Accurate Human-Limb Segmentation in RGB-D images for Intelligent Mobility Assistance Robots

Siddhartha Chandra[1,2] Stavros Tsogkas[2] Iasonas Kokkinos[1,2]
siddhartha.chandra@inria.fr stavros.tsogkas@centralesupelec.fr iasonas.kokkinos@inria.fr
[1] INRIA GALEN, Paris, France [2] Centrale Supélec, Paris, France

Abstract

Mobility impairment is one of the biggest challenges faced by elderly people in today's society. The inability to move about freely poses severe restrictions on their independence and general quality of life. This work is dedicated to developing intelligent robotic platforms that assist users to move without requiring a human attendant. This work was done in the context of an EU project involved in developing an intelligent robot for elderly user assistance. The robot is equipped with a Kinect sensor, and the vision-component of the project has the responsibility of locating the user, estimating the user's pose, and recognizing gestures by the user. All these goals can take advantage of a method that accurately segments human-limbs in the colour (RGB) and depth (D) images captured by the Kinect sensor. We exploit recent advances in deep-learning to develop a system that performs accurate semantic segmentation of human limbs using colour and depth images. Our novel technical contributions are the following: 1) we describe a scheme for manual annotation of videos, that eliminates the need to annotate segmentation masks in every single frame; 2) we extend a state of the art deep learning system for semantic segmentation, to exploit diverse RGB and depth data, in a single framework for training and testing; 3) we evaluate different variants of our system and demonstrate promising performance, as well the contribution of diverse data, on our in-house Human-Limb dataset. Our method is very efficient, running at 8 frames per second on a GPU.

1. Introduction

Mobility impairment is widespread among elderly people today. Recent studies indicate that 20% of people between the age of $70-85$ years, and 50% of people above the age of 85 years report difficulties accomplishing basic activities of daily life [21]. Mobility impairment not only restricts independence severely, it is also detrimental to the self-esteem of the individuals and drastically limits their quality of life. Studies in demographics show that mobility impairment among people over the age of 65 is constantly increasing in industrialized countries [5]. The goal of this work is the design and implementation of intelligent mobility assistance robots, or *nurse-type robots*, that help the user accomplish basic everyday tasks such as standing up, sitting down, and moving about, avoiding stationary physical obstacles. Users would be able to interact with these robots through pre-determined gestures, and be able to draw the robot closer, or farther, and indicate the type assistance they need. The computer vision component of such systems would exploit information, in the form of RGB-D images, from Kinect-sensors to allow the system (a) locate the user, (b) estimate their pose, and (c) understand their gestures. All these tasks involve recognizing landmarks/parts on the human body. While locations of these parts would give the system a fair indication of the user's location, the relative orientations of these parts would enable the system to recognize pose/ gesture. To this end, we develop a deep-learning based solution to human-limb segmentation, which provides accurate pixel-wise class labels to each pixel in the images captured by the camera. Figure 2 shows one prac-

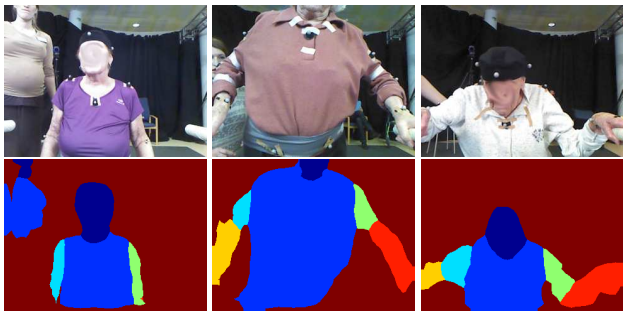

Figure 1. Our system takes RGB or RGB-D images as input and produces a parsing of the person's limbs, torso, and head.

tical usecase of our solution. Human-limb segmentations, in conjunction with the depth field, can be used to estimate $3-D$ surfaces of body parts of the user. Normals of these surfaces can be used to determine if the user's pose is unstable. In case of instability, the robot can assist the user achieve a stable position by exerting a physical force of the optimal magnitude in the optimal direction.

