NRC·CNRC

NRC Publications Archive
Archives des publications du CNRC

Increasing the robustness of CNN-based human body segmentation in range images by modeling sensor-specific artifacts

Seoud, Lama; Boisvert, Jonathan; Drouin, Marc-Antoine; Picard, Michel; Godin, Guy

For the publisher's version, please access the DOI link below./ Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

Publisher's version / Version de l'éditeur:
https://doi.org/10.1007/978-3-030-11015-4_55
Computer Vision – ECCV 2018 Workshops, pp. 729-743, 2019-01-23

NRC Publications Archive Record / Notice des Archives des publications du CNRC :
https://nrc-publications.canada.ca/eng/view/object/?id=4153a938-f232-4a90-a208-70b3848fef9c
https://publications-cnrc.canada.ca/fra/voir/objet/?id=4153a938-f232-4a90-a208-70b3848fef9c

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at
https://nrc-publications.canada.ca/eng/copyright
READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site
https://publications-cnrc.canada.ca/fra/droits
LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

National Research Council Canada
Conseil national de recherches Canada

Canada

This ECCV 2018 workshop paper, provided here by the Computer Vision Foundation, is the author-created version. The content of this paper is identical to the content of the officially published ECCV 2018 LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/eccv

Increasing the robustness of CNN-based human body segmentation in range images by modeling sensor-specific artifacts

Lama Seoud, Jonathan Boisvert, Marc-Antoine Drouin, Michel Picard, and Guy Godin

National Research Council, 1200 Montreal Road, Ottawa, ON, K1A 0R6, Canada
{firstname.lastname}@nrc-cnrc.gc.ca

Abstract. This paper addresses the problem of human body parts segmentation in range images acquired using a structured-light imaging system. We propose a solution based on a fully convolutional neural network trained on realistic synthetic data that were simulated in a way that closely emulates our structured-light imaging system with its inherent artifacts such as occlusions, noise and missing data. The results on synthetic test data demonstrate quantitatively the performance of our method in identifying 33 body parts, with negligible confusion between the front and back sides of the body and between the left and right limbs. Our experiments highlight the importance of sensor-specific data augmentation in the training set to improve the robustness of the segmentation. Most importantly, when applied to range data actually acquired by our system, the method was capable of accurately segmenting the different body parts with inter-frame consistency in real-time.

Keywords: Human body segmentation, structured-light imaging, convolutional neural network

1 Introduction

Despite a long history of publications on the matter, human pose estimation and human body segmentation is still a challenging research subject. Challenges come from the large variations in pose, shape, viewpoint, lighting and clothing. Nevertheless, it is a key step in human motion analysis which finds application in a large variety of fields. In advanced manufacturing, robots or machines need postural information on the human they are interacting with in order to collaborate safely and effectively. In healthcare, physiotherapy can be performed remotely by examining the kinematics recorded by a marker-less vision system while a patient is at home doing his/her exercises.

Range sensors have drawn much interest for human activity related research since they provide explicit 3D information about the shape, and that is invariant to clothing color, skin color and illumination changes compared to RGB cameras, and facilitates background subtraction. Among the existing range sensing technologies, structured-light sensors offer the advantages of high resolution

combined with high speed, compared to time-of-flight cameras or stereoscopic reconstruction systems, making it more practical for real-time applications. Additionally, structured-light systems are typically more affordable.

However, triangulation-based systems (which include structured-light sensors) generate shadows or occlusions in the image when parts of the scene cannot be seen by both the projector and the camera. Those occlusions depend on the shape of the object being imaged, but also on the structured-light system design characteristics (distance between projector and camera, the triangulation angle, lens focals, *etc.*). Missing points and measurement noise are also dependent both on the object characteristics and on the design of the 3D measuring system. For instance, rapid movements and/or dark or patterned clothes may generate holes or missing data in the images. These artifacts inherent to this kind of sensor add a level of difficulty to the task of human body segmentation that is typically not addressed in the literature.

In this work, we address the problem of real-time human body segmentation from range images acquired by a high resolution structured-light imaging system. The challenge toward this goal is to design a segmentation model that is able to reason about 3D spatial information and, at the same time, is robust to artifacts inherent to the structured-light system, such as occlusions or triangulation shadows, noise and missing data.

To address this challenge, our contributions are as follows. First, we propose a domain-specific data augmentation strategy that closely simulates the actual acquisition scenario with the same intrinsic parameters as our sensor and the artifacts it generates. Second, we adapt the fully convolutional network of [20] to range images of the human body in order for it to transfer its learning toward 3D spatial information instead of light intensities. Third, we quantitatively demonstrate the importance of simulating sensor-specific artifacts in the training set to improve the robustness of the segmentation of actual range images.

