Human Recognition in RGBD Combining Object Detectors and Conditional Random Fields

Konstantinos Amplianitis[1], Ronny Hänsch[2] and Ralf Reulke[1]
[1]*Computer Vision Group, Humboldt Universität zu Berlin, Rudower Chaussee 25, 12489 Berlin, Germany*
[2]*Computer Vision and Remote Sensing Group, Technische Universität Berlin, Marchstr. 23, 10587 Berlin, Germany*

Keywords: Deformable Part Models, RGBD Data, Conditional Random Fields, Graph Cuts, Human Recognition.

Abstract: This paper addresses the problem of detecting and segmenting human instances in a point cloud. Both fields have been well studied during the last decades showing impressive results, not only in accuracy but also in computational performance. With the rapid use of depth sensors, a resurgent need for improving existing state-of-the-art algorithms, integrating depth information as an additional constraint became more ostensible. Current challenges involve combining RGB and depth information for reasoning about location and spatial extent of the object of interest. We make use of an improved deformable part model algorithm, allowing to deform the individual parts across multiple scales, approximating the location of the person in the scene and a conditional random field energy function for specifying the object's spatial extent. Our proposed energy function models up to pairwise relations defined in the RGBD domain, enforcing label consistency for regions sharing similar unary and pairwise measurements. Experimental results show that our proposed energy function provides a fairly precise segmentation even when the resulting detection box is imprecise. Reasoning about the detection algorithm could potentially enhance the quality of the detection box allowing capturing the object of interest as a whole.

1 INTRODUCTION

Existing work on RGB and RGBD space has concentrated more on semantic segmentation and scene understanding, i.e. which pixel/voxel should be assigned to which object label. In spite of the fact that even though this is an interesting field of research, many applications require additional knowledge of the dynamic, moving objects in the scene. This procedure requires information about an approximate *location* and *spatial extent* of these objects in the scene. For the human vision, perception has a physical meaning, as it involves a natural process performed in the cerebral cortex of the brain. For computer vision, this natural process was scientifically treated as two separate and independent tasks. As was mentioned by (Hariharan et al., 2014), neither the detection box nor object regions can produce a compelling output representation but rather they should be able to complement each other.

The availability of commodity depth sensors such as the Kinect, paved the way for researchers to improve the state-of-the-art detection and segmentation algorithms and increase the richness of available information. In this paper, we try to bring together detection and segmentation as a recognition task, to infer about the location and spatial extent of an object in RGBD space. Localization is performed using an accelerated version of the deformable part algorithm model proposed by (Dubout and Fleuret, 2013) which allows deformation of the individual detection parts in a multi scale fashion. Furthermore, we propose a conditional random field energy function modelling up to a pairwise 4-neighborhood relation in RGBD space. The unary potentials are defined by a probabilistic framework using a simple shape prior which biases the segmentation towards human shapes and a decision tree ensemble trained on RGB features. Likewise, the edge potentials are modelled using any of the following RGBD feature functions: Canny edges, RGB color distance, 3D Euclidean distance and surface normals. Energy function is submodular and can efficiently be solved in polynomial time using graph cuts.

We have developed a fully automatic approach for detecting and segmenting human instances in a point cloud without requiring any hard biases such as prior knowledge for the graph cut. We also show that using

Amplianitis, K., Hänsch, R. and Reulke, R.
Human Recognition in RGBD Combining Object Detectors and Conditional Random Fields.
DOI: 10.5220/0005786006550663
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 655-663
ISBN: 978-989-758-175-5

any of the aforementioned edge potentials in the energy function, does not influence as much the quality of the segmentation but mostly depends on the results of the unary potentials.

One of the main drawbacks of our approach is the fact that depending on the distance of the camera to the object but also the quality of the classifier, a detection box may or may not capture the complete body and hence can lead to an incomplete segmentation. At this point, we have to stress that potential failure to an imperfect detection box does not render the segmentation method less accurate. This is clear from the qualitative results in the experimental section.