Our contributions are threefold. We first create the Human-Limb-dataset from videos captured by a Kinect sensor, and devise a scheme for annotating pixelwise class-labels that does not require us to annotate every single frame in the video. Second, we combine the state of the art Deeplab network for semantic segmentation [6] in RGB images, with the state of the art Alexnet-HHA network [13] on object detection in depth images to train a single network with the objective of pixelwise semantic segmentation in RGB-D images. Finally, we explain how we can re-purpose a deep CNN using Caffe [14], to use diverse data (RGB + depth fields) in a unified framework for training and testing. Importantly, our method is also very efficient, running at 8 frames per second.

Figure 3 gives a visual summary of our framework. We begin with a brief review of recent advances in semantic segmentation, followed by describing our contributions, and finally report empirical results on the Human-Limb dataset.

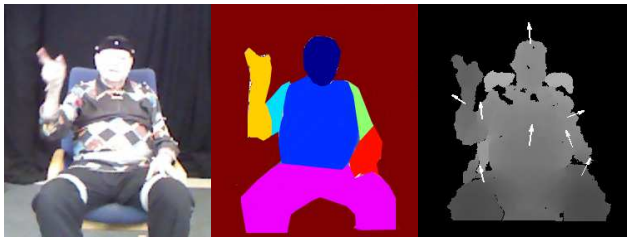

Figure 2. Human-limb segmentations, alongside depth, can be exploited to fit surfaces corresponding to parts of the body. Normals to these surfaces can then be estimated, and these normals give an indication of the stability of the user's pose. In the event of the user being in an unstable pose, the robot can use knowledge of these normals to help the user gain a stable pose by exerting the necessary physical force in the optimal direction.

2. Related Work

Semantic segmentation of body parts, and fine-grained semantic segmentation have attracted increased interest in recent years. Bo and Fowlkes [3] sequentially merge superpixels obtained from an oversegmentation, to form larger parts, such as head, arms, clothes etc., and parse pedestrians. Simple constraints, (such as that *head* appears above *upper body* and that *hair* appears above *head*) enforce a consistent layout of parts in resulting segmentations

In [9] Eslami *et al*. introduced the Shape Boltzmann Machine (SBM), an hierarchical generative model that can generate realistic samples from the underlying object shape distribution. This model was extended to deal with multiple region labels (parts) in [8] and coupled with a model for part appearances.

Lu *et al*. formulate car parsing as a landmark identification problem [20]. In [25] the authors draw inspiration from previous work on hierarchical models for objects [15, 16, 26, 27] and propose a mixture of compositional models that represent horses and cows in terms of their boundaries and the boundaries of semantic parts. Their algorithm starts by segmenting large parts first, such as head, neck, torso, and moves on to segment legs, which are deformable and thus much more difficult to segment.

Many works have used depth images to complement RGB information for scene labelling. Ren *et al*. convert local similarities (kernels) to patch descriptors, and incorporate context information using superpixel MRFs and segmentation trees [22]. Couprie *et al*. [7] adopt a multiscale convolutional neural network to learn features directly from RGB-D images, whereas Wang *et al*. attempt to learn visual patterns from RGB and depth in a joint manner via an unsupervised learning framework [18, 24]. They sample RGB-D patches and feed them as input to a two-layer stacked structure. The output of their method is a collection of superpixels that combine features from different sampled patches. As a final step, they train linear SVM classifiers to assign scene labels to each superpixel.

Gupta *et al*. generalize the popular contour detection and hierarchical segmentation *gPb* algorithm [1] to make effective use of depth information. In [12], they design features based on local geometric contour information, shape, size and appearance of superpixels, and train classifiers to assign each superpixel a semantic class label. In [13], they follow an object-centric approach. They propose a three-channel encoding of a raw depth field, and modify R-CNN [11] so as to exploit this new type of information. Detection obtained from this modified R-CNN system are then used to compute additional features for superpixels and improve the semantic segmentation system of [12].