2 Related Work

Most previous work on human pose estimation use as input either a single RGB image or a RGB video sequence [5,9,15,19,23,24]. Even though these images contain rich information, the sensitivity of RGB sensors to illumination changes and the presence of texture that interferes with geometric features affect the robustness of RGB image-based human pose estimation. With the advent and wide accessibility of range sensors, research on range image-based or RGBD-based (color and depth) methods [7,8,16,21,2] has become very active.

Furthermore, the literature related to human body pose estimation can be categorized into generative and discriminative methods. Generative methods consist in fitting a human body shape template or prior to the input data points, using some optimization procedures, making them considerably time-consuming. Point clouds obtained by range sensing motivate the use of different variants of the iterative closest point algorithm, such as in [7]. A subclass of generative approaches groups methods that use part-based models where the human body is

represented as a skeleton with different body parts connected by joints-imposed constraints or kinematics constraints, such as the popular pictorial structural model [5,19] or some Markov Random Field based graphical model to impose spatial constraints [23].

On the other hand, discriminative methods consist in directly identifying a mapping between the input image and the body parts or joints. Among these methods, some aim at detecting the joints by regression methods [22,24], or identifying and classifying interest points [16], or segmenting the body into its individual parts using a pixel-level classification [2,11,15,21]. Because it does not take into account the kinematic properties of the human body configuration, those methods may result in incoherent body parts segmentation. Nevertheless, machine learning approaches, either random forests [21,22] or deep convolutional neural networks [2,8,9,11,15,24] have proved that, with sufficiently large training datasets, the global distribution of body parts is somehow implicitly learned by the classification model. The biggest advantage of discriminative methods over model-based method is in execution time, making them more suitable for real-time applications.

Machine learning approaches rely heavily on the size and quality of the training data. Several papers have addressed the limited availability of segmented range and RGB images of the human body and proposed datasets of synthetic training data [4,14,21,25,27]. Generally, synthetic range images are generated using some motion capture sequences with retargeting of different body shapes and standard computer graphics techniques. While authors emphasize the importance of having a variety of shapes and poses, they tend to neglect the artifacts introduced by actual range sensing systems, limiting the generalization performance on real data.

However, a better modeling of the sensor characteristics has already been shown to improve performances of a different learning task applied to RGBD image. For instance, Planche et al. [17] demonstrated improved performances in the determination of the position and orientation of isolated rigid objects using this kind of approach. It should therefore provide gains for body segmentation methods as well since those sensor characteristics can significantly alter the appearance of body parts in range images. Furthermore, the complex relationship between occlusions and the changing shape of a deformable object (such as the human body) might provide insight that a CNN (Convolutional Neural Network) can use to boost performances.

3 Method

In this section, we first describe our imaging system based on structured light, the configuration of which is then used to synthetically generate realistic range images from existing 3D human body meshes. For those synthetic images to be realistic, we define and simulate occlusions, noise and missing data. The labeling of the different body parts is then described, followed by the deep neural network configuration and its training. Finally, we provide an overview of our

experimental setup for the evaluation of our method on both synthetic and real data.

3.1 Structured-light imaging system

Our structured-light imaging system [6] uses a high-resolution projector and a Emergent Vision Technologies camera working at 360 fps. The standoff distance is 2.5 m and the system baseline is 0.75m. The system was designed to cover the volume of an adult performing large amplitude movements. The focal lengths are respectively 12 and 12.5 mm for the projection and collection lenses. The lateral resolution of the system is 1mm and the range uncertainty is sub-millimetric.

In our experiments, we configured the system to use 5 phase shift patterns for range measurements and 7 binary patterns for the phase unwrapping. The system can generate 2M points range images at 30Hz. The configuration of our system is illustrated in Figure 1.

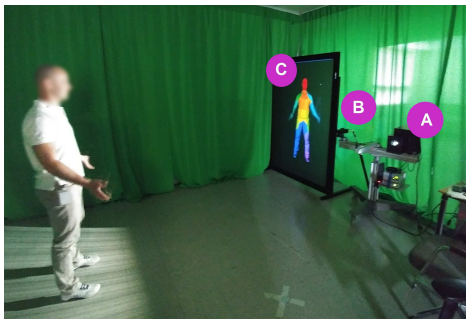

Fig. 1. Set-up of our 3D human body imaging system with the structured-light projector (A), the camera (B) and the real-time body part segmentation projected on a large screen (C).