2 RELATED WORK

Our proposed method is inspired by the work of (Ladicky et al., 2010), (Vibhav Vineet and Torr, 2011), (Teichman et al., 2013) and (Lai et al., 2012). (Ladicky et al., 2010) are the first who combined object detectors with CRFs for jointly estimating the class category, location and spatial extent of objects/regions in a scene. This work was later on used by (Vibhav Vineet and Torr, 2011) for approaching the problem of human instance segmentation. Specifically, they proposed a CRF energy function for integrating instant level information such as shape prior and exemplar histograms, biasing the segmentation towards human shape. Incorporating higher level image representations, (Shu et al., 2013) introduced a method for improving generic detectors and subsequently extracting object regions, using a superpixel-based Bag-of-Words model.

As Convolutional Neural Networks (CNNs) are becoming more famous in the computer vision community, literature in this area of research is still limited but noteworthy. To the best of our knowledge, the work of (Hariharan et al., 2014) was the first attempt towards simultaneously detecting and segmenting objects in an image. Their algorithm is based on classifying region proposals, using features extracted from both the bounding box of the region and the region foreground integrated in a jointly trained CNN.

In the RGBD domain, (Lai et al., 2012) proposed a view-based approach for segmenting objects in a point cloud generated by a depth sensor. A sliding window detector trained from different object views is used for assigning class probabilities to every image pixel. Then, they performed an MRF inference over the projected probabilities in voxel space, combining cues from different views for labeling the scene. To the best of our knowledge, this work is conceptually closer to ours. Moreover, (Teichman et al., 2013) proposed a semi-automatic approach for segmenting deformable objects in RGBD space, providing an initial seed as a prior hard constraint for inferring the segmentation. His approach makes use of a rich set of features defined in RGBD space. To the best of our knowledge, recent work in the field is the one of (Gupta et al., 2014) who studied the problem of object detection and segmentation in RGBD by combining an RGB feature-based CNN with a depth feature-based CNN in an SVM classifier.

3 OUR APPROACH

3.1 Energy Function

We begin by representing every pixel in the image as a random variable. Each of these random variables is assigned a label from the binary label set $Y = \{0,1\}$ where 0 corresponds to the background and 1 to the foreground. Let $X = \{x_1, x_2, \ldots, x_N\}$ be a discrete random field defined over a set of pixels $V = \{1,2,\ldots,N\}$. Every $x_i \in X$ associated with a pixel $i \in V$, is assigned a value y_i from the label set Y.

A Conditional Random Field is a discriminative undirected probabilistic graphical model, used to predict the values of the latent (unobserved) variables given a set of observed variables. Mathematically, this is expressed by the a posteriori probability $\mathbb{P}(\mathbf{y}|x)$ where $\mathbf{y} \in Y^n$ is the segmentation for an image of n pixels and x represents a set of features computed (in our case) from an RGBD frame. According to the Hammersley-Clifford theorem (Lafferty et al., 2001), a Conditional Random Field can be expressed in the form of a Gibbs distribution in the following way:

$$\mathbb{P}(\mathbf{y}|x) = \frac{1}{Z(x)} \exp\left(-E(\mathbf{y},x)\right) \quad (1)$$

where Z is known as the normalized or partition function. The energy function $E(\mathbf{y},x)$, corresponds to a Gibbs submodular energy function and contains all the factors for the unary and pairwise potentials which we will extensively go through in the upcoming sections. Our proposed energy function takes the form:

$$E(\mathbf{y},x) = w_{\mathcal{N}} \sum_{j \in V} \psi_j(\mathbf{y},x) + w_{\mathcal{E}} \sum_{(j,k) \in N_j} \psi_{jk}(\mathbf{y},x) \quad (2)$$

where $\psi_j(y,x)$ is a node potential function defined in our framework by the product of two conditionally independent events introduced in Section 3.4.1, $\psi_{jk}(y,x)$ is an edge potential function capturing different pairwise relations in RGBD space as discussed

in Section 3.4.2 and $w_{\mathcal{N}}$,$w_{\mathcal{E}}$ are the node and edge potential weights respectively. Our proposed energy function uses a 4-neighbourhood relation.