Recently, Banica and Sminchisescu modified CPMC [4] to account for both intensity and depth discontinuities [2]. A pool of figure-ground segmentations is created and a ranker is trained to distinguish which of those segments correspond to objects. A ranking process selects a valid and compact subset of segments, while local, hand-crafted features, as well as learned CNN features, are used to classify the retained segmentation masks.

Our work combines ideas from [12, 13] and recent works on semantic segmentations of objects that employ deep convolutional neural networks [6, 19]. We augment the Deeplab system to take advantage of depth information, and change the labelling task from object segmentation to hu-

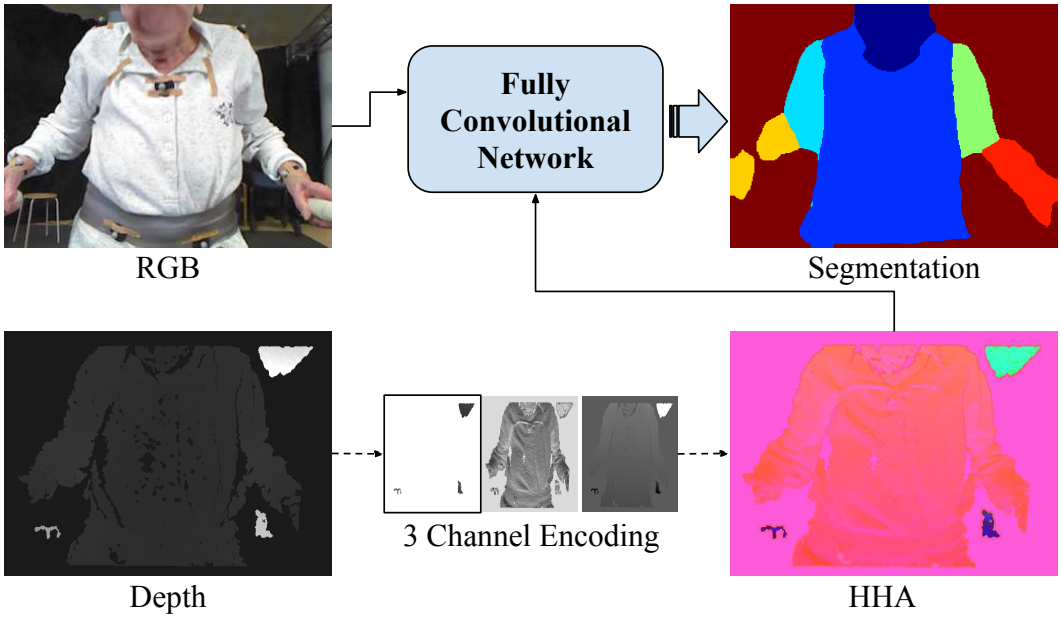

Figure 3. Our method combines RGB and depth information to train a fully convolutional deep neural network for human limb semantic segmentation.

man limb parsing. Training and testing is performed using a single network that can seamlessly switch from only using RGB or depth images to the augmented RGB-D inputs.

3. Dataset

Our data-acquisition setup consisted of a Kinect sensor mounted on top of a passive rollator, as in [21]. The recording process involved capturing RGB-D videos using the open-source Robotics Operating System (ROS) software capturing human subjects in a set of predefined use-cases and scenarios. Example colour and depth images are shown in Figure 4, along with manually annotated ground-truth.

Our Human-Limb dataset consists of six recorded videos, each containing a different human subject performing a sit-stand-sit activity several times. These six videos correspond to 2618 RGB-D frames in total. We split the dataset into 1399 training and 1219 testing frames: images of three human subjects are used to construct the training set, while images of the remaining three subjects are used to construct the testing set. We annotate each of these images with 7 labels, namely, (1) head, (2) torso, (3) left upper arm, (4) right upper arm, (5) left lower arm, (6) right lower arm, and (7) background. In Section 3.1, we describe the scheme we employed for efficient annotation, exploiting the visual similarity of frames in a video.