The size of the resulting range images is 1920 x 988 pixels, however, in this work, we down-sampled the images by a factor of 3 for a faster training.

3.2 Building a realistic set of synthetic data

To generate training data for our model, we used a methodology inspired from the work of [27] but with a particular emphasis on replicating the output of our structured-light imaging system.

We collected 3D meshes of the human body from 3 publicly available datasets: SCAPE [1], MIT [26] and CAESAR [18]. The first one consists of 71 meshes of the same unclothed subject (labeled as A) in different poses. The second one consists of 2 clothed subjects (labeled as B and C) in 4 different motion sequences each

(walking, jumping, crane, squatting, bouncing and hand-standing), for a total of 825 and 850 meshes for B and C respectively. The third dataset comprises the meshes of 583 minimally clothed subjects in the same canonical posture.

Posture variability is covered by SCAPE and MIT meshes, while inter-subject variability is mostly covered by CAESAR meshes. Having both clothed and unclothed subjects adds a level of invariance to clothing.

For each 3D mesh, we simulated 54 different range images: the mesh is placed iteratively at 2.15m, 2.55m, and 2.95m from the camera along its focal axis and for each position, the mesh is rendered from 18 different viewpoints around the model (10° rotation between each pair of consecutive views). This enforces an invariance to distance and viewpoint.

In a real acquisition system, range images are affected by artifacts proper to the imaging system itself. Using a structured-light sensor, light occlusions are present when part of the scene is not illuminated by the projector (Figure 2). On top of that, the acquisition can be affected by noisy range values resulting from the calibration of the system. Finally, because of dark patterned clothes and/or rapid movements, some data might be missing in the range image.

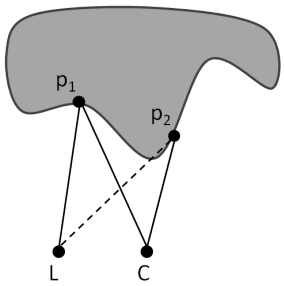

Fig. 2. Light occlusions : top view of a structured-light imaging system with the light projector (L) and the camera (C). Point p_1 is illuminated by L and seen by C, thus it rendered as a foreground pixel in the range image. Point p_2 is seen by C but not illuminated by L, this occlusion results in a background valued pixel in the range image.

In order for our classification model to be robust to those inherent artifacts, we simulated the presence of occlusions, missing data, and measurement noise in our training:

- Light occlusions modeling: For structured-light specific occlusions, we considered two frame buffers, one that emulates the light projector and the other the camera, and we considered only the pixels that are rendered in both buffers.
- Measurement noise modeling: To simulate range measurement noise, additive Gaussian noise is added to the foreground pixels. For each image a standard

deviation is randomly selected between 0 mm (noise) and 100 mm (which is considerably more then the expected noise on our real data).

– Missing data modeling: By missing data in range images, we mean pixels that do not correspond to any light occlusions, but whose value was two erroneous due to rapid motion or dark patterned clothes and thus discarded from the range image. To model those "missing data" in a given range image, we randomly removed a random percentage of pixels from the foreground and replaced it by background values. The pixel removal rate is randomly chosen between 0%, 5% and 50%.

To evaluate the effect of each of these artifacts, we created five variations of the dataset, each one modeling a different combination of artifacts (see Table 1).

Table 1. Description of differences between the datasets used in the experiments.

Dataset	Description
X_{none}	no sensor-specific artifacts modeled
X_{occ}	only occlusions modeled
X_{noise}	only measurement noise modeled
X_{md}	only missing data modeled
X_{all}	occlusions, measurement noise and missing data modeled

In total, we generated 5 sets of 125 766 range images of size 640x329 pixels with the range encoded on 16 bits. Figure 3(top row) illustrates some examples of the simulated synthetic data.

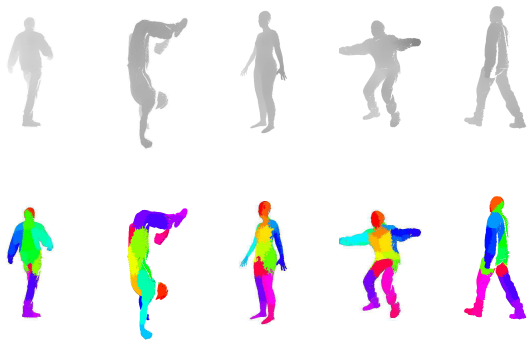

Fig. 3. Examples of simulated synthetic range images (top row) with their corresponding segmentation into 33 body parts (bottom row). Each color correspond to a distinct body part. Note the presence of occlusions (for instance the self-occlusion of the left leg in the first column or the disconnected leg and head in the second column).

