereo pairs of images are used in INRIA 3D Movie dataset (Alahari et al. 2013), which contains sequences from 3D movies. Such sequences show people performing a broad variety of activities from a range of orientations and with different levels of occlusions. A comparison of existing multi-modal datasets focused on human body related approaches is provided in Table 1. As one can see, there is a lack of datasets that combine RGB, depth, and thermal modalities focused on the human body segmentation task, like the one we propose in this paper.

3 The RGB–Depth–Thermal Dataset

The proposed dataset features a total of 11,537 frames divided into three indoor scenes, of which 5724 are annotated. Having pictured sample imagery of the three scenes in Fig. 1, we also show their corresponding number of annotated frames and depth range in Table 2. Activity in scene 1 and 3 uses the full depth range of the Kinect® sensor, whereas activity in scene 2 is constrained to a depth range of ±0.250 m in order to suppress the parallax between the two physical sensors. Scenes 1 and 2 are situated in a closed meeting room with little natural light to disturb the sense of depth, while scene 3 is situated in an area with wide windows and a substantial amount of sunlight. The human subjects are walking, reading, using their phones, and, in some cases, interacting with each other. In all scenes, at least one of the humans interacts with a heated object in order to complicate the extraction of humans in the thermal domain. Examples of heated objects in the scene are radiator pipes, boilers, toasters, and mugs.

3.1 Acquisition

The RGB–D–T data stream is recorded using a Microsoft® Kinect® for XBOX360, which captures the RGB and depth image streams, and an AXIS Q1922 thermal camera. The resolution of the imagery is fixed at 640 × 480 pixels. As seen in Fig. 2, the cameras are vertically aligned in order to reduce perspective distortion.

The image streams are captured using custom recording software that invokes the Kinect for Windows® and AXIS Media Control SDKs. The integration of the two SDKs enables the cameras to be calibrated against the same system clock, which enables the post-capture temporal alignment of the image streams. Both cameras are able to record at 30 FPS. However, the dataset is recorded at 15 FPS due to recording software performance constraints.

3.2 Multi-modal Calibration

The calibration of the thermal and RGB cameras was accomplished using a thermal-visible calibration device inspired by Vidas et al. (2012). The calibration device consists of two parts: we use an A3-sized 10 mm polystyrene foam board as a backdrop and a board of the same size with cut-out

Table 1 Comparison of multi-modal datasets aimed for human body related approaches in order of release

Dataset	Data format	Video Seq.	Annotation	Scenario	Purpose
ETISEO Project (Nghiem et al. 2007)	RGB–T	Yes	Bounding box	Indoor/outdoor	Video surveillance
IRIS Thermal/Visible Face Database (Abidi 2007)	RGB–T	No	–	Indoor	Face detection
OSU Color-Thermal Database (Davis and Sharma 2007)	RGB–T	Yes	Bounding box	Outdoor	Object detection
RGB–D People Dataset (Spinello and Arras 2011)	RGB–D	Yes	Bounding box	Indoor	Human detection
H2View Dataset (Sheasby et al. 2012)	RGB–D (stereo)	Yes	Segmentation masks, Ground-truth depth, Human pose	Indoor	3D pose estimation
LIRIS Human Activities Dataset (Wolf et al. 2012)	RGB–D	Yes	Bounding box, Activity class	Indoor	Human activity recognition
RGB–D Person Re-identification Dataset (Barbosa et al. 2012)	RGB–D	Yes	Foreground masks, Skeleton, 3D mesh	Indoor	Person re-identification
VAP RGB–D Face Dataset (Hg et al. 2012)	RGB–D	No	Pose class	Indoor	Face detection, Pose estimation
Biwi Kinect Head Pose Database (Fanelli et al. 2013)	RGB–D	Yes	Head 3D position, Head rotation	Indoor	Head pose estimation
Cornell Activity Datasets (Koppula et al. 2013)	RGB–D	Yes	Bounding box, Activity class, skeleton	Indoor	Human activity recognition
Eurecom Kinect Face Dataset (Huynh et al. 2013)	RGB–D	No	6 facial landmarks, Person information	Indoor	Face recognition
Inria 3D Movie Dataset (Alahari et al. 2013)	RGB–D (stereo)	Yes	Bounding box, Human pose, Segmentation masks	Indoor/outdoor	Human detection, Human segmentation, Pose estimation
RGB–D–T Facial Database (Nikisins et al. 2014)	RGB–D–T	No	Bounding box	Indoor	Face recognition
Our Proposal	RGB–D–T	Yes	Pixel-level	Indoor	Human detection, Human segmentation, Person re-identification

Table 2 Annotated number of frames and spatial constraints of the scenes in meters (m)

Scene	Frames	Annotated frames	Depth range (m)
1	4693	1767	1–4
2	2216	2016	1.4–1.9
3	4628	1941	1–4

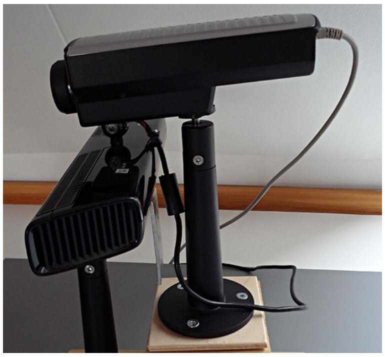

Fig. 2 Camera configuration. The RGB and thermal sensor are vertically aligned

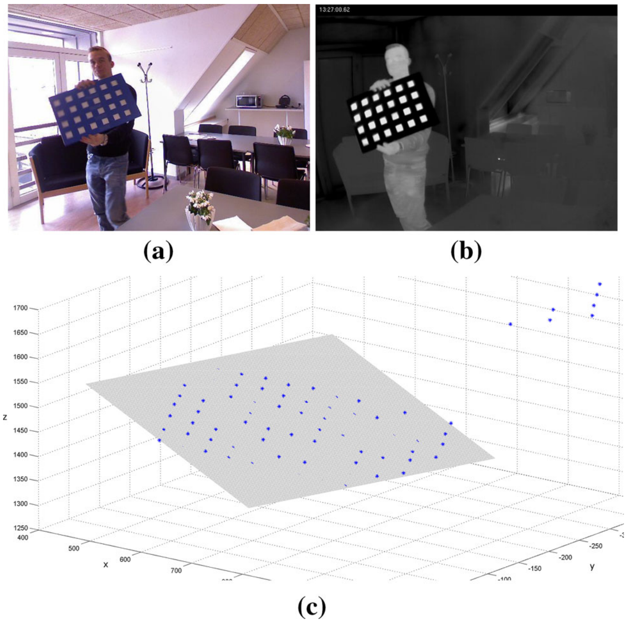

Fig. 3 The calibration device as seen by the (**a**) RGB and (**b**) thermal camera. The corresponding points in world coordinates and the plane, which induces a homography, are overlayed in (**c**). Noise in the depth information accounts for the outliers in (**c**)

squares as the checkerboard. Before using the calibration device, we heat the backdrop and keep the checkerboard plate at room temperature, thus maintaining a suitable thermal contrast when joined, as seen in Fig. 3. Using the Camera Calibration Toolbox of Bouguet (2004), we are able to extract corresponding points in the thermal and RGB modalities. The sets of corresponding points are used to undistort both image streams and for the subsequent registration of the modalities.

3.3 Registration

The depth sensor of the Kinect® is factory registered to the RGB camera and a point-to-point correspondence is obtained from the SDK. The registration is static and might therefore be saved in two look-up-tables for RGB ⇔ depth.

The registration from RGB ⇒ thermal, $\mathbf{x} \Rightarrow \mathbf{x}'$, is handled using a weighted set of multiple homographies based on the approximate distance to the view that the homography represents. By using multiple homographies, we can compensate for parallax at different depths. However, the spatial dependency of the registration implies that no fixed, global registration or look-up-table is possible, thus inducing a unique mapping for each pixel at different depths.

Homographies relating RGB and thermal modalities are generated from a minimum of 50 views of the calibration device scattered throughout each scene. One view of the calibration device induces 96 sets of corresponding points in the RGB and thermal modality (Fig. 3c), from which a homography is computed using a RANSAC-based method. The acquired homography and the registration it establishes are only accurate for points on the plane that are spanned by the particular view of the calibration device. To register an arbitrary point of the scene, $\mathbf{x} \Rightarrow \mathbf{x}'$, the 8 closest homographies are weighted according to this scheme:

1. For all J views of the calibration device, calculate the 3D centre of the K extracted points in the image plane:

$$\overline{\mathbf{X}}_j = \frac{\sum_{k=1}^{K} \mathbf{X}_{k_j}}{K} = \frac{\sum_{k=1}^{K} \mathbf{P}^{+} \mathbf{x}_{k_j}}{K}. \tag{1}$$

   The depth coordinate of $\mathbf{X}$ is estimated from the registered point in the depth image. $\mathbf{P}^{+}$ is the pseudoinverse of the projection matrix.
2. Find the distance from the reprojected point $\mathbf{X}$ to the homography centres:

$$\omega(j) = |\mathbf{X} - \overline{\mathbf{X}}_j|. \tag{2}$$

3. Centre a 3D coordinate system around the reprojected point $\mathbf{X}$ and find $\min \omega(j)$ for each octant of the coordinate system. Set $\omega(j) = 0$ for all other weights. Normalize the weights:

$$\omega^{*}(j) = \frac{\omega(j)}{\sum_{j=1}^{J} \omega(j)}. \tag{3}$$

4. Perform the registration $\mathbf{x} \Rightarrow \mathbf{x}'$ by using a weighted sum of the homographies:

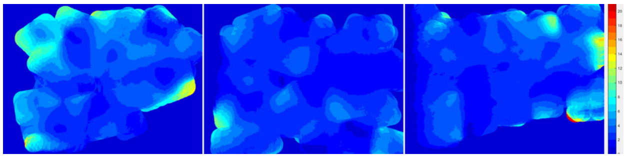

Fig. 4 Average registration error, RGB (**a**) ⇒ thermal (**b**), of the three dataset sequences, averaged over the depth range of the Kinect. The errors are shown in image coordinates and are computed from multiple views of the calibration device. Registrations errors are more prominent in the boundaries of the images

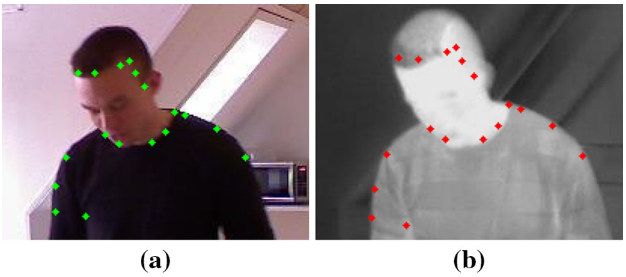

Fig. 5 Example of RGB (**a**) ⇒ thermal (**b**) registration

$$\mathbf{x}' = \sum_{j=1}^{J} \omega^*(j)\, \mathbf{H}_j \mathbf{x}, \tag{4}$$

where $\mathbf{H}_j$ is the homography induced by the j^{th} view of the calibration device.