3.1. Annotation

Annotating a video/sequence of frames with pixel-wise labels can be a tedious task. One key observation specific to this task is that videos typically contain a lot of visually similar frames. While consecutive frames in a video are usually similar, in our recordings, the human subjects perform the same task (sit-stand-sit), repeatedly, increasing redundancy among frames.

To avoid annotating very similar frames more than once, we use a clustering-based approach to reduce the dataset into a set of representative frames. We annotate each of these representative frames using the open source image editor GIMP (`www.gimp.org`). Finally, the ground truth labels for each frame in the dataset are transferred from the annotation of the visually most similar frame from the set of representative frames.

The clustering approach we use to determine the representative frames consists of the following steps: we begin by computing HOG [10] features on all colour image frames in our dataset. Next, we compute euclidean distances in the HOG space between each pair of frames, and lastly, we employ a greedy strategy to group frames into clusters,using a threshold on the distance in the HOG space: any two frames within a distance less than d_{thresh} are merged into the same cluster. The hyper-parameter d_{thresh} is manually picked, ensuring that all the frames in each cluster look very similar. One frame from each cluster is randomly picked to be the cluster representative. Figure 5 shows some example clusters returned by our algorithm. On average, we obtained 28 clusters per video in our dataset. In the end we made a

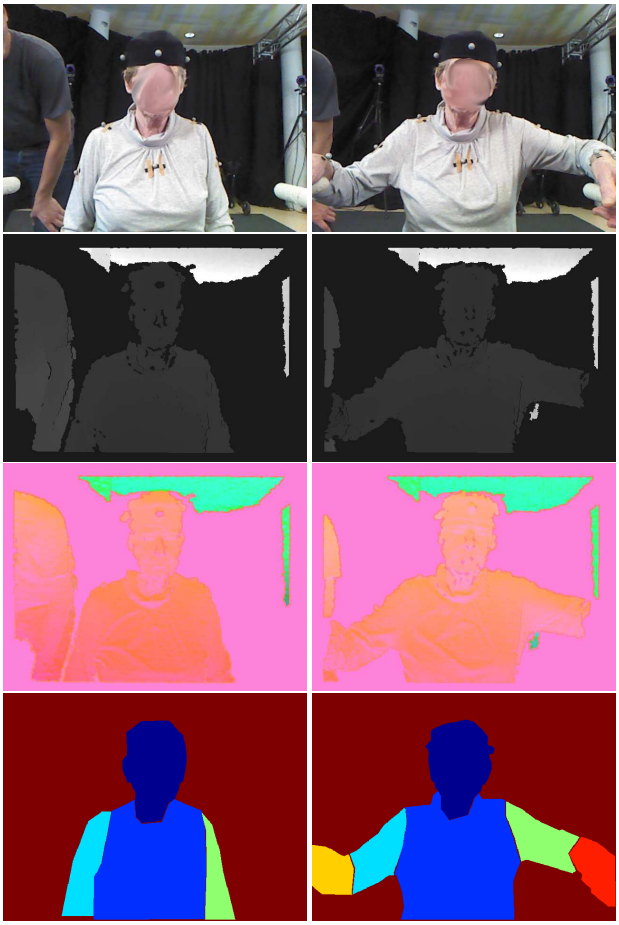

Figure 4. Example images from the dataset. Row 1 shows the RGB images. Row 2 shows the depth images. Row 3 shows the depth images encoded to 3 channel HHA [13], and Row 4 shows the pixelwise ground truth class labels.

final pass of the full dataset, to visually assess the quality of the transferred annotations.

4. Human-Limb Segmentation in RGB-D Images

Given the dataset of RGB-D images and pixelwise class-labels, we now describe our deep learning framework for human-limb segmentation. Our approach builds on top of the Deeplab network [6], which is a state of the art method for semantic segmentation on RGB images. In this work, we propose an extension to the Deeplab architecture, that allows it to exploit RGB-D images. We begin by describing the original network, followed by our extension.

4.1. Deeplab Network for Semantic Segmentation in RGB images

In recent years, Convolutional Neural Networks have shown unprecedented performance in object classification and detection [11, 17, 23], and they have already become the standard learning machinery for many other computer vision tasks. Recently, it was shown that CNNs can also be used in a fully convolutional setting for semantic segmentation [6, 19].