3.3 Annotating the data

To annotate the different body parts in the range images, we used anatomical landmarks identified as salient points on the 3D meshes. For the meshes in CAESAR, the coordinates of 33 anatomical landmarks are provided with each mesh. For SCAPE and MIT, we manually identified those same landmarks on one mesh for each of the subjects A, B and C; we then used the fact that the meshes of each subject in different postures share the same topology to propagate those landmarks coordinates to the remaining meshes of these subjects.

To define body parts from those anatomical landmarks, we used the fast marching closest neighbor algorithm [12] that aggregates the vertices that are the closest to the landmarks in terms of geodesic distance on the triangular mesh. For each landmark or resulting body part, a distinct label and color are attributed.

Finally, for each simulated range map, an image of the same size is generated with the corresponding labels encoded on 8 bits. The background is set to an arbitrary value that will be ignored in the remaining process. Some examples of segmented images are provided in Figure 3(bottom row).

3.4 Network description

We approach the problem of body parts tracking from range images as a pixel classification problem. Thus, to perform the segmentation, we defined a dense fully convolutional neural network [20] derived from the Alexnet [13]. The architecture of our network is detailed in Figure 4.

To do so, we first removed the final classification layer of the Alexnet and replaced all the fully connected layers of the Alexnet by convolutional layers with a stride of 1 and a kernel size of 1. At the end of the network, before the final classification, we added *conv8*, a 1x1 convolutional layer with 16 outputs in order to extract for each pixel a 16D features vector and *upfeat*, a backward convolution (deconvolution) layer to bilinearly up-sample the coarse outputs to pixel-wise outputs that are the same size as the input image.

Finally, the last convolutional layer, *score*, provides for each pixel 33 outputs corresponding to the 33 body parts considered in our segmentation problem. Also, because the original Alexnet aims at classifying 3-channel RGB images, we had to modify the first convolutional layer, *conv1* to adapt it so that it takes only one channel (the range).

3.5 Network training

From each of the 5 datasets of simulated range images, 97 146 images are selected as follows and used for training the network : for subject A, we used the first 57 postures for training and the remaining 14 postures for testing, for subjects B and C, we used the jumping and the squatting sequences respectively for testing and the remaining for training, for the CAESAR dataset, we used the first 467 subjects for training and the remaining for the test. This way, we ensure that

Layers	Parameters	Weight initialisation
score	Convolution (k=1, s=1, o=33)	Gaussian
upfeat	Deconvolution (k=63, s=16, o=16) + Crop	Bilinear
conv8	Convolution (k=1, s=1, o=16) + ReLU	Gaussian
conv7	Convolution (k=1, s=1, o=4096) + ReLU + Dropout	Adapted from Alexnet
conv6	Convolution (k=6, s=1, o=4096) + ReLU + Dropout	Adapted from Alexnet
pool5	Pooling (max, k=3, s=2)	
conv5	Convolution (k=3, s=1, o=256) + ReLU	Alexnet
conv4	Convolution (k=3, s=1, o=384) + ReLU	Alexnet
conv3	Convolution (k=3, s=1, o=384) + ReLU	Alexnet
pool2	Pooling (max, k=3, s=2)	
conv2	Convolution (k=5, s=1, o=256) + ReLU	Alexnet
pool1	Pooling (max, k=3, s=2)	
conv1	Convolution (k=11, s=4, o=96) + ReLU	Adapted from Alexnet

Fig. 4. Architecture of the fully-convolutional neural network trained for body parts segmentation. Starting from the deepest layers at the bottom and going up to shallower layers. k: kernel size, s: stride, o: number of outputs

all the range images generated from one mesh are considered together either for training or for testing the network.

Instead of training our network from scratch, we performed a transfer learning. We started the training with the weights of the Alexnet and we fine-tuned it by applying a gradient of learning rates: for the deepest layers corresponding to more generic filters, we applied a smaller learning rate (0.01 times the learning rate of the last layer), and for the shallowest layers corresponding to application specific features, we gradually increased the learning rate.

For the first layer of our network, *conv1*, we averaged the original weights of the 3-channels in the Alexnet to generate 1-channel filters. And for the additional layers that are not part of the original Alexnet, we initialized the weights using a Gaussian distribution with a standard deviation of 0.01. The deconvolution layer implements a bilinear up-sampling filter and its weights are kept frozen during the training.