For registering thermal points, the absence of depth information means that points are reprojected at a fixed distance, inducing parallax for points at different depths. Thus, the registration framework may be written:

$$\text{depth} \Leftrightarrow \text{RGB} \Rightarrow \text{thermal} \tag{5}$$

The accuracy of the registration of RGB ⇒ thermal is mainly dependent on:

1. The distance in space to the nearest homography.
2. The synchronization of thermal and RGB cameras. At 15 FPS, the maximal theoretical temporal misalignment between frames is thus 34 ms.
3. The accuracy of the depth estimate.

A quantitative view of the registration accuracy is provided in Fig. 4. An example of the registration for Scene 3 is seen in Fig. 5.

3.4 Annotation

The acquired videos were manually annotated frame by frame in a custom annotation program called Pixel Annotator. The dataset contains a large number of frames spread over a number of different sequences. All sequences have three modalities: RGB, depth, and thermal. The focus of the annotation is on the people in the scene and a mask-based annotation philosophy was employed. This means that each person is covered by a mask and each mask (person) has a unique ID that is consistent over all frames. In this way the dataset can be used not only for subject segmentation, but also for tracking and re-identification purposes. Since the main purpose of the dataset is segmentation, it was necessary to use a pixel-level annotation scheme. Examples of the annotation and registered annotated masks are shown in Fig. 7.

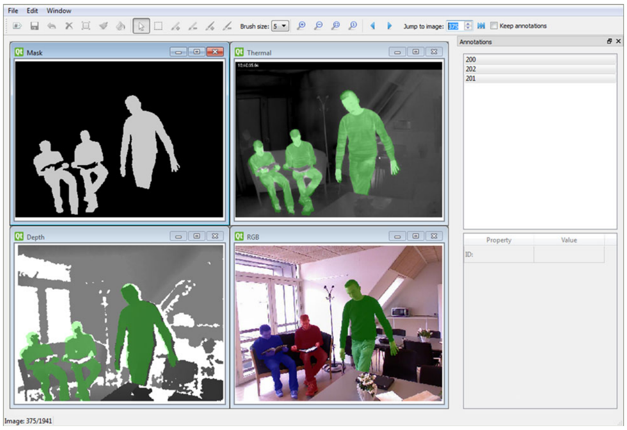

Fig. 6 Pixel Annotator showing the RGB masks and the corresponding, registered masks in the other views

Pixel Annotator provides a view of each modality with the current mask overlaid, as well as a raw view of the mask (see Fig. 6). It implements the registration algorithm described above so that the annotator can judge whether the mask fits in all modalities. Because the modalities are registered to each other, there are not specific masks for any given modality but rather a single mask for all (Fig. 7).

Each annotation can be initialized with an automatic segmentation using the GrabCut algorithm (Rother et al. 2004) to get it quickly off the ground. Pixel Annotator then provides pixel-wise editing functions to further refine the mask. Each annotation is associated with a numerical ID and can have an arbitrary number of property fields associated with it. They can be boolean or contain strings so that advanced annotation can take place, from simple occluded/not occluded fields to fields describing the current activity. Pixel Annotator is written in C++ on the Qt framework and is fully cross-platform compatible.

The dataset and the registration algorithm is freely available at http://www.vap.aau.dk/. Since we subdivided the several scenes into 10 variable-length sequences in order to carry out our baseline experiments, we also provide the parti-

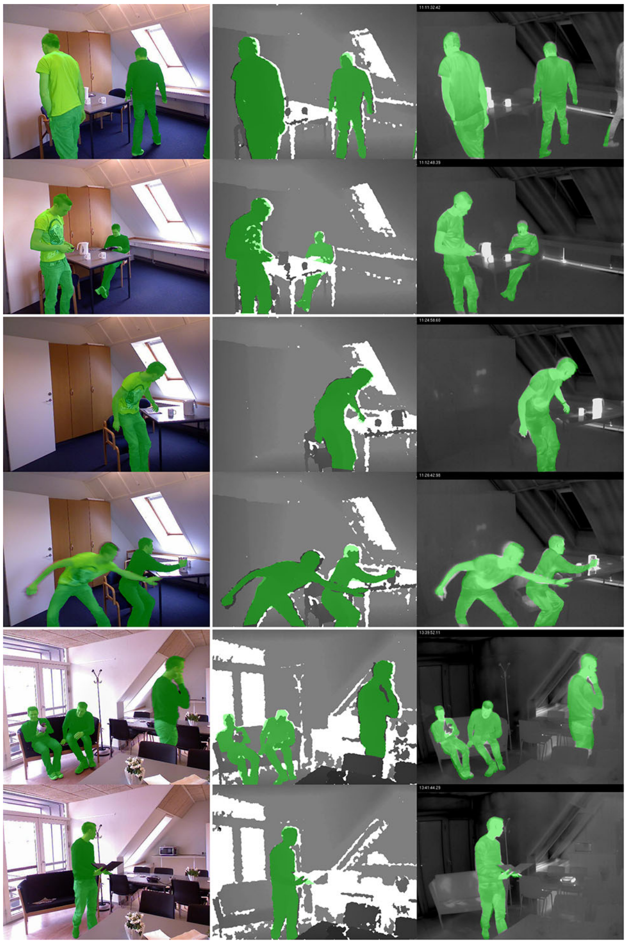

Fig. 7 Examples of the annotated imagery for two views in each of the three scenes. The RGB modality is manually annotated and the corresponding mask is registered to the depth and thermal modalities. The causes of registration misalignment of the masks are motion blur and noisy depth information, which induce parallax in the thermal modality

tionings in a file along with the dataset. We refer the reader to Sect. 5 for more details about the evaluation of the baseline.

4 Multi-modal Human Body Segmentation

We propose a baseline methodology to segment human subjects automatically in multi-modal video sequences. The first step of our method focuses on reducing the spatial search space by estimating the scene background to extract the foreground regions of interest in each one of the modalities. Note that such regions may belong to human or non-human entities, so in order to perform an accurate classification we describe them using modality-specific state-of-the-art feature descriptors. The obtained features are then used to learn probabilistic models in order to predict which foreground regions actually belong to human subjects. Predictions obtained from the different models are then fused using a learning-based approach. Figure 8 depicts the different stages of the method.

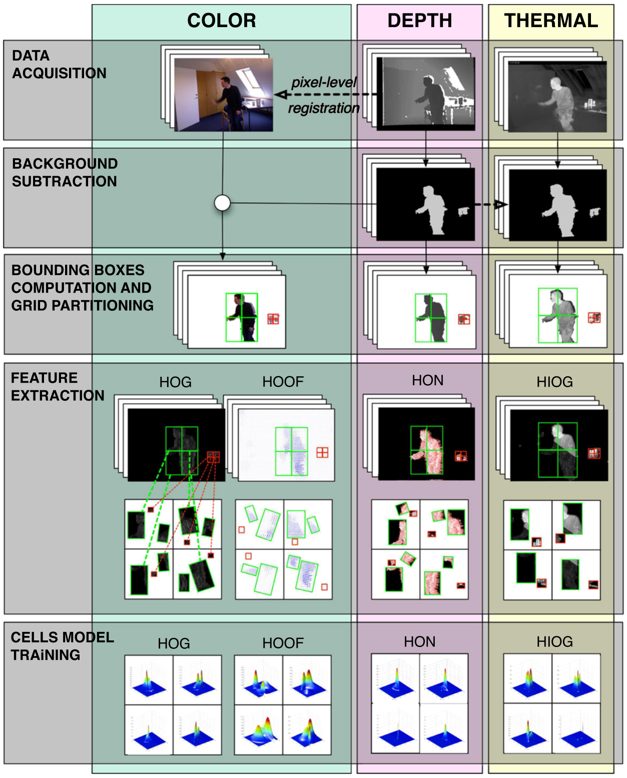

Fig. 8 The main steps of the proposed baseline method, before reaching the fusion step

4.1 Extraction of Masks and Regions of Interest

The first step of our baseline is to reduce the search space. For this task, we learn a model of the background and perform background subtraction.

4.1.1 Background Subtraction

A widely used approach for background modeling in this context is GMM, which assigns a mixture of gaussians per pixel with a fixed number of components (Bouwmans et al. 2008). Sometimes the 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 intruding objects that stand still for a period of time to vanish, so a non-adaptive approach is more convenient in our case.

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 algo-

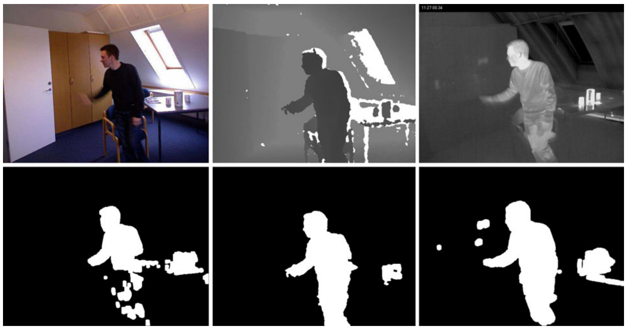

Fig. 9 Background subtraction for different visual modalities of the same scene (RGB, depth, and thermal respectively)

rithms 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, in addition to the inconvenience of temperature changes in objects. A halo effect can also be observed around warm items. Regarding depth-based approaches, they may produce misdetections due to the presence of foreground objects at a depth similar to that of the background. 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 caused by the separation of the infrared emitter and sensor of the Kinect® device. However, they are more robust when it comes to lighting artifacts and shadows. A comparison is shown in Fig. 9, where the actual foreground objects are the humans and the objects on the table. As one can see, RGB fails at extracting the human legs because they are of a similar color to the chair in the back. The thermal cue segments the human body more accurately, but it includes some undesired reflections and illuminates the jar and mugs with a surrounding halo. The pipe tube is also extracted as foreground due to temperature changes over 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, thereby allowing us to use the same masks for both modalities. Let us consider the depth value of a pixel at frame i as $z^{(i)}$. The background model $p(z^{(i)}|B)$ – where B represents the background—is estimated from a training set of depth images represented by $\mathcal{Z}$ using the T first frames of a sequence such that $\mathcal{Z} = \{z_1^{(i)}, \ldots, z_T^{(i)}\}$. This way, the estimated model is denoted by $\hat{p}(z^{(i)}|\mathcal{Z}, B)$, modeled as a mixture of gaussians. We use the method presented in Zivkovic (2004), which uses an on-line clustering algorithm that constantly adapts the number of components of the mixture for each pixel during the learning stage.

4.1.2 Extraction of Regions of Interest

Once the binary foreground masks are obtained, a 2D connected component analysis is performed using basic mathematical morphological operators. We also set 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.

A region of interest should contain a separated person or object. However, different subjects or objects may overlap in space, resulting in a bigger component that contains more than one item. For this reason, each component has to be analyzed to find each item separately in order to obtain the correct bounding boxes that surround them.

One of the advantages of the depth cue is that we can use the depth value in each pixel to know whether an item is farther than another. We can assume that a given connected component denotes just one item if there is no rapid change in the disparity distribution and it has a low standard deviation. For those components that do have a greater standard deviation, and assuming a bimodal distribution—two items in that connected component—, Otsu's method (Otsu 1975) can be used to split the blob in two classes such that their intra-class variance is minimal.