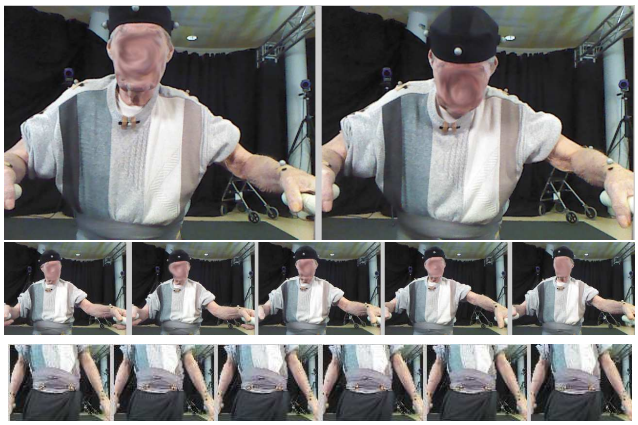

Figure 5. Example clusters returned by our clustering method, which puts two images together in the same cluster if the euclidean distance between them in the HOG space is below a threshold. We exploit the visual similarity of images in the same cluster by annotating one image from the cluster and propagating its annotation to other images in the cluster.

One of the challenges faces when using convolutional networks for semantic segmentation is the downsampling factor. Due to repeated max-pooling and downsampling, the spatial resolution of an input image is reduced as it is propagated through the network. The popular Alexnet [17] has a down-sampling factor of 16. This means if we use a 128×128 image as input, we will get 8×8 activations (ignoring the last dimension for brevity–we are only concerned with the spatial dimensions in this example). While this is not a problem for tasks like image classification and object detection, it is undesirable for tasks such as image segmentation where we require a class label for each of the 128×128 pixels. Output maps typically have to be scaled 16 times in a post processing procedure, thus introducing approximation errors. The Deeplab network proposed by Chen *et al*. [6], reduces the down-sampling factor to 8 by introducing holes in the convolution kernel. In this work, we build on the fully convolutional Deeplab CNN and adapt it for the semantic segmentation of human limbs. A schematic representation of the Deeplab network is shown in Figure 6.

The Deeplab network is based on the very deep model for object classification by Simonyan *et al*. [23], after dropping the fully connected layers. This results in a fully convolutional network, preserving spatial information that is normally scraped away by fully connected layers. More precisely, the Deeplab network consists of 16 convolutional layers, interleaved with 5 pooling layers. It also has 15 normalization layers, and 2 dropout layers. We initialize the

weights using the VGG pretrained network and finetune the network for semantic segmentation of human limbs. For optimization, we use standard stochastic gradient descent, minimizing a cross-entropy loss, averaged over all image positions. As in [6], we also use the post-processing step of applying a fully-connected CRF to obtain sharper masks, rather than simply relying on the CNN coarse responses. Our experiments indicate that using CRF leads to moderate improvements in segmentation accuracy.

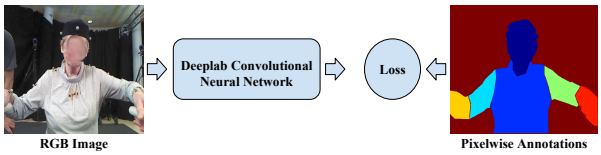

Figure 6. Schematic representation of the Deeplab network, using RGB images as input.

4.2. Alexnet-HHA Network for Semantic Segmentation in depth images

The Deeplab network described in the previous subsection uses RGB images. However, we would like to exploit the available depth frames in addition to the RGB images. A straightforward approach to exploit depth information for our task, would be to simply add the raw depth field as an extra channel in the RGB input images during training. This is problematic for two reasons: first, the range of values for the depth channel is very different than the one for RGB inputs, which means that a calibration step could be needed to account for this discrepancy during training. Second, instead of taking advantage of the initialized weights in the pre-trained model and simply finetune the network with a few passes over the dataset, we would be forced to retrain the full network, a procedure that usually takes several days.