The training is performed using the stochastic gradient descent, with a fixed last layer learning rate of 10^{-2}, a momentum of 0.99 and a weight decay of 5×10^{-4}. At each iteration, we used a gradient accumulation across 20 images. We run the training over 70 000 iterations. A softmax loss layer computes the cost function while ignoring all the pixels that belong to the background of the input image.

We used the Caffe framework [10] to implement the training on an 12GB NVIDIA GeForce GTX TITAN X GPU.

3.6 Experimental setup

We performed a two-step evaluation of the proposed method. First, to quantify the classification accuracy on simulated data, we considered the remaining 28 620 images of the X_{all} dataset as our synthetic test set. To weigh each body part equally despite their varying sizes, the accuracy is reported as the average per-class segmentation accuracy, as in [21]. Test images always include random additive noise, missing data, and occlusions. We also evaluate the relationship between the confidence probability associated with each classified pixel (obtained by softmax) and the classification accuracy.

In a second step, we applied our method to real data sequences acquired with our structured-light imaging system and we qualitatively evaluated the inter-frame consistency, considering that the processing is performed independently on each frame. We also report the overall processing time.

4 Results and discussion

In this section, we report and analyze the results obtained on both the synthetic test dataset and the real data acquired by our structured-light 3D imaging system. We also evaluate the robustness to noise and missing data as well as the computational performance of the proposed method.

4.1 Results on the synthetic test set

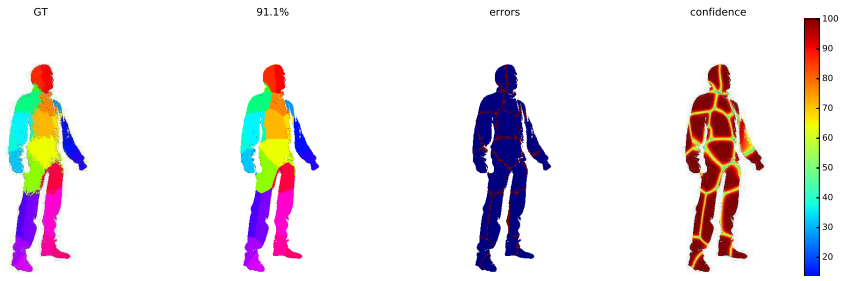

Fig. 5. Confidence in the segmentation. From left to right, the ground truth segmentation (GT), the result of the CNN segmentation, the error image (blue for correct, red for error in the segmentation) and the confidence image computed as a probability using the softmax function and expressed in percentage.

Over the entire synthetic test set, we report a global average per-class segmentation accuracy of 81.3% when the network is trained using the X_{all} dataset. Figure 5 provide an example of the results. Most of the errors are located at the edges of the segmentation. This is due in part to the ground truth annotations. In

fact, when rendering the annotated faces of the meshes, the 2D projection creates triangular patterns at the edges of the body parts. However, when computing the confidence in the segmentation using the softmax function at the output of the network for a random test image (Figure 5), we see that the confidence on the edges of the segmentation is the lowest. Thus, if we apply a threshold of 70% on the confidence values, we get an average per-class segmentation accuracy of 97.8% for this image instead of 91.1% , while discarding only 18% of the pixels in the test set.

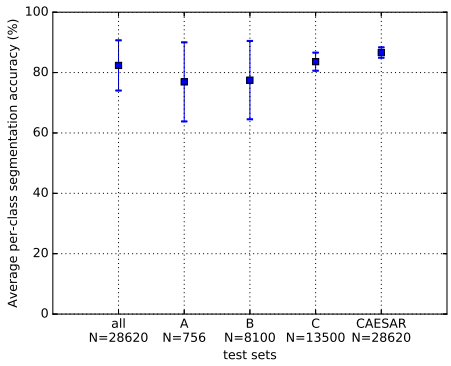

Fig. 6. Average ± one standard deviation per-class segmentation accuracy for the whole test set and the different subsets.

Figure 6 reports the average per-class segmentation accuracy for the whole test set and for the test subsets corresponding to subjects A, B, C separately and the CAESAR test set, as well as the per-class accuracy on the whole test set. The best results are obtained on the CAESAR test set where the variation is essentially in the human shape, the pose being similar for all subjects in the database. On the contrary, the segmentation accuracy for the test set from SCAPE is only of 78.1%, indicating that the model is more robust to inter-subject shape variations than to intra-subject pose variations.