For such purposes, we define $\mathbf{c}$ as a vector containing the depth range values that correspond to a given connected component, with mean $\mu_{\mathbf{c}}$ and standard deviation $\sigma_{\mathbf{c}}$, and σ_{otsu} as a parameter that defines the maximum $\sigma_{\mathbf{c}}$ allowed to not apply Otsu. Note that erroneous or out-of-range 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 $\sigma_{\mathbf{c}} > \sigma_{\text{otsu}}$, Otsu 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. In addition, rectangular bounding boxes are generated encapsulating such items individually over time, whose function is to denote the regions of interest of a given foreground mask.

4.1.3 Correspondence to Other Modalities

As stated in Sect. 4.1.1, depth and color cues use the same foreground masks, so we can take advantage of the same bounding boxes for both modalities. Foreground masks for the thermal modality are computed using the provided registration algorithm with the depth/color foreground masks as input. For each frame, each item is registered individually to the thermal modality and then merged into one mask, thus preserving the same item ID for the depth/color foreground masks. In this way, we achieve a one-to-one straightforward correspondence between items of all modalities, and the constraint of having the same number of items in all the modalities is fulfilled. Bounding boxes are generated in the same way depth modality is, which, although they do not have the same coordinates, denote the same regions of interest. Henceforth, we use R to refer to such regions and $F = \{F^{\text{color}}, F^{\text{depth}}, F^{\text{thermal}}\}$ to refer to the set of foreground masks.

4.1.4 Tagging Regions of Interest

The extracted regions of interest are further analyzed to decide whether they belong to objects or subjects. In order to train and test the models and determine final accuracy results, we need to have a ground-truth labeling of the bounding boxes in addition to the ground-truth masks.

This labeling is done in a semiautomatic manner. First, we extract bounding boxes from regions of interest of ground-truth masks, compare them to those extracted previously from the foreground masks F, and compute the overlap between them. Defining y_r as the label applied to the r region of interest, the automatic labeling is therefore applied as follows:

$$y_r = \begin{cases} 0 \;\; \text{(Object)} & \text{if} \;\; \text{overlap} \leq \lambda_1 \\ -1 \;\text{(Unknown)} & \text{if} \;\; \lambda_1 < \text{overlap} < \lambda_2 \\ 1 \;\; \text{(Subject)} & \text{if} \;\; \text{overlap} \geq \lambda_2 \end{cases} \quad (6)$$

In this way, regions with low overlap are considered to be objects, whereas those with high overlap are classified as subjects. A special category named *unknown* has been added to denote those regions that do not lend themselves to direct classification, such as regions with subjects holding objects, multiple overlapping subjects, and so on.

However, such conditions may not always hold, since some regions whose overlap value is lower than λ_1 compared to the ground-truth masks could actually be part of human beings. For this reason we reviewed the applied labels manually to check for possible mislabelling.

4.2 Grid Partitioning

Given the accuracy of the registration, particularly because of the depth-to-thermal transformation, we are not able to make an exact pixel-to-pixel correspondence. Instead, the association is made among greater information units: grid cells. In the context of this work, a grid cell is the unit of information processed in the feature extraction and classification procedures.

Each region of interest $r \in R$ associated with either a segmented subject or object is partitioned in a grid of $n \times m$ cells. Let G_r denote a grid, which in turn is a set of cells, corresponding to the region of interest r. Hence, we write G_{rij} to refer to the position (i, j) in the r-th region, such that $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$.

Furthermore, a grid cell G_{rij} consists of a set of multichannel images $\{\mathbf{G}_{rij}^{(c)} \mid \forall c \in \mathcal{C}\}$, corresponding to the set of cues $\mathcal{C} = \{$“color”, “motion”, “depth”, “thermal”$\}$. Accordingly, $\{\mathbf{G}_{rij}^{(c)} \mid \forall r \in R\}$, i.e. the set of (i, j)-cells in the c cue, is indicated by $G_{ij}^{(c)}$.

The next section provides the details about the feature extraction processes on the different visual modalities at cell level.

4.3 Feature Extraction

Each cue in $\mathcal{C}$ involves its own specific feature extraction/ description processes. For this purpose, we define the feature extraction function f such that $f\colon \mathbb{R}^{n\times m} \rightarrow \mathbb{R}^{\delta}$. Accordingly, $\mathbf{G} \xrightarrow{\mathbb{R}^{n\times m}} \mathbf{d}$, where $\mathbf{d}$ is a δ-dimensional vector, representing the description of $\mathbf{G}$ in a certain feature space (the output space of f). For the color modality two kinds of descriptions are extracted for each cell—HOG and HOFs—, whereas in the depth and thermal modality the histogram of oriented normals (HON) and histogram of intensities and oriented gradients (HIOG) are used respectively. Hence, we define a set of four different kinds of descriptions $\mathcal{D} = \{\text{HOG}, \text{HOF}, \text{HON}, \text{HIOG}\}$. In this way, for a particular cell G_{rij}, we extract the set of descriptions $D_{rij} = \{f_d(\mathbf{G}_{rij}^{(c)}) \mid c = \varpi(d), \forall d \in \mathcal{D}\} = \{\mathbf{d}_{rij}^{(d)} \mid \forall d \in \mathcal{D}\}$. The function $\varpi(\cdot)$ simply returns the cue corresponding to a given description.

4.3.1 Color Modality

The color imagery is the most popular modality and has been extensively used to extract a range of different feature descriptions.

Histogram of oriented gradients (HOG) For the color cue, we make the most of the original implementation of HOG but with a lower descriptor dimension than the original by not overlapping the HOG blocks. For the gradient computations, we use RGB color space with no gamma correction and the Sobel kernel.

The gradient orientation is therefore determined for each pixel by considering the pixel’s dominant channel and quantized in a histogram over each HOG-cell (note that we are not referring to our cells), evenly spacing orientation values in the range [0°, 180°]. HOG-cells’ histograms in each HOG-block are concatenated and L2-normalized. Finally,

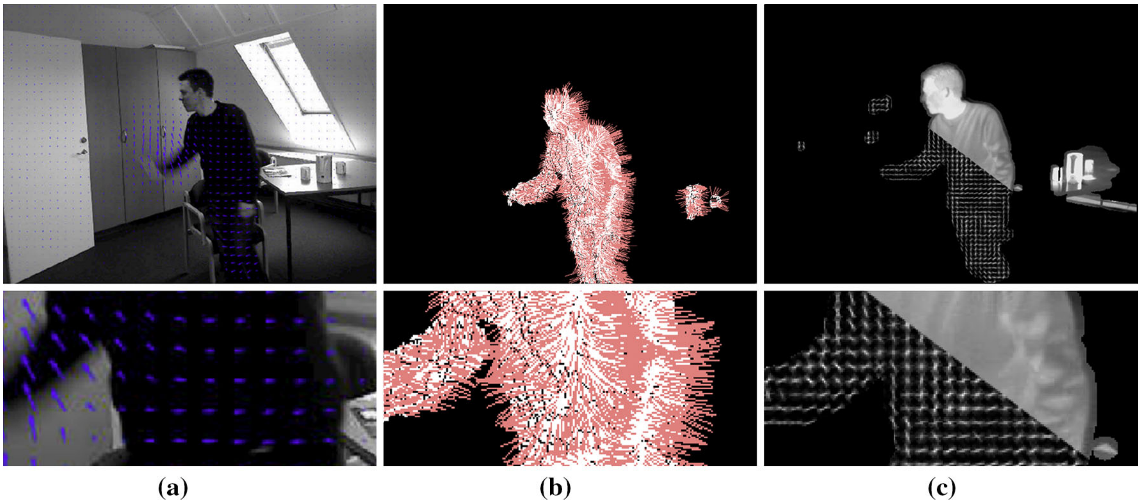

Fig. 10 Example of descriptors computed in a frame for the different modalities: (**a**) represents the motion vectors using a forward scheme; that is, the optical flow orientation gives insight into where the person is going in the next frame; (**b**) the computed surface normal vectors; and (**c**) the thermal intensities and thermal gradients' orientations

normalized HOG-block histograms are concatenated in the κ-bin histogram that we use for our cell classification.

Histogram of Optical Flow (HOF) The color cue also allows us to obtain motion information by computing the dense optical flow and describing the distribution of the resultant vectors. The optical-flow vectors of the whole image can be computed using the luminance information of image pairs with the Gunnar Farnebäck's algorithm (Farnebäck 2003). In particular, we use the available implementation in OpenCV,[1] which is based on modeling the neighborhoods of each pixel of two consecutive frames by quadratic polynomials. This implementation allows a wide range of parameterizations, which are specified in Sect. 5.

The resulting motion vectors, which are shown in Fig. 10, are masked and quantized to produce weighted votes for local motion based on their magnitude, taking into account only those motion vectors that fall inside the G^{color} grids. Such votes are locally accumulated into a ν-bin histogram over each grid cell according to the signed (0°–360°) vector orientations. In contrast to HOG, HOF uses signed optical flow since the orientation information provides more discriminative power.

4.3.2 Depth Modality

The grid cells in the depth modality G^{depth} are depth dense maps represented as planar images of pixels that measure depth values in millimeters. From this depth representation (projective coordinates) it is possible to obtain the "real world" coordinates by using the intrinsic parameters of the depth sensor. This new representation, which can be seen as a 3D point cloud structure $\mathcal{P}$, offers the possibility of measuring actual euclidean distances – those that can be measured in the real world.

After completing the former conversion, we propose to compute the surface normals for each particular point cloud $\mathcal{P}_{rij}$ (representing an arbitrary grid cell $\mathbf{G}_{rij}^{\text{depth}}$) and their distribution of angles summarized in a δ-bin histogram that describes the cell from the depth modality point of view.

Histogram of oriented depth normals (HON) In order to describe an arbitrary point cloud $\mathcal{P}_{rij}$, the surface normal vector for each 3D point must be computed first. The normal 3D vector at a given point $\mathbf{p} = (p_x, p_y, p_z) \in \mathcal{P}$ can be seen as a problem of determining the normal of a 3D 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:

$$\mathbf{q} \triangleq \bar{\mathbf{p}} = \frac{1}{|\mathcal{K}|} \sum_{\mathbf{p} \in \mathcal{K}} \mathbf{p}. \tag{7}$$

The solution of $\mathbf{n}$ can be then approximated as the smallest eigenvector of the covariance matrix $C \in \mathbb{R}^{3\times 3}$ of the points in $\mathcal{P}_{\mathbf{p}}^{\mathcal{K}}$.

The sign of $\mathbf{n}$ can be either positive or negative, and it cannot be disambiguated from the calculations. We adopt the convention of consistently re-orienting all computed normal vectors towards the depth sensor's viewpoint direction $\mathbf{z}$. Moreover, a neighborhood radius parameter determines the cardinality of $\mathcal{K}$, i.e. the number of points used to compute the normal vector in each of the points in $\mathcal{P}$. The computed normal vectors over a human body region is shown in Fig. 10. Points are illustrated in white, whereas normal vectors are red

[1] This is an implementation of the work of Bradski and Kaehler (2008), which can be found at http://code.opencv.org.

lines (instead of arrows to ease the visualization). The next step is to build the histogram describing the distribution of the normal vectors' orientations.