Given a set of features x, Eq. 1 can efficiently be solved by finding a labeling $\mathbf{y} \in Y^n$ that maximizes the Maximum A Posteriori (MAP) inference, satisfying the following statement:

$$\underset{\mathbf{y}\in\mathcal{L}}{\text{maximize}}\ \mathbb{P}(\mathbf{y}|x) = \underset{\mathbf{y}\in\mathcal{L}}{\text{minimize}}\ E(\mathbf{y},x) \qquad (3)$$

As we will analyze in Section 3.3, equation 1 cannot be solved exactly due to the existence of the partition function in the gradient. A solution to this, is to solve the energy function exactly using graph cuts. According to (Boykov and Kolmogorov, 2004), if an energy function is submodular, then it can be solved exactly and in polynomial time using graph cuts.

3.2 Object Detection

We employ an accelerated version of the deformable part-based detector (Felzenszwalb et al., 2010) introduced by (Dubout and Fleuret, 2013) with the exact same performance but with convolutions of an order of magnitude faster than the original version. Due to the non rigidness of the human figure, training over different parts can generate better prediction outputs for the position and size of the object. Let $D = \{d_1,\ldots,d_n\}$ represent the amount of detections found in an image and $S = \{s_1,\ldots,s_n\}$ the corresponding detection scores. In order to ensure that the detector will find all true positives, a lower detection threshold t_d is required. This will produce a large amount of false positives but will guarantee all true positive solutions. For eliminating all false positives and preserving only the correct detection outputs, we convert the detection scores into conditional probabilities. Platt scaling (Platt, 1999) (also known as *Platt calibration*) is used to relate the detection scores with the conditional probabilities according to the following regression formulation:

$$\mathbb{P}(c|s_i) = \frac{1}{1+\exp(A*s_i+B)}\ \ \forall s_i \in S, c \in C \qquad (4)$$

where A, B are the parameters of the sigmoid and can be found by minimising the negative log likelihood of the training or validation set and $C = \{c_B, c_F\}$ correspond to the foreground/background classes.

In order to obtain a background probability for every detection rectangle, the following formulation should hold:

$$\mathbb{P}(c_B|s) + \mathbb{P}(c_F|s) = 1,\ \ \forall s \in S \qquad (5)$$

Detection rectangles with probabilities smaller than a predefined probability threshold are classified as background.

3.3 Learning

Learning process involves finding a set of weights w returning the lowest energy $E(\mathbf{y},x)$ for the current graph. If $x = \{x_1,x_2,\ldots,x_n\}$ corresponds to a set of RGBD features computed as shown by Algorithm 1 and $D = \{(\mathbf{y_1},d_1),(\mathbf{y_2},d_2),\ldots,(\mathbf{y_n},d_n)\}$ corresponds to a set of training images, where $\mathbf{y_n}$ represents the ground truth image and d_n an RGBD frame, the goal is to learn the parameters that maximize the likelihood:

$$\max_{w} \prod_{n} \mathbb{P}(y_n|x_n) \qquad (6)$$

As already stated in Section 3.1, it is not feasible to solve the objective condition 6 exactly, due to the existence of the partition function in the gradient. The partition function $Z(x) = \sum_{\mathbf{y}} \exp(E(\mathbf{y},x))$ has an exponential number of constraints, as it sums up all possible 2^n solutions, where n the number of pixels in the image, making this problem intractable and computationally very expensive.

Algorithm 1: Generate RGBD features.

Require: $S = \{(\mathbf{y_1},d_1),(\mathbf{y_2},d_2),\ldots,(\mathbf{y_n},d_n)\}$
1: $\mathscr{D} = \emptyset$
2: **for** $(\mathbf{y_m},d_m) \in S$ **do**
3: Compute all features x_m

$$\mathscr{D} := \mathscr{D} \cup \{(\mathbf{y_m},x_m)\}$$

4: **end for**
5: **return** $\mathscr{D}$

Tsochantaridis in (Tsochantaridis et al., 2005) proposed an approach known as the *structured support vector machine (SSVM)* which tries to solve the objective condition 6 using margin maximization optimization techniques. This method was later on used by (Szummer et al., 2008) for the purpose of image segmentation.

In this paper, we employ the one-slack margin rescaling SSVM introduced by (Joachims et al., 2009) which efficiently solves the minimization problem. The learning process is presented by Algorithm 2. Here, C and ε are constant values, $w = \{w_{\mathcal{N}}, w_{\mathcal{E}}\}$ are the weights that have to be optimized for a given training set $\mathscr{D}$, ξ is a slack variable and Δ represents a loss function (in our case, a Hamming loss). Within this learning process, we enforce that the ground truth

Algorithm 2: Structured SVM.