Gupta *et al*. [13] propose an alternative approach: instead of taking the depth fields at face value, they encode each pixel value of the depth fields with a 3×1 vector representing horizontal disparity(H), height above the ground(H) and the angle(A) the pixel's local surface normal makes with the inferred gravity direction. As a result, an $M \times N$ depth field is transformed into an $M \times N \times 3$, *HHA* encoding, and values are linearly scaled to map observed values across the training dataset to the 0 to 255 range. Authors in [13] finetune the popular Alexnet [17] using HHA images for the task of object detection. This representation encodes properties of geocentric pose a CNN would have difficulty learning directly from a depth image, especially when very limited training data is available. Rows 2,3 in Figure 4 show two example depth images and their corresponding HHA representations respectively.

In this work, we take advantage of the pretrained Alexnet-HHA in a similar way. We start by encoding the depth frames in our dataset to the HHA representation, and use them to finetune a modified version of Alexnet-HHA for the task of semantic segmentation. We optimize performance on the training set using the same cross-entropy loss function as in [6]. The original Alexnet has 5 convolutional layers, 3 pooling layers, and several normalization and dropout layers. As stated in section 4.1, Alexnet has a downsampling factor of 16. To retain a downsampling factor of 8 while being able to use the pretrained network weights, we discard any convolutional / fully connected layers after the second convolutional layer. A schematic representation of the Alexnet-HHA network is shown in Figure 7.

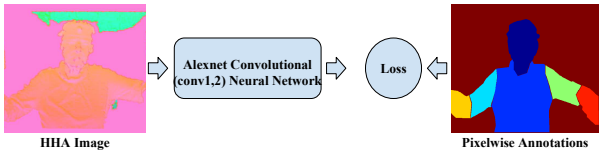

Figure 7. Schematic representation of our modified Alexnet-HHA network that uses a 3-channel encoding of the depth fields as input.

4.3. Human-limb Segmentation in RGB-D images

Having described our extensions of the Deeplab and Alexnet-HHA networks for semantic segmentation in RGB and depth images respectively, we now turn to the task of combining them to train a single network which learns from colour and depth images simultaneously. To this end we combine the two networks by concatenating their penultimate layers, and then using a single loss layer on top. A schematic representation of our combined network is shown in Figure 8.

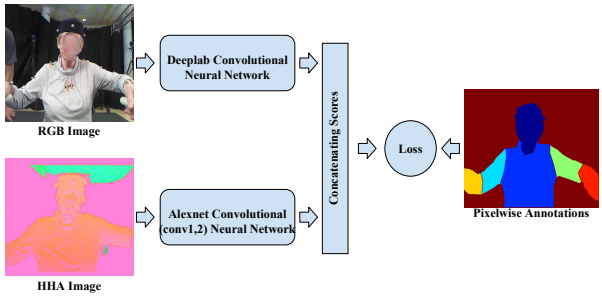

Figure 8. We combine RGB and HHA-encoded depth information in a single deep network.

4.4. Learning from RGB, HHA, RGB+HHA images

Proliferation of depth sensors is a recent phenomenon. Consequently, we have a lot of publicly available datasets with colour images, and comparatively fewer depth image datasets. From a technical standpoint, it is desirable to have a single network that can exploit RGB-D images, or resort to using only one type of information when the other is not available. To incorporate this flexibility in our network, we implement a novel *adaptiveConcatenation* layer

in *Caffe* [14]. The *adaptiveConcatenation* layer accepts a flag alongside each image which indicates whether the input image is RGB only, HHA only or RGB+HHA. During the feed-forward phase, the *adaptiveConcatenation* layer ignores penultimate layer activations of the HHA component if the image is RGB only, and ignores the penultimate layer activations of the RGB component if the layer is HHA only. Similarly, during the back-propagation phase, if the image is RGB only, the loss is back-propagated only to the Deeplab network component. If the image is HHA only, the loss is back-propagated only to the Alexnet-HHA component. Otherwise, the loss is backpropagated through both components of the network. In our experiments, we demonstrate the utility of this layer, by using images of outdoor scenes from RGB only INRIA person database to augment our training set with more background examples.