A normal vector is expressed in spherical coordinates using three parameters: the radius, the inclination θ, and the azimuth φ. In our case, the radius is a constant value, so this parameter can be omitted. Regarding θ and φ, the cartesian-to-spherical coordinate transformation is calculated as:

$$\theta = \arctan\left(\frac{n_z}{n_y}\right), \quad \varphi = \arccos\frac{\sqrt{(n_y^2 + n_z^2)}}{n_x}. \tag{8}$$

Therefore, a 3D normal vector can be characterized by a pair (θ, φ) and the depth description of a cell consists of a pair of δ_θ-bin and δ_φ-bin histograms (such that $\delta = \delta_\theta + \delta_\varphi$), L1-normalized and concatenated, describing the two angular distributions of the body surface normals within the cell.

4.3.3 Thermal Modality

Whereas neither raw values of color intensity nor depth values of a pixel provide especially meaningful information for the human detection task, raw values of thermal intensity on their own are much more informative.

Histogram of thermal intensities and oriented gradients (HIOG) The descriptor obtained from a cell in the thermal cue $\mathbf{G}_{rij}^{\text{thermal}}$ is the concatenation of two histograms. The first one is a histogram summarizing the thermal intensities, which spread across the interval $[0, 255]$. The second histogram summarizes the orientations of thermal gradients. Such gradients, computed by convolving a first derivative kernel in both directions, are binned in a histogram weighted by their magnitude. Finally, the two histograms are L1-normalized and concatenated. We used α_i bins for the intensities and α_g bins for the gradients' orientations.

4.4 Uni-modal (Description-Level) Classification

Since we wish to segment human body regions, we need to distinguish those from the other foreground regions segmented by the background subtraction algorithm. One way to tackle this task is from an uni-modal perspective.

From the previous step, each grid cell has been described using each and every description in $\mathcal{D}$. For the purpose of classification, we train a GMM for every cell (i, j) and description in $\mathcal{D}$. For a particular description d, we thereby obtain the set of GMM models $\mathcal{M}^{(d)} = \{\mathcal{M}_{ij}^{(d)} \mid \forall i \in \{1, \ldots, n\}, \forall j \in \{1, \ldots, m\}\}$.

For predicting a new unseen region r to be either a subject or an object according to d, it is first partitioned into G_r, the cells' contents $\{\mathbf{G}_{rij}^{\varpi(d)}\}_{\forall i,j}$ are described, and the $n \times m$ feature vectors representing the region in the d-space, $\{\mathbf{d}_{rij}^{(d)}\}_{\forall i,j}$, are evaluated in the corresponding mixtures' PDFs. The log-likelihood value associated with the (i, j)-th feature vector, $\mathbf{d}_{rij}^{(d)}$, is thus the one in the most probable component in the mixture $\mathcal{M}_{ij}^{(d)}$. Formally, we denote this log-likelihood value as $\ell_{rij}^{(d)}$. Eventually, the category – either subject or object – of the (i, j) cell according to d can be predicted by comparing the standardized log-likelihood $\hat{\ell}_{rij}^{(d)}$ with an experimentally selected threshold value $\tau_{ij}^{(d)}$.

However, given that we can have a different category prediction for each cell, we first need to reach a consensus among cells. In order to do this, we convert the standardized log-likelihoods to confidence-like terms. This transformation consists of centering $\{\hat{\ell}_{rij}^{(d)} \mid \forall r \in R\}$ to $\tau_{ij}^{(d)}$ and scaling the centered values by a deviation-like term that is simply the mean squared difference in the sample with respect to $\tau_{ij}^{(d)}$. This way, we eventually come up with the confidence-like terms $\{\varrho_{rij}^{(d)} \mid \forall r \in R\}$ that conveniently differ in their sign depending on the category label: a negative sign for objects and a positive one for subjects; thus, the more negative (or positive) the value is, the more confidently we can categorize it as an object (or a subject).

Finally, the consensus among the cells of a certain region r can be attained by a voting scheme. For this purpose, we define the grid consensus function $g(r; d)$ as follows:

$$v_r^{(d,-)} = \sum_{i,j} \mathbb{1}\{\varrho_{rij}^{(d)} < 0\}, \quad v_r^{(d,+)} = \sum_{i,j} \mathbb{1}\{\varrho_{rij}^{(d)} > 0\} \tag{9}$$

$$\bar{\varrho}_r^{(d,-)} = \frac{1}{v_r^{(d,-)}} \sum_{(i,j) \mid \varrho_{rij}^{(d)} < 0} \varrho_{rij}^{(d)}, \tag{10}$$

$$\bar{\varrho}_r^{(d,+)} = \frac{1}{v_r^{(d,+)}} \sum_{(i,j) \mid \varrho_{rij}^{(d)} > 0} \varrho_{rij}^{(d)} \tag{11}$$

$$g(r; d) = \begin{cases} 0 & \text{if } v_r^{(d,-)} > v_r^{(d,+)} \\ \mathbb{1}\left\{|\bar{\varrho}_r^{(d,-)}| < |\bar{\varrho}_r^{(d,+)}|\right\} & \text{if } v_r^{(d,-)} = v_r^{(d,+)} \\ 1 & \text{if } v_r^{(d,-)} < v_r^{(d,+)} \end{cases}, \tag{12}$$

where $v_r^{(d,-)}$ and $v_r^{(d,+)}$ keep count of the votes of the r grid cells for object (negative confidence) and subject (positive confidence), respectively. $\bar{\varrho}_r^{(d,-)}$ and $\bar{\varrho}_r^{(d,+)}$ are the averages of negative and positive confidences, respectively. In the case of a draw, the magnitude of the mean confidences obtained for both categories are compared. Since confidence values ϱ are centered at the decision threshold τ, these can be interpreted as a margin distance. From these calculations, the cells' decisions can be aggregated and the category of a grid r determined from each of the descriptions' point of view.

4.5 Multi-modal Fusion

Our hypothesis is that the fusion of different modalities and descriptors, potentially providing a more informative and richer representation of the scenario, can improve the final segmentation result.

4.5.1 Learning-based Fusion Approach

As before, the category of a grid r should be predicted. However, instead of just relying on individual descriptions, we exploit the confidences ϱ provided by the GMMs in the different cells and types of description altogether. This approach follows the Stacked Learning scheme (Cohen 2005; Puertas et al. 2013), which involves training a new learning algorithm by combining previous predictions obtained with other learning algorithms. More precisely, each grid r is represented by a vector $\mathbf{v}_r$ of confidences:

$$\mathbf{v}_r = (\varrho^{(1)}_{r11}, \ldots, \varrho^{(1)}_{rNM}, \ldots, \varrho^{(|\mathcal{D}|)}_{r11}, \ldots, \varrho^{(|\mathcal{D}|)}_{rNM}, y_r), \qquad (13)$$

where y_r is the actual category of the r grid. Using such representation of the confidences in the different grid cells and modalities, we build a data sample containing the R feature vectors of this kind. In this way, any supervised learning algorithm can be used to learn from these data and infer more reliable predictions than using individual descriptions and defined voting scheme for cells' consensus. For this purpose, we use a Random Forest classifier (Breiman 2001) after an experimental evaluation of different state-of-the-art classifiers.

5 Evaluation

We test our approach in the novel RGB–D–T dataset and compare it to other state-of-the-art approaches. First we detail the experimental methodology and evaluation parameters and then provide the experiments' results and a discussion about them.

5.1 Experimental Methodology and Validation Measures

We divided the dataset into 10 continuous sequences, as listed in Table 3, and performed a leave-one-sequence-out cross-validation so as to compute the out-of-sample segmentation overlap. The unequal length of the sequences stems from the posture variability criterion followed: to ensure that very similar postures are not repeated in the different folds (i.e. sequences).

In addition, we performed a model selection in each training partition in order to find the optimal values for the GMMs' experimental parameters: k (number of components in the mixture), τ (decision threshold), and ϵ (stopping criterion for fitting the mixtures). We provide more detailed information about their values in Sect. 5.2. Although we used the leave-one-sequence-out cross-validation strategy again, we applied it this time to the remaining $N-1$ training sequences. In each inner fold, a grid search was carried out to measure the performance of each combination (k, τ, ϵ). The optimal combination, i.e., the one that showed the best average across the 10×9 model selections, was used to train the final model eventually validated in the corresponding test sequence.

Table 3 Division of the scenes into 10 sequences (or partitions) of different length

Sequence id.	Scene id.	No. frames	Start–end frame
1	1	134	00001–00134
2		905	00135–01638
3		762	01639–02400
4	2	247	00001–00247
5		816	00248–01063
6		463	01064–01526
7		690	01527–02216
8	3	142	00001–00142
9		848	00143–01449
10		951	01450–02400

The parameters of the supervised classifiers in the learning-based fusions were selected following the same validation procedure as above but considered the vectors of stacked confidences instead of the original descriptors. While the selection of k, τ, and ϵ was sufficiently exhaustive, given their nature, the parameters involved in these supervised learning algorithms often require more exhaustive searches to fine-tune their values. In order to find the best parameters while keeping the number of combinations manageable, we performed a two-level grid search, which consisted of a first coarse grid search followed by a second narrow grid search around the coarse optimal values.

As previously mentioned, we computed an overlap measure in order to evaluate the performance of our baseline. The overlap was first computed per person-ID and frame, and then averaged across all IDs in that frame. For the computation, we used intersection-over-union $\frac{|A \cap B|}{|A \cup B|}$, where A is a ground-truth region with a certain person-ID and B the region of prediction with its pixels coinciding with those of A. Having computed the overlaps at frame-level, the overlap of a sequence is thereby calculated as the mean overlap of all those frames containing at least one blob, whether it be in the ground-truth or in the prediction mask.

As stated in Sect. 4.1.1, 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. To tackle this problem, a *do not care region* (DCR) is often used. A DCR simply defines a border region of pixels over the silhouette contours in both the prediction and contour masks that are not taken into account for the overlap computation. In this way, we can compare the effect of using a growing DCR to the actual overlap.

5.2 Parameters and Settings

We experimentally set $\lambda_1 = 0.1$ and $\lambda_2 = 0.6$ for the automatic tagging of regions of interest. We also set $\sigma_{\text{otsu}} = 8.3$ for a connected component area of at least 0.1 % of the image and $\sigma_{\text{otsu}} = 12$ for other cases. These settings were established in order to maintain a trade-off between finding the maximum number of overlapping people situations without dividing a subject in different regions, depending on the variation of depth of the body parts.

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, with $m = 2$ and $n = 2$.

For the HOG descriptor, each grid cell was resized to 64×128 pixels and divided in rectangular blocks of 32×32 pixels, which were, in turn, divided into rectangular local spatial regions of 16×16 pixels. We also set $\kappa = 9$. The information of each local spatial region is concatenated, resulting in a vector of 36 values per HOG-block. This brings the final vector size of a grid cell to 4 HOG-blocks vertically × 2 HOG-blocks horizontally × 4 HOG-cells per block × 9 bins per HOG-cell, making a total of 288 components/dimensions. To further reduce the vector length and avoid the curse of dimensionality, we applied PCA to such vector, retaining 95 % of the information. This way, the number of components of the feature vectors from all descriptions do not differ greatly.

In order to compute optical flow, we fixed the parameters of the given implementation based on the best-performing ones from the tests performed in Brkić et al. (2013). Specifically, we set the average window size to 2, the size of the pixel neighborhood considered when finding polynomial expansion in each pixel 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 values. For the motion descriptor (HOF), we defined $\nu = 8$ to produce an 8-dimensional feature vector.