Unfortunately, because there is currently no unified benchmark for the anatomic segmentation of full human bodies from range images, it is impossible to have a fair comparison to previous work. Still, for the sake of comparison, an average per-class segmentation accuracy of 60% was achieved at best using handcrafted features and randomized decision forests [21], bearing in mind that their dataset and their classes are different.

4.2 Robustness to noise

By adding Gaussian noise randomly on the synthetic test images, we evaluated the robustness of our network to the presence of noise. Figure 7 illustrates

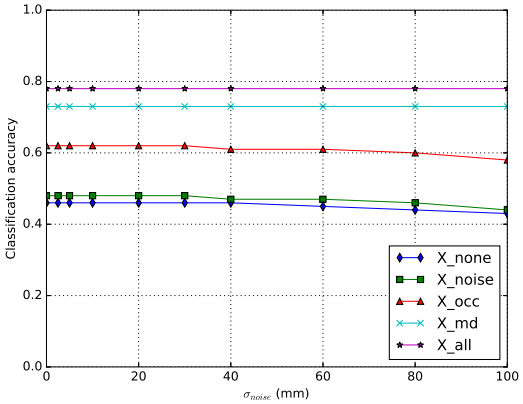

Fig. 7. Robustness to Gaussian noise reported as the average of the segmentation accuracy evaluated on the synthetic test set based on training with each of the five training set variations (see Table 1 for details).

the average segmentation accuracy computed on the whole test set for different amounts of simulated Gaussian noise. This result clearly demonstrates that the system is highly robust to Gaussian noise in the range maps, even when its standard deviation reaches 40 mm, which is significantly larger than the range accuracy of our system (<1mm). This robustness to noise is implicitly embedded in the architecture of the CNN itself, particularly at the pooling layers. Furthermore, it appears that simulating sensor-specific artifacts improved performances at varying levels with missing data simulation and occlusions providing the greatest individual gains.

4.3 Robustness to missing data

By randomly removing data from the range images in the synthetic test set, we evaluated the robustness of our network to missing data. We trained the network with five variants of the training set presented earlier (see Table 1). Figure 8 illustrates the segmentation accuracy computed on the whole test set for different amounts of missing data. This result clearly demonstrates the importance of simulating missing data and occlusions (to a lesser degree) during the training to increase the robustness of the network. To our knowledge, no previous work has analyzed the effects of missing data on the human body parts segmentation from range and/or RGB images, even though robustness to missing data is an important feature especially when dealing with real range image acquisition where rapid movements and/or dark clothes can generate holes in the image.

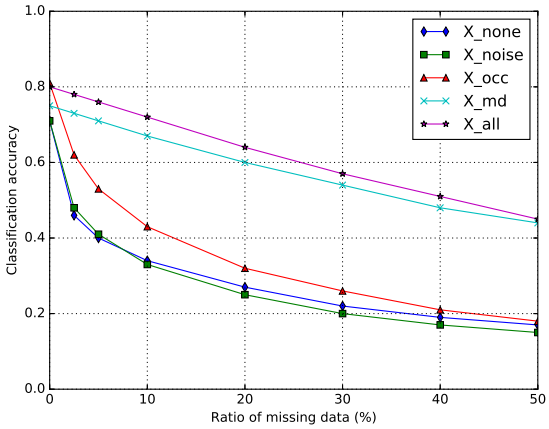

Fig. 8. Robustness to random missing data, reported as the average of the segmentation accuracy evaluated on the synthetic test set. Training was performed using the five training set variations (see Table 1 for details).

4.4 Results on a real data sequence

The results on the synthetic datasets demonstrate quantitatively the performance of the proposed segmentation network. We are also interested in evaluating qualitatively the performance in a real scenario, with real range images acquired using our own structured-light imaging system. The effect of considering the sensor-specific artifacts on real data in practice is illustrated in Figure 9 for 3 range images acquired with our structured-light imaging system. Qualitatively, the worst segmentation results are obtained when no occlusions and no missing data are considered in the training set. These results on real test data emphasize even more the importance of a good modeling of the sensor's characteristics in the simulation of training data, and hence in the training of the CNN.