5. Experiments

We use our method for the task of semantic segmentation of human limbs. All experiments are conducted on the Human-Limb Dataset, which is described in detail in section 3. We compare four different setups in our empirical evaluation, (1) Deeplab network using RGB images, (2) Alexnet-HHA network using HHA images, (3) Deeplab+Alexnet-HHA network (described in section 4.3) using RGB+HHA images, and (4) Deeplab+Alexnet-HHA network using RGB+HHA images from our dataset alongside 1218 RGB images from the INRIA person database, containing outdoor images for additional background examples. The last method uses the *adaptiveConcatenation* layer introduced in section 4.4.

While training/testing all networks, we resize input images to 242×322 pixels. Since we have 7 labels (6 human body parts + 1 background label), our networks are designed to output 7 scores per pixel, one corresponding to each label. As pointed out in sections 4.1, 4.2, our networks have a downsampling factor of 8. Due to this downsampling of spatial dimensions, the size of the final output score map is $31 \times 41 \times 7$. Thus, during training we also downsample the ground truth label maps to 31×41 pixels by *nearest neighbour* interpolation, to avoid averaging artifacts. At test time, we resize the output scores to the original image size of 242×322 pixels via *bilinear interpolation*. To get the labels, we compute the softmax probabilities of each label from the scores, and pick the label with the highest probability. We also employ dense CRF to refine segmentations provided by our networks, as in [6]. The dense CRF framework serves as a post-processing step on our networks' output scores, before computing the softmax-probabilities for class labels.

We report quantitative results of our evaluation in table 5. We use the accuracy of segmentation as the evaluation metric for our experiments. In table 5, we report results both with and without the dense CRF. Our results indicate that using dense CRF improves results, boosting the networks' capability to capture fine details. Figure 9 shows some qualitative results on the test images.

Our implementation is built on top of the *Caffe* library, and our testing time for RGB+HHA is about 8 frames-per-second on an NVIDIA-Tesla K40 GPU.

Method	Softmax	Dense CRF
RGB	86.62	87.26
HHA	83.09	84.94
RGB+HHA	86.99	87.45
(RGB + HHA) + INRIA	**88.12**	**88.84**

Table 1. Segmentation Accuracies of our system for the four combinations of input types. We also show the effect of post-processing the scores using dense CRF on the performance of these methods.

6. Discussion

In this paper we presented a method for combining information from RGB images and depth images to train a system for semantic segmentation of human body parts. Our system is flexible, allowing us to take advantage of the two types of inputs, or fall back on using just one type, when the other is not available. We also introduced a novel set of body part annotations, and used it to evaluate the performance of our approach. We demonstrate promising results and, as a future direction, we plan to explore the application of body-part semantic segmentation for action recognition and pose estimation.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *PAMI*, 2011.

[2] D. Banica and C. Sminchisescu. Second-order constrained parametric proposals and sequential search-based structured prediction for semantic segmentation in rgb-d images. In *CVPR*, pages 3517–3526, 2015.

[3] Y. Bo and C. C. Fowlkes. Shape-based pedestrian parsing. In *CVPR*, pages 2265–2272. IEEE, 2011.

[4] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *TPAMI*, 34(7):1312–1328, 2012.

[5] U. Census. The elderly population. 2010.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[7] C. Couprie, C. Farabet, L. Najman, and Y. LeCun. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013.