For the depth descriptors (HON), we defined $\delta_\theta = 8$ and $\delta_\varphi = 8$, whereas for the thermal descriptors (HIOG), we defined $\upsilon_{\text{i}} = 8$ and $\upsilon_{\text{g}} = 8$, as they are standard values often used in the literature.

In the GMM-related experiments, we set $k = \{2, 4, 6, 8, 10, 12\}$ and $\tau = \{-3, -2.5, -2, -1.5, -1.25, -1, -0.75, -0.5, -0.4, \ldots, 0.5, 0.75, 1, 1.25, 1.5, 2, 2.5, 3\}$. In order to avoid overfitting problems, we also optimized the termination criterion of the Expectation-Maximization algorithm used for training the GMMs, $\epsilon = \{1e{-}2, 1e{-}3, 1e{-}4, 1e{-}5\}$.

Among many existing state-of-the-art supervised learning algorithms able to perform the fusion, we tested the following: Adaptive Boosting, Multi-Layer Perceptron (with both sigmoidal and radial basis activation functions), Support Vector Machines (linear and radial basis function kernels), and Random Forest. In the AdaBoost experiment, we selected the number of possible weak classifiers and the weight trimming rates among $\{10, 20, 50, 100, 200, 500, 1000\}$ and $\{0, 0.7, 0.75, 0.8, \ldots, 1\}$, respectively; in the MLP, we chose the number of neurons of the hidden layer among $\{2, 5, 10, 15, \ldots, 50, 60, 70, \ldots, 100\}$; in the SVM, we tested the regularization and the gamma parameters within $\{1e - 7, 1e - 6, \ldots, 1e4\}$ and in $\{1e - 7, 1e - 6, \ldots, 1e2\}$; and finally, in the RF we selected the maximum depth of the trees from $\{2, 4, 8, 16, 32, 64\}$, the maximum number of trees from $\{1, 2, 4, 8, 16, 32, 64, 128\}$, and the proportion of random variables to consider in node split from $\{0.05, 0.1, 0.2, 0.4, 0.8, 1\}$.

Regarding the DCR size, we tested several values (number of pixels) in the interval $[2 \cdot 0 + 1, \ldots, 2 \cdot 8 + 1]$.

In addition, and to better capture the posture variability, we augmented the training data by including the mirrored versions of the regions of interest along the vertical axis, as well as the original ones. Nonetheless, at the test stage, we considered only original regions of interest.

5.3 Experiments

In this section, we illustrate the performance of our baseline in terms of overlap after carrying out an extensive experiment. First, we illustrate the performance of the different descriptions (HOG, HOF, HON, and HIOG). Second, we compare the best description to the learning-based fusions. Third, we show the performance of the baseline in the different sequences (test partitions). Fourth, we compare the evaluation of the baseline using the color/depth ground-truth masks vs. the thermal ones. And fifth, we compare our baseline to two standard techniques of the state of the art performing segmentation in the different modalities. In all cases we measure the overlap in function of the DCR size and compare it to color/depth ground-truth masks, unless otherwise stated.

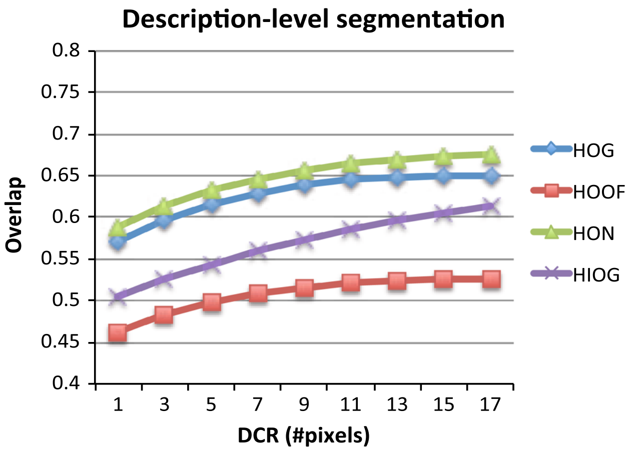

Fig. 11 Results obtained from the different individual descriptions (HOG, HOF, HON, and HIOG) in terms of overlap

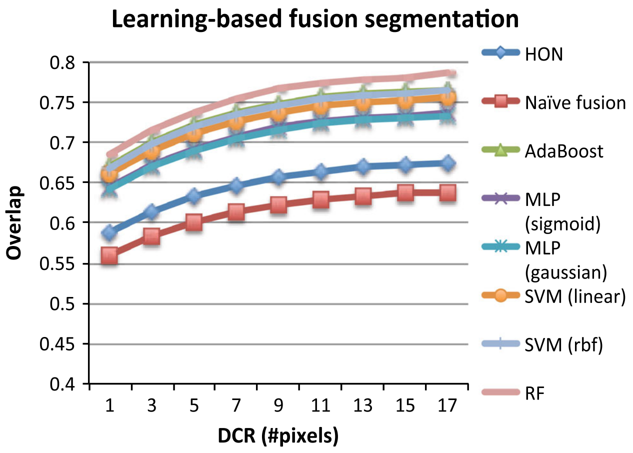

Fig. 12 Results obtained from the best individual descriptions (HON), a naive fusion, and different learning-based fusions, in terms of overlap

5.3.1 Experiment: HOG, HOF, HON, and HIOG Descriptions

We evaluated the performance of the proposed descriptions (HOG, HOF, HON, and HIOG) when predicting on their own. The overlap results shown in Fig. 11, where each descriptor overlap index is computed with respect to their specific modality ground-truth masks, demonstrate the superior performance of the HON descriptor computed in the depth modality, which reach 67.5 % of overlap and improve by 14 % (on average for the different DCR sizes) the results of the worst performing description. The HOG description in the color modality came in a close second (65 %), achieving 2.5 % less overlap than HON (in average). The worst results were obtained by the motion cue in this case, probably because they were uninformative when dealing with static postures, which are abundant in our data. Despite this, it is able to segment people while achieving more than 50 % of such a pessimistic measure as overlap. Note, also, the different upward trend of HIOG in the thermal modality. We discuss this phenomenon, which is due to the color-to-thermal registration, in Sect. 5.4.

5.3.2 Experiment: Learning-based Fusion

In the second experiment, we compared the learning-based fusion with different classifiers against both the best performing description (HON) and a naive fusion we designed in order to give more credit to the better performance of the learning-based fusions. The naive fusion simply averages the cells confidences along the different modalities and then aggregates the averaged cell confidences as described in Sect. 4.4.

Figure 12 shows that the best performing method was the Random Forest classifier (up to 78.6 % of overlap), which thus became our choice for the baseline. This supposed an improvement over HON of 10 % (on average). On the other hand, the worst performing fusion (MLP with gaussian activation function) also presented an improvement over HON, but only of 5 % (on average).

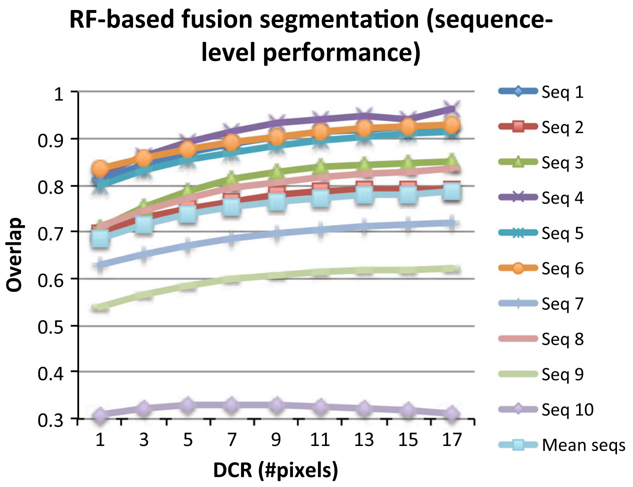

Fig. 13 Results obtained from the RF-based fusion (the best learning-based fusion) in terms of overlap for the different sequences

The naive fusion resulted in an overlap of 63.9 %, which was substantially lower than both HON and HOG.

Once the best classifier for the learning-based fusion was determined, we measured separately the performance of our baseline on the different sequences. Figure 13 depicts the performance in the sequences. Notice that there is a large difference in performance across the evaluated sequences. Four of them—*Seq.1*, *Seq.4*, *Seq.5*, and *Seq.6*—exhibit saturation on the improvement of performance around 90 % at DCR of 11–13 pixels. Four others—*Seq.2*, *Seq.3*, *Seq.7*, and *Seq.8*—are closer to the mean performance *Mean seqs*. And two of them—*Seq.9* and *Seq.10*—suffer a more severe

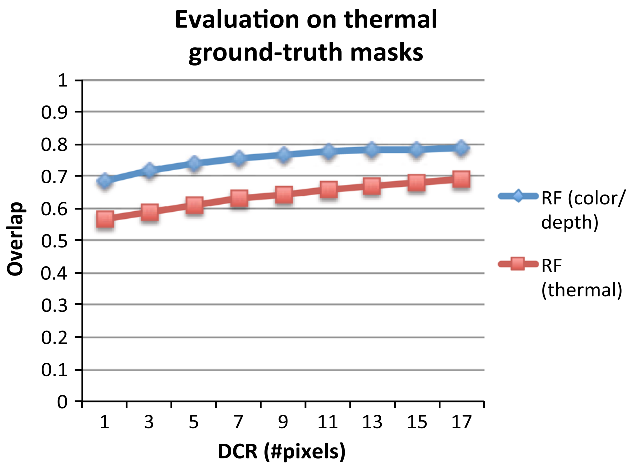

Fig. 14 Comparison of performance measuring the overlap in the thermal registered masks against the manually annotated masks from color/depth

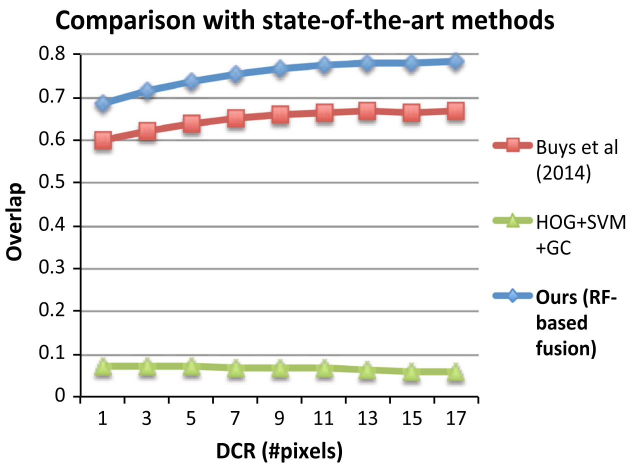

Fig. 15 Comparison of our baseline (using RF-based fusion) with other state-of-the-art approaches that perform human body segmentation from color imagery (HOG + SVM + GC) and depth maps (Buys et al. 2014)

drop in performance, especially *Seq.10*. We discuss plausible reasons for this further on in the paper.

5.3.3 Experiment: Evaluation on Thermal Ground-Truth Masks

In addition, we measured the performance of our most successful approach on the thermal masks in order to quantitatively measure the decrease in performance caused by the misalignment in the thermal-to-color registration. Figure 14 reveals a relatively small decrease in performance. This fact somehow justifies the slightly poorer performance of HIOG in respect to HON and HOG, as previously depicted in Sect. 5.3.1, and why any thermal-related descriptors would pay a price when evaluated in the thermal ground-truth.