A video illustrating the real-time body parts segmentation on a live full motion sequence is available at : https://youtu.be/2aEbHqwKlmg. The subject performs successively a jump, a squat, two 360° rotations, and ends in a final posture with crossing arms on the chest. Qualitatively, the different body parts are accurately segmented in most frames with remarkably little jitter. Even in the two full 360° rotations, the performance in differentiating the front and the back sides of the body, as well as the left and right limbs is very satisfactory, discarding the need for a tracking algorithm as opposed to the conclusions in [21]. During the full rotations, when the subject is perpendicular to the baseline of the system, the arms are correctly segmented even though there is self-occlusion with the rest of the body. However, one evident segmentation error is noted in the crossed arms final posture where the forearms are vanishing in the chest in the segmented images. This particular failure mode has also been raised in

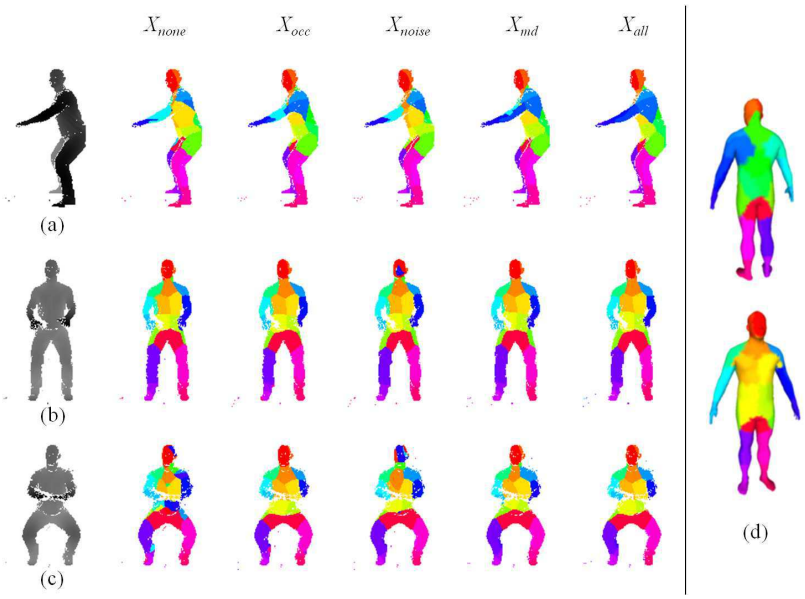

Fig. 9. For each image (a, b and c), the results of the segmentation using the CNN trained on each of the 5 training set variations (see Table 1 for details) are illustrated from left to right respectively. In the absence of ground truth for those actual range images, a template body segmentation is provided on the right (d) to qualitatively compare the segmentation results.

previous work [21] and we believe that this is partly due to the lack of similar postures in the training set.

4.5 Processing time

We evaluated the performance of the network on the same hardware as the training, an 12GB NVIDIA GeForce GTX TITAN X GPU. We recorded the processing time required to perform a segmentation, considering the data already loaded on the memory. We recorded an average of 61 ms ($\pm$ 1.2 ms) on the whole test set with the average image size being 640x329 pixels. This results shows that method would be suitable for real-time applications.

5 Conclusion

We presented in this paper a deep learning approach for human body parts segmentation from range images. Not only did it yield semantically accurate results in synthetic test data, but we demonstrated its performance in a real scenario with images acquired with a high-resolution structured-light imaging system. We have also demonstrated the importance of having a realistic sensor-specific training set to improve the robustness of the segmentation to artifacts such as occlusions, noise and missing data which affect the range images acquired by a structured-light system in particular.

The proposed data-augmentation strategy is specific to structured-light imaging systems. Of course, depending on the acquisition system, sensor-specific artifacts are quite different. Time-of-flight sensors, for example, suffer from the multiple paths problem, whereas passive stereo systems deal with non-uniform noise depending on the texture of the object being imaged.

Compared to previous work, we considered a segmentation of the body into 33 parts, which is a finer granularity than most of the previous work [2,8,11,15]. In the future, we aim at further refining the granularity to identify a dense mapping between the range images and a 3D body template.

Among the remaining challenges are the self-occlusions, such as the crossed-arms posture and the postures that are very different than the one used for the training. However, since our objective is to segment image sequences, it would be interesting to investigate some temporal constraints to regularize the segmentation especially for the frames where the posture is unseen by the network.

In its current implementation, the segmentation of a single frame of size 640x329 requires around 60 ms which is suitable for almost real-time applications. In future work, we plan on using the full resolution images from our high-resolution sensor (1920 x 988), which will require more attention to the processing time. The combination of high resolution, high accuracy and high speed of our acquisition system and our segmentation module will open the door to the analysis of tiny and rapid movements, which is currently a challenge for existing commercial range sensors [3].