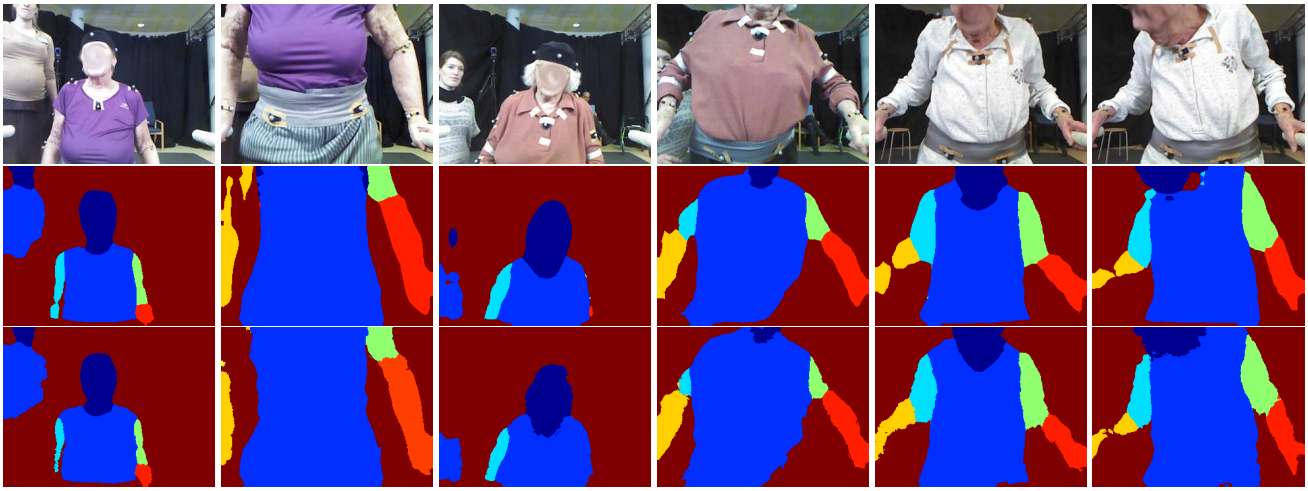

Figure 9. Segmentation Results of our RGB network on unseen images. The first row shows the rgb images, the second row shows the segmentation results of our network, and the third row shows the results after using dense CRF. Best viewed in colour.

[8] S. Eslami and C. Williams. A generative model for parts-based object segmentation. In *NIPS*, 2012.

[9] S. A. Eslami, N. Heess, C. K. Williams, and J. Winn. The shape boltzmann machine: a strong model of object shape. *IJCV*, 107(2):155–176, 2014.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. PAMI*, 2010.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587. IEEE, 2014.

[12] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 564–571. IEEE, 2013.

[13] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *Computer Vision–ECCV 2014*, pages 345–360. Springer, 2014.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[15] Y. Jin and S. Geman. Context and hierarchy in a probabilistic image model. In *CVPR*, volume 2, pages 2145–2152. IEEE, 2006.

[16] I. Kokkinos and A. Yuille. Inference and learning with hierarchical shape models. *International Journal of Computer Vision*, 93(2):201–225, 2011.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[18] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng. Ica with reconstruction cost for efficient overcomplete feature learning. In *Advances in Neural Information Processing Systems*, pages 1017–1025, 2011.

[19] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *arXiv preprint arXiv:1411.4038*, 2014.

[20] W. Lu, X. Lian, and A. Yuille. Parsing semantic parts of cars using graphical models and segment appearance consistency. *BMVC*, 2014.

[21] X. S. Papageorgiou, C. S. Tzafestas, P. Maragos, G. Pavlakos, G. Chalvatzaki, G. Moustris, I. Kokkinos, A. Peer, B. Stanczyk, E.-S. Fotinea, and E. Efthimiou. Advances in intelligent mobility assistance robot integrating multimodal sensory processing. *Universal Access in Human-Computer Interaction. Aging and Assistive Environments*, 2010.

[22] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2759–2766. IEEE, 2012.

[23] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[24] A. Wang, J. Lu, G. Wang, J. Cai, and T.-J. Cham. Multi-modal unsupervised feature learning for rgb-d scene labeling. In *ECCV*, pages 453–467. Springer, 2014.

[25] J. Wang and A. Yuille. Semantic part segmentation using compositional model combining shape and appearance. In *CVPR*, 2015.

[26] L. Zhu, Y. Chen, and A. Yuille. Learning a hierarchical deformable template for rapid deformable object parsing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(6):1029–1043, 2010.

[27] S.-C. Zhu and D. Mumford. *A stochastic grammar of images*. Now Publishers Inc, 2007.