5.3.4 Experiment: Comparison to State-of-the-Art Approaches

Since there is no approach that uses the three modalities for human body segmentation, we compared our baseline with two successful state-of-the-art approaches for such task performing in either the color or the depth cue.

One was the work of Buys et al. (2014), which performs solely on the depth modality. This work, based on that of Shotton et al. (2011), describes depth pixels by a set of depth-invariant features generated from the normalized depth differences at pairs of random offsets in respect to the evaluated pixel. From this description, a Random Forest classifier is able to classify each pixel as a body part. In our experiments, we used the open-source implementation made available as part of the Point Cloud Library[2] along with a set of pre-trained trees.[3] In this way we were able to ensure that the method was not relying on tracking techniques—for a fairer comparison to our approach—as would have been the case with the implementation of Shotton et al. (2011) found in the Kinect SDK.[4] Furthermore, we took advantage of the extracted foreground masks from Sect. 4.1.1 in order to apply the body part detector only to foreground pixels; this way, we avoided the apparition of false body part detections all around the scene.

We also compared our approach with that of HOG + SVM + GC (GrabCut) for people segmentation in the color modality. We used the OpenCV available implementations, which are based on the original algorithms (Dalal and Triggs 2005; Rother et al. 2004). The HOG + SVM combination, in particular, detects people as bounding boxes, and the inner dense silhouettes are then segmented by means of GC. The latter is applied in an automatic fashion, learning the GMMs of 70 % of the bounding box as *Probably Foreground* and the rest as *Probably Background*.

Both approaches were trained in independent but larger datasets that ensured more variation than if they had been trained in our dataset. As shown in Fig. 15, our approach outperformed the other baselines when applied to our dataset.

Our baseline largely improved the HOG + SVM + GC approach. However, Buys et al. (2014) achieved a result comparable to ours, with a maximum overlap of 67.1 %. Despite that, our approach also improved this one by more than 10 %.

[2] http://pointclouds.org/documentation/tutorials/gpu_people.php.

[3] https://github.com/PointCloudLibrary/data/tree/master/people/results.

[4] Shotton et al. (2011) specified in the “Acknowledgements” section that the tracking system of Kinect SDK was built based on the research they presented in the paper.

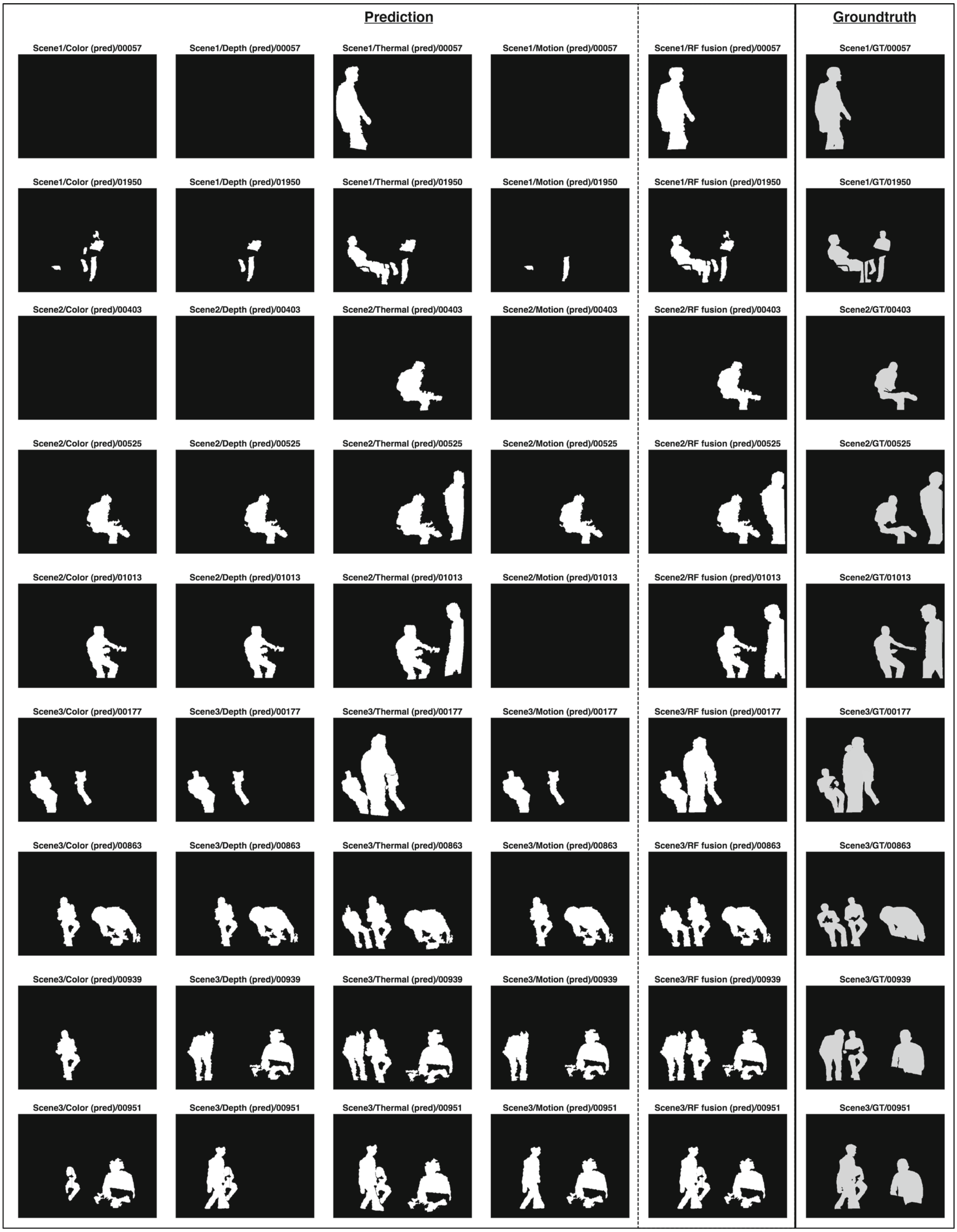
Prediction
Groundtruth
Scene1/Color (pred)/00057
Scene1/Depth (pred)/00057
Scene1/Thermal (pred)/00057
Scene1/Motion (pred)/00057
Scene1/RF fusion (pred)/00057
Scene1/GT/00057
Scene1/Color (pred)/01950
Scene1/Depth (pred)/01950
Scene1/Thermal (pred)/01950
Scene1/Motion (pred)/01950
Scene1/RF fusion (pred)/01950
Scene1/GT/01950
Scene2/Color (pred)/00403
Scene2/Depth (pred)/00403
Scene2/Thermal (pred)/00403
Scene2/Motion (pred)/00403
Scene2/RF fusion (pred)/00403
Scene2/GT/00403
Scene2/Color (pred)/00525
Scene2/Depth (pred)/00525
Scene2/Thermal (pred)/00525
Scene2/Motion (pred)/00525
Scene2/RF fusion (pred)/00525
Scene2/GT/00525
Scene2/Color (pred)/01013
Scene2/Depth (pred)/01013
Scene2/Thermal (pred)/01013
Scene2/Motion (pred)/01013
Scene2/RF fusion (pred)/01013
Scene2/GT/01013
Scene3/Color (pred)/00177
Scene3/Depth (pred)/00177
Scene3/Thermal (pred)/00177
Scene3/Motion (pred)/00177
Scene3/RF fusion (pred)/00177
Scene3/GT/00177
Scene3/Color (pred)/00863
Scene3/Depth (pred)/00863
Scene3/Thermal (pred)/00863
Scene3/Motion (pred)/00863
Scene3/RF fusion (pred)/00863
Scene3/GT/00863
Scene3/Color (pred)/00939
Scene3/Depth (pred)/00939
Scene3/Thermal (pred)/00939
Scene3/Motion (pred)/00939
Scene3/RF fusion (pred)/00939
Scene3/GT/00939
Scene3/Color (pred)/00951
Scene3/Depth (pred)/00951
Scene3/Thermal (pred)/00951
Scene3/Motion (pred)/00951
Scene3/RF fusion (pred)/00951
Scene3/GT/00951

Fig. 16 Qualitative results illustrating the importance of the thermal cue, with each row representing a frame. For each frame, we show the human prediction masks obtained from the different descriptions separately, in addition to the prediction from the fusion approach using a Random Forest classifier. From left to right, the predictions using: Color (HOG), Depth (HON), Thermal (HIOG), Motion (HOF), and RF-based fusion. The last column corresponds to the segmentation ground-truth mask. On top of each binary image, we indicate "sequence name"/"modality name" (or GT if ground-truth)/"frame ID"

5.4 Discussion

The results we obtained showed that 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.

Among the modalities included in our approach, we considered the thermal modality to be of great importance. One cannot guarantee human presence just because of large thermal intensity readings, since many non-human entities such as animals or unanimated objects can emit a considerable amount of heat. However, relatively low thermal intensities are, indeed, highly likely to imply the absence of human presence. This leads, in our case, to the classification of that region as a background category. Hence, in the context of human-background classification, we can consider this “human heat” prior a valuable piece of information that, used together with the thermal gradients and later fused with other cues, enhances the overall performance of our method. In Fig. 16, we illustrated some situations in which the thermal contribution was of great importance to a proper segmentation. Nonetheless, we found the use of the modalities altogether to be very important for the segmentation task.

The set of simple yet reliable descriptions extracted from the multiple cues produced errors somehow uncorrelated. This could be seen in the qualitative results.[5] Our initial assumption was that the learning-based fusion should be able to take advantage of this lack of correlation and thus improve individual results. The quantitative results illustrated in Sect. 5.3.2 confirmed the validity of our initial assumption. The RF-based fusion, in particular, improved the individual descriptions by 25 % on average when compared to HOF (the worst description) and 10 % when comparing to HON (the best description). Moreover, the importance of the learning process in the fusion step was also assessed comparing the results of the learning-based approach to a more naive fusion of confidences.

The selection of the best classifier also proved to be crucial, doubling the improvement of performance with respect to HON when choosing RF over a MLP with gaussian activation function (from 5 to 10 %). In fact, a SVM classifier with linear kernel performed surprisingly well, demonstrating the stacked vectors of confidences to be linearly separable features. Yet the RF classifier increased the overlap results 2.5 % (on average) with respect to the linear SVM, showing that there was still room for improvement.

We also studied the performance of each of the sequences. In 7 out of 10 sequences, results were above the mean. The poor performance in one of them, *Seq. 10*, reduced the *Mean seqs* overlaps by almost 5 % (on average). After checking the predicted masks, we noticed a false positive on a chair’s back region, which appeared quite static during the whole sequence and was a relatively big image region—because it was close to the camera. The difficulty level of this sequence can be better seen qualitatively in the last two rows of Fig. 1. As mentioned before, this scenario contains wide windows with a large amount of sunlight, which may disturb the depth data. Moreover, the color of the subject’s jumper is extremely similar to the color of the couch, making it difficult for the color modality. Another interesting effect is the heat mark that the subject bodies left on the couch in the thermal modality, which may be mistaken for a real subject.