References

1. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: Shape Completion and Animation of People. In: ACM Transactions on Graphics - Proceedings of ACM SIGGRAPH 2005. pp. 408–416 (2005)
2. Chandra, S., Tsogkas, S., Kokkinos, I.: Accurate Human-Limb Segmentation in RGB-D images for Intelligent Mobility Assistance Robots. In: IEEE International Conference on Computer Vision (ICCV). pp. 44–50 (2015)
3. Chen, L., Wei, H., Ferryman, J.: A survey of human motion analysis using depth imagery. Pattern Recognition Letters **34**(15), 1995–2006 (2013)
4. Chen, W., Tu, C., Su, H., Lischinski, D., Cohen-or, D., Wang, Z., Chen, B.: Synthesizing Training Images for Boosting Human 3D Pose Estimation. In: International Conference on 3D Vision. pp. 479–488 (2016)
5. Dantone, M., Gall, J., Leistner, C.: Human Pose Estimation using Body Parts Dependent Joint Regressors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3041–3048 (2013)
6. Drouin, M.A., Blais, F., Godin, G.: High Resolution Projector for 3D Imaging. In: International Conference on 3D Vision (3DV). vol. 1, pp. 337–344 (2014)
7. Ganapathi, V., Plagemann, C., Koller, D., Thrun, S.: Real-Time Human Pose Tracking from Range Data. In: European Conference on Computer Vision (ECCV). pp. 738–751 (2012)
8. Haque, A., Peng, B., Luo, Z., Alahi, A., Yeung, S., Fei-Fei, L.: Towards viewpoint invariant 3D human pose estimation. In: European Conference on Computer Vision (ECCV). pp. 160–177 (2016)
9. Jain, A., Tompson, J., LeCun, Y., Bregler, C.: MoDeep: A Deep Learning Framework Using Motion Features for Human Pose Estimation. In: Asian Conference on Computer Vision (ACCV). pp. 302–315 (2014)
10. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional Architecture for Fast Feature Embedding. In: ACM international conference on Multimedia. pp. 675–678 (2014)
11. Jiu, M., Wolf, C., Taylor, G., Baskurt, A.: Human body part estimation from depth images via spatially-constrained deep learning. Pattern Recognition Letters **50**, 122–129 (2014)
12. Kimmel, R., Sethian, J.A.: Computing geodesic paths on manifolds. In: Proceedings of the National Academy of Sciences of the United States of America. vol. 95, pp. 8431–8435 (1998)
13. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet Classification with Deep Convolutional Neural Networks. In: Advances in neural information processing systems (NIPS). pp. 1097–1105 (2012)
14. Nishi, K., Miura, J.: Generation of human depth images with body part labels for complex human pose recognition. Pattern Recognition **71**, 402–413 (2017)
15. Oliveira, G.L., Valada, A., Bollen, C., Burgard, W., Brox, T.: Deep learning for human part discovery in images. In: Proceedings - IEEE International Conference on Robotics and Automation. pp. 1634–1641 (2016)
16. Plagemann, C., Ganapathi, V., Koller, D., Thrun, S.: Real-time identification and localization of body parts from depth images. In: Proceedings - IEEE International Conference on Robotics and Automation. pp. 3108–3113 (2010)
17. Planche, B., Wu, Z., Ma, K., Sun, S., Kluckner, S., Lehmann, O., Chen, T., Hutter, A., Zakharov, S., Kosch, H., Ernst, J.: DepthSynth: Real-time realistic synthetic data generation from CAD models for 2.5D recognition. In: International Conference on 3D Vision (3DV) (2017)

18. Robinette, K.M., Daanen, H., Paquet, E.: The CAESAR project: a 3-D surface anthropometry survey. In: Second International Conference on 3D Digital Imaging and Modeling, 3DIM. pp. 380–386 (1999)
19. Sapp, B., Taskar, B.: MODEC : Multimodal Decomposable Models for Human Pose Estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3674–3681 (2013)
20. Shelhamer, E., Long, J., Darrell, T.: Fully Convolutional Networks for Semantic Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(4), 640–651 (2016)
21. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Communications of the ACM **56**(1), 119–135 (2013)
22. Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., Blake, A.: Efficient Human Pose Estimation from Single Depth Images. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(12), 2821–2840 (2013)
23. Tompson, J., Jain, A., LeCun, Y., Bregler, C.: Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. In: Advances in neural information processing systems (NIPS). pp. 1799–1807 (2014)
24. Toshev, A., Szegedy, C.: DeepPose: Human Pose Estimation via Deep Neural Networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1653–1660 (2014)
25. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from Synthetic Humans. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
26. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. ACM Transactions on Graphics **27**(3), 1–9 (2008)
27. Wei, L., Huang, Q., Ceylan, D., Vouga, E., Li, H.: Dense Human Body Correspondences Using Convolutional Networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1544–1553 (2016)