Accurate pixel-level segmentation is a complex task in state-of-the-art techniques. In these scenarios, a DCR is often considered. In our case, experiments showed marginal improvements for DCR sizes greater than 11 pixels, except for the case of thermal modality, which exhibited a particular upward trend. It is important to note that thermal descriptions cannot reach overlap values as good as the other descriptions. The reason for this is that the binary masks F^{thermal} were created from F^{depth} using the registration algorithm, which cannot be accurate up to pixel level, in such a way that the ground-truth and registered masks differ slightly, especially on the left and right sides of the image. As one can observe, this misalignment caused by the registration algorithm introduced an additional error to the depth’s halo effect, which kept being palliated with the biggest DCR sizes.

It is also worth discussing the causes of some misclassifications that we noticed. One of the problems originates at the beginning of the chain. Since background subtraction reduces the search space, it may reject some actual person parts. This happens mainly when a person is situated at the same depth as something that belongs to the background model. This could be improved by combining the different modalities in order to learn the background model. Furthermore, the contours of the foreground binary masks may not be perfect, either. One possible solution would be to apply GrabCut or other post-segmentation approaches to refine and smooth the contours, which in turn would improve segmentation accuracy. Another issue is that some regions considered *unknown*—mostly those generated when one person overlaps other—differ considerably from those that are used to train the different models. Hence, the classification of such regions is not a trivial task.

6 Conclusions

We first introduced a novel RGB–Depth–Thermal dataset of video sequences, which contains several subjects interacting with everyday objects, along with a registration algorithm and the manual pixel-level annotations of human masks. Second, we proposed a multi-modal human body segmentation

[5] Check the video included as supplementary material in which some qualitative results are shown, named trimodal_seg_results.mp4.

approach using the registered RGB–Depth–Thermal data as a preprocessing step for human activity recognition tasks.

The registration algorithm registered the different data modalities using multiple homographies generated from several views of the proposed calibration device. The segmentation baseline segmented the people appearing in a set of 10 trimmed video sequences out of the three recorded scenes. It consisted of, first, a non-adaptive background subtraction approach in order to extract the regions of interest that deviate from the depth-background model previously learned. The regions from the different modalities were partitioned in a grid of cells. The cell were then described in the corresponding modalities using state-of-the-art image feature descriptors. HOG and HOF were computed on RGB color imagery, a histogram of intensity gradients on thermal, and histograms of normal vectors' orientations on depth. For each cell and modality, we modeled the distribution of descriptions using a GMM. During the prediction phase, cells were evaluated in the corresponding GMMs and the obtained likelihoods turned into confidence-like terms and stacked in a feature vector representation. A supervised learning algorithm, such as Random Forest, learned to categorize such representation into human or non-human regions.

In the end, we found notable performance improvements with the proposed learning-based fusion strategies in comparison to each isolated modality, and Random Forest obtained the best results. Furthermore, our baseline outperformed different state-of-the-art uni-modal segmentation methods, hence demonstrating the power of multi-modal fusion.

Acknowledgments This work was partly supported by the Spanish Project TIN2013-43478-P. The work of Albert Clapés was supported by SUR-DEC of the Generalitat de Catalunya and FSE. We would like to thank Anders Jørgensen for his valuable help in capturing the dataset.

References

Abidi, B. (2007). IRIS thermal/visible face database. DOE University Research Program in Robotics under grant DOE-DE-FG02-86NE37968

Alahari, K., Seguin, G., Sivic, J., & Laptev, I. (2013). Pose estimation and segmentation of people in 3D movies. In *IEEE international conference on computer vision (ICCV 2013)*.

Alpert, S., Galun, M., Basri, R., & Brandt, A. (2007). Image segmentation by probabilistic bottom-up aggregation and cue integration. In *IEEE conference on computer vision and pattern recognition, 2007 (CVPR '07)* (pp. 1–8). doi:10.1109/CVPR.2007.383017.

Andriluka, M., Roth, S., & Schiele, B. (2009). Pictorial structures revisited: people detection and articulated pose estimation. In *IEEE conference on computer vision and pattern recognition, 2009 (CVPR 2009)* (pp. 1014–1021).

Andriluka, M., Roth, S., & Schiele, B. (2010). Monocular 3D pose estimation and tracking by detection. In *IEEE conference on computer vision and pattern recognition, 2010 (CVPR 2010)* (pp. 623–630).

Barbosa, I.B., Cristani, M., Del Bue, A., Bazzani, L., & Murino, V. (2012). Re-identification with RGB-D sensors. In *Computer vision ECCV 2012. Workshops and demonstrations* (pp. 433-442). Berlin: Springer.

Bertozzi, M., Broggi, A., Gomez, C.H., Fedriga, R.I., Vezzoni, G., & Del Rose, M. (2007). Pedestrian detection in far infrared images based on the use of probabilistic templates. In *Intelligent vehicles symposium. 2007 IEEE* (pp. 327–332). Piscataway: IEEE.

Bouguet, J. Y. (2004). Camera calibration toolbox for matlab.

Bourdev, L., & Malik, J. (2009). Poselets: body part detectors trained using 3D human pose annotations. In *IEEE 12th international conference on computer vision, 2009* (pp. 1365–1372).

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

Bouwmans, T., El Baf, F., Vachon, B., et al. (2008). Background modeling using mixture of gaussians for foreground detection: A survey. *Recent Patents on Computer Science*, *1*(3), 219–237.

Boykov, Y. Y., & Jolly, M. P. (2001). Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In *Proceedings of eighth IEEE international conference on computer vision, 2001 (ICCV 2001)* (Vol. 1, pp. 105–112).

Bradski, G., & Kaehler, A. (2008). *Learning OpenCV: Computer vision with the OpenCV library*. Sebastopo: O'reilly.

Bray, M., Kohli, P., & Torr, P.H.S. (2006). Posecut: Simultaneous segmentation and 3D pose estimation of humans using dynamic graph-cuts. In *Computer vision–ECCV 2006* (pp. 642–655). Berlin: Springer.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brkić, K., Rašić, S., Pinz, A., Šegvić, S., & Kalafatić, Z. (2013). Combining spatio-temporal appearance descriptors and optical flow for human action recognition in video data. arXiv:1310.0308.

Buys, K., Cagniart, C., Baksheev, A., De Laet, T., De Schutter, J., & Pantofaru, C. (2014). An adaptable system for RGB-D based human body detection and pose estimation. *Journal of Visual Communication and Image Representation*, *25*(1), 39–52.

Camplani, M., & Salgado, L. (2014). Background foreground segmentation with RGB-D Kinect data: An efficient combination of classifiers. *Journal of Visual Communication and Image Representation*, *25*(1), 122–136.

Carson, C., Belongie, S., Greenspan, H., & Malik, J. (2002). Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(8), 1026–1038.

Charles, J., Everingham, M. (2011). Learning shape models for monocular human pose estimation from the Microsoft Xbox Kinect. In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)* (pp. 1202–1208).

Chun, S.Y., Lee, C.S. (2013). Applications of human motion tracking: Smart lighting control. In *2013 IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 387–392).

Clapés, A., Reyes, M., & Escalera, S. (2012). User identification and object recognition in clutter scenes based on RGB-Depth analysis. In *Articulated motion and deformable objects* (pp. 1–11). Berlin: Springer.

Cohen, W. W. (2005). *Stacked sequential learning*. DTIC Document: Technical report.

Dai, C., Zheng, Y., & Li, X. (2007). Pedestrian detection and tracking in infrared imagery using shape and appearance. *Computer Vision and Image Understanding*, *106*(2), 288–299.

Dalal, N., Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE computer society conference on computer vision and pattern recognition, 2005* (CVPR 2005) (Vol. 1, pp. 886–893).

Dalal, N., Triggs, B., Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *Computer vision–ECCV 2006* (pp. 428–441) Berlin: Springer.

Davis, J. W., & Sharma, V. (2004). Robust background-subtraction for person detection in thermal imagery. In *IEEE international workshop on object tracking and classification beyond the visible spectrum.*

Davis, J. W., & Sharma, V. (2007). Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, *106*(2), 162–182.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2012). The PASCAL visual object classes challenge 2012 results. See http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Fanelli, G., Dantone, M., Gall, J., Fossati, A., & Van Gool, L. (2013). Random forests for real time 3D face analysis. *International Journal of Computer Vision*, *101*(3), 437–458.

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Image analysis* (pp. 363–370) Berlin: Springer.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(9), 1627–1645.

Fernández-Caballero, A., Castillo, J. C., Serrano-Cuerda, J., & Maldonado-Bascón, S. (2011). Real-time human segmentation in infrared videos. *Expert Systems with Applications*, *38*(3), 2577–2584.

Fernández-Sánchez, E. J., Díaz, J., & Ros, E. (2013). Background subtraction based on color and depth using active sensors. *Sensors*, *13*(7), 8895–8915.

Fidler, S., Mottaghi, R., Yuille, A., & Urtasun, R. (2013). Bottom-up segmentation for top-down detection. In *2013 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3294–3301).

Gade, R., & Moeslund, T. B. (2014). Thermal cameras and applications: A survey. *Machine Vision and Applications*, *25*(1), 245–262.

Gade, R., Jorgensen, A., & Moeslund, T. B. (2013). Long-term occupancy analysis using graph-based optimisation in thermal imagery. In *2013 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3698–3705).

Giordano, D., Palazzo, S., & Spampinato, C. (2014). Kernel density estimation using joint spatial-color-depth data for background modeling. In *2014 22nd international conference on pattern recognition (ICPR)* (pp. 4388–4393). Piscataway: IEEE.

Girshick, R. B., Felzenszwalb, P. F., & Mcallester, D.A. (2011). Object detection with grammar models. In *Advances in neural information processing systems* (pp. 442–450).

Gordon, G., Darrell, T., Harville, M., & Woodfill, J. (1999). Background estimation and removal based on range and color. In *IEEE computer society conference on computer vision and pattern recognition, 1999* (Vol. 2).

Gulshan, V., Lempitsky, V., & Zisserman, A. (2011). Humanising grabCut: learning to segment humans using the Kinect. In 2011 IEEE International conference on computer vision workshops (ICCV workshops) (pp. 1127–1133).

Hernández-Vela, A., Bautista, M. A., Perez-Sala, X., Ponce, V., Baró, X., Pujol, O., et al. (2012a). BoVDW: Bag-of-Visual-and-Depth-Words for gesture recognition. In *2012 21st International conference on pattern recognition (vICPR)* (pp. 449–452). Piscataway: IEEE.

Hernández-Vela, A., Zlateva, N., Marinov, A., Reyes, M., Radeva, P., Dimov, D., Escalera, S. (2012b). Graph cuts optimization for multi-limb human segmentation in depth maps. In *2012 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 726–732).

Hg, R. I., Jasek, P., Rofidal, C., Nasrollahi, K., Moeslund, T. B., Tranchet, G., et al. (2012). An RGB-D database using Microsoft's Kinect for Windows for face detection. In *2012 eighth international conference on signal image technology and internet based systems (SITIS)* (pp. 42–46). Piscataway: IEEE.

Holt, B., Ong, E.J., Cooper, H., & Bowden, R. (2011). Putting the pieces together: Connected poselets for human pose estimation. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)* (pp. 1196–1201).

Huynh, T., Min, R., & Dugelay, J. L. (2013). An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data. In *Computer vision-ACCV 2012 workshops* (pp. 133–145). Berlin: Springer.

Irani, R., Nasrollahi, K., Oliu, M., Corneanu, C., Escalera, S., Bahnsen, C., Lundtoft, D., Moeslund, T. B., Pedersen, T., Klitgaa, M.L., & Petrini, L. (2015). Spatiotemporal analysis of rgb-d-t facial images for multi-modal pain level recognition. In *IEEE conference on computer vision and pattern recognition workshop*.

Koppula, H. S., Gupta, R., & Saxena, A. (2013). Learning human activities and object affordances from RGB-D videos. *The International Journal of Robotics Research*, *32*(8), 951–970.

Kumar, M.P., Ton, P. H. S., & Zisserman, A. (2005). Obj cut. In *IEEE computer society conference on computer vision and pattern recognition, 2005 (CVPR 2005)* (Vol. 1, pp. 18–25).

Ladický, L., Sturgess, P., Alahari, K., Russell, C., & Torr, P. H. S. (2010). What, where and how many? combining object detectors and crfs. In *Computer vision–ECCV 2010* (pp. 424–437) Berlin: Springer.

Leibe, B., Leonardis, A., & Schiele, B. (2004). Combined object categorization and segmentation with an implicit shape model. In *Workshop on statistical learning in computer vision, ECCV* (Vol. 2, p. 7).

Leibe, B., Leonardis, A., & Schiele, B. (2008). Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, *77*(1–3), 259–289.

Levin, A., & Weiss, Y. (2006). Learning to combine bottom-up and top-down segmentation. In *Computer vision–ECCV 2006* (pp. 581–594). Berlin: Springer.

Leykin, A., & Hammoud, R. (2006). Robust multi-pedestrian tracking in thermal-visible surveillance videos. In *IEEE conference on computer vision and pattern recognition workshop 2006. (CVPRW'06)* (p. 136).

Leykin, A., Ran, Y., & Hammoud, R. (2007). Thermal-visible video fusion for moving target tracking and pedestrian classification. In *IEEE conference on computer vision and pattern recognition, 2007.* (CVPR'07) (pp. 1–8).

Lin, Z., Davis, L.S., Doermann, D., & DeMenthon, D. (2007). An interactive approach to pose-assisted and appearance-based segmentation of humans. In *IEEE 11th international conference on computer vision, 2007* (ICCV 2007) (pp 1–8).

Lopes, O., Reyes, M., Escalera, S., & Gonzalez, J. (2014). Spherical blurred shape model for 3D object and pose recognition: Quantitative analysis and hci applications in smart environments.

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of eighth IEEE international conference on computer vision, 2001 (ICCV 2001)* (Vol. 2, pp. 416–423).

Mittal, A., Zhao, L., & Davis, L. S. (2003). Human body pose estimation using silhouette shape analysis. In *Proceedings of IEEE conference on advanced video and signal based surveillance, 2003* (pp 263–270).

Moeslund, T. B. (2011). *Visual analysis of humans: Looking at people*. London: Springer.

Møgelmose, A., Bahnsen, C., Moeslund, T., Clapés, A., & Escalera, S. (2013). Tri-modal person re-identification with rgb, depth and thermal features. In *IEEE conference on computer vision and pattern recognition workshops (CVPRW), 2013* (pp. 301–307). doi:10.1109/CVPRW.2013.52.

Mori, G., Ren, X., Efros, A. A., & Malik, J. (2004). Recovering human body configurations: combining segmentation and recognition. In *Proceedings of the 2004 IEEE computer society conference on* *computer vision and pattern recognition, 2004* (CVPR 2004) (Vol. 2, pp. II-326).

Nghiem, A.T., Bremond, F., Thonnat, M., & Valentin, V. (2007). ETISEO, performance evaluation for video surveillance systems. In *IEEE conference on advanced video signal based surveillance, 2007* (AVSS 2007) (pp. 476–481).

Nikisins, O., Nasrollahi, K., Greitans, M., & Moeslund, T. (2014). Rgb-d-t based face recognition. In *2014 22nd international conference on pattern recognition (ICPR)* (pp. 1716–1721).

Olmeda, D., de la Escalera, A., & Armingol, J. M. (2012). Contrast invariant features for human detection in far infrared images. In *2012 IEEE on Intelligent Vehicles Symposium (IV)* (pp. 117–122).

Oreifej, O., Liu, Z. (2013). Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *2013 IEEE conference on computer vision and pattern recognition (CVPR)*. (pp. 716–723).

Otsu, N. (1975). A threshold selection method from gray-level histograms. *Automatica*, *11*(285–296), 23–27.

Pirsiavash, H., Ramanan, D. (2012). Steerable part models. In *2012 IEEE conference on computer vision and pattern recognition (CVPR)*. (pp. 3226–3233).

Plagemann, C., Ganapathi, V., Koller, D., Thrun, S. (2010). Real-time identification and localization of body parts from depth images. In *2010 IEEE international conference on robotics and automation (ICRA)*. (pp. 3108–3113).

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing*, *28*, 976–990. doi:10.1016/j.imavis.2009.11.014.

Puertas, E., Escalera, S., Pujol, O. (2013). Generalized multi-scale stacked sequential learning for multi-class classification. *Pattern Analysis and Applications*, 1–15

Pugeault, N., Bowden, R. (2011). Spelling it out: Real-time asl fingerspelling recognition. In *2011 IEEE International conference on computer vision workshops (ICCV workshops)*. (pp. 1114–1119).

Ramanan, D. (2006). Learning to parse images of articulated bodies. In *Advances in neural information processing systems*. (pp. 1129–1136).

Rother, C., Kolmogorov, V., Blake, A. (2004). Grabcut: interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*. (Vol. 23, pp. 309–314). ACM.

Scharwächter, T., Enzweiler, M., Franke, U., Roth, S. (2013). Efficient multi-cue scene segmentation. In *Pattern Recognition*. (pp. 435–445).

Schwarz, L.A., Mkhitaryan, A., Mateus, D., Navab, N. (2011). Estimating human 3D pose from time-of-flight images based on geodesic distances and optical flow. In *2011 IEEE International Conference on Automatic Face &amp; Gesture Recognition and Workshops (FG 2011)*. (pp. 700–706).

Sheasby, G., Warrell, J., Zhang, Y., Crook, N., Torr, P.H.S. (2012). Simultaneous human segmentation, depth and pose estimation via dual decomposition. In *British Machine Vision Conference, Student Workshop, BMVW*.

Sheasby, G., Valentin, J., Crook, N., Torr, P. (2013). A robust stereo prior for human segmentation. In *Computer Vision–ACCV 2012*. (pp 94–107). Berlin: Springer.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888–905.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. (CVPR '11). (pp. 1297–1304). Washington, DC: IEEE Computer Society. doi:10.1109/CVPR.2011.5995316

Spinello, L., Arras, K.O. (2011). People detection in RGB-D data. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. (pp. 3838–3843).

Stauffer, C., Grimson, W.E.L. (1999). Adaptive background mixture models for real-time tracking. In *IEEE Compututer Society Conference on Computer Vision and Pattern Recognition, 1999*. (Vol. 2)

Stefańczyk, M., & Kasprzak, W. (2012). Multimodal segmentation of dense depth maps and associated color information. In *Computer vision and graphics*. (pp. 626–632). Berlin: Springer.

Suard, F., Rakotomamonjy, A., Bensrhair, A., Broggi, A. (2006). Pedestrian detection using infrared images and histograms of oriented gradients. In *Intelligent Vehicles Symposium, 2006 IEEE*. (pp. 206–212).

Susperregi, L., Martínez-Otzeta, J.M., Ansuategui, A., Ibarguren, A., Sierra, B. (2013). RGB-D, laser and thermal sensor fusion for people following in a mobile robot. International Journal of Advanced Robotic Systems, 10.

Teichman, A., & Thrun, S. (2013). Learning to segment and track in RGB-D. *Algorithmic Found* (pp. 575–590). Robot. X: Springer.

Vidas, S., Lakemond, R., Denman, S., Fookes, C., Sridharan, S., & Wark, T. (2012). A mask-based approach for the geometric calibration of thermal-infrared cameras. *IEEE Transactions on Instrumentation and Measurement*, *61*(6), 1625–1635.

Vineet, V., Sheasby, G., Warrell, J., & Torr, P. H. S. (2013). PoseField: An efficient mean-field based method for joint estimation of human pose, segmentation, and depth. In *Energy minimization methods in computer vision and pattern recognition*. (pp. 180–194). Berlin: Springer.

Viola, P., Jones, M. J., & Snow, D. (2005). Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, *63*(2), 153–161.

Wang, L., Qiao, Y., Tang, X. (2013). Motionlets: mid-level 3D parts for human motion recognition. In *2013 IEEE conference on computer vision and pattern recognition (CVPR)*. (pp. 2674–2681).

Wang, W., Zhang, J., Shen, C. (2010). Improved human detection and classification in thermal images. In *2010 17th IEEE International Conference on Image Processing (ICIP)*. (pp. 2313–2316).

Wang, Y., Tran, D., Liao, Z. (2011). Learning hierarchical poselets for human parsing. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (pp. 1705–1712).

Windheuser, T., Schlickewei, U., Schmidt, F.R., Cremers, D. (2011). Geometrically consistent elastic matching of 3D shapes: a linear programming solution. In *2011 IEEE international conference on computer vision (ICCV)*. (pp. 2134–2141).

Wolf, C., Mille, J., Lombardi, E., Celiktutan, O., Jiu, M., Baccouche, M., Dellandréa, E., Bichot, C.E., Garcia, C., Sankur, B. (2012). The LIRIS human activities dataset and the ICPR 2012 human activities recognition and localization competition. In *LIRIS Umr 5205 CNRS/INSA Lyon/Universite'Claude Bernard Lyon 1/Universite'Lumie 're Lyon 2/E'cole Cent*.

Xia, L., Chen, C.C., Aggarwal, J.K. (2011). Human detection using depth information by kinect. In *2011 IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW)*. (pp. 15–22).

Yang, Y., Ramanan, D. (2011). Articulated pose estimation with flexible mixtures-of-parts. In *2011 IEEE conference on computer vision and pattern recognition (CVPR)*. (pp. 1385–1392).

Yang, Y., & Ramanan, D. (2013). Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2878–2890.

Yao, B., Fei-Fei, L. (2010). Grouplet: a structured image representation for recognizing human and object interactions. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (pp. 9–16).

Zhang, L., Wu, B., Nevatia, R. (2007). Pedestrian detection in infrared images based on local shape features. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. (CVPR'07)*. (pp. 1–8).

Zhao, J., Sen-ching, S.C. (2012). Human segmentation by geometrically fusing visible-light and thermal imageries. Multimedia Tools and Applications, 1–29.

Zhu, L., Chen, Y., Lu, Y., Lin, C., Yuille, A. (2008). Max margin and/or graph learning for parsing the human body. In *IEEE conference on computer vision and pattern recognition, 2008. (CVPR 2008)*. (pp. 1–8).

Zivkovic, Z. (2004). Improved adaptive Gaussian mixture model for background subtraction. In *Proceedings of the 17th international conference on pattern recognition, 2004. (ICPR 2004).* (Vol. 2, pp 28–31).