Require: A set of training examples $\mathscr{D}$, constant values C, ε

$\mathscr{W} \leftarrow \emptyset$

repeat

Update the parameters w to maximize the margin

$$\begin{aligned} \underset{w,\,\xi}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 + C\,\xi \\ \text{subject to} \quad & w \geq 0,\ \xi \geq 0 \\ & \frac{1}{M}\sum_{m=1}^{M} E(\hat{\mathbf{y}}_\mathbf{m}, x_m) \;-\; E(\mathbf{y}_\mathbf{m}, x_m) \geq \frac{1}{M}\sum_{m=1}^{M} \Delta(\mathbf{y}_\mathbf{m}, \hat{\mathbf{y}}_\mathbf{m}) \;-\; \xi \\ & \forall\ (\hat{\mathbf{y}}_\mathbf{1}, \ldots, \hat{\mathbf{y}}_\mathbf{M}) \in \mathscr{W} \end{aligned} \tag{7}$$

for $(\mathbf{y}_\mathbf{m}, x_m) \in \mathscr{D}$ **do**

$\hat{\mathbf{y}}_\mathbf{m} \leftarrow argmin_y E(\mathbf{y}, x_m)$

end for

$\mathscr{W} \leftarrow \mathscr{W} \cup \{\hat{\mathbf{y}}_\mathbf{1}, \ldots, \hat{\mathbf{y}}_\mathbf{M}\}$

until $\frac{1}{M}\sum_{m=1}^{M} \Delta(\mathbf{y}_\mathbf{m}, \hat{\mathbf{y}}_\mathbf{m}) \;-\; E(\hat{\mathbf{y}}_\mathbf{m}, x_m) \;+\; E(\mathbf{y}_\mathbf{m}, x_m) \;\leq\; \xi + \varepsilon$

energy will have the lowest value from all other labelings. If this constraint is not satisfied, or if the margin is not achieved, this label solution will be added in the constraint set. This process continues until the values of the weights have converged. According to (Joachims et al., 2009), the objective function is quadratic to w and linear to the constraints, also known as a *quadratic programming problem*. Main advantage is that it is a convex quadratic problem and can guarantee a global minimum. We implement the Nesterov non linear quadratic optimization algorithm, which is part of a family of algorithms known as *interior point solvers*, for minimizing the objective function 7.

3.4 Potentials

3.4.1 Node Potentials

Every pixel in the image should be classified as foreground or background label, based on a cost defined in the unary term of energy function 2. In this framework, the cost is expressed by the product of two conditionally independent probability events, formulated as follows:

$$\psi_j(\mathbf{y}, x) = \begin{cases} p_1(x_j)p_2(x_j), & \text{if } y_j = 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $p_1(x_j)$ is the probability of pixel x_j to be assign a foreground label according to a learned prior shape probability map (see Algorithm 3) and $p_2(x_j)$ refers to the probability of pixel x_j to belong to the foreground, based on the probability outcome of a decision tree classifier, trained on RGB features.

Algorithm 3: Generate shape prior map.

Require: A sequence of label images $\mathbf{y}_\mathbf{m}$ and corresponding RGB images I_m of a person in the scene:

$$\mathscr{S} = \{(\mathbf{y}_\mathbf{1}, I_1), (\mathbf{y}_\mathbf{2}, I_2), \ldots, (\mathbf{y}_\mathbf{n}, I_n)\}$$

1: $\mathscr{R} = \emptyset$
2: **for** $(\mathbf{y}_\mathbf{m}, I_m) \in \mathscr{S}$ **do**
3: Get detection rectangle from image I
4: Extract the corresponding rectangle from label image $\mathbf{y}_\mathbf{m}$
5: Resize rectangle to a 128×64 sized image r_m
6: $$\mathscr{R} := \mathscr{R} \cup \{r_m\}$$
7: **end for**
8: **return** The probability map of $\mathscr{R}$

Shape Prior - The probability $p_1(x_j)$ of pixel x_j to be assigned to the foreground class is based on a learned prior shape probability map. Every detection rectangle contains regions of pixels which do not correspond to the object of interest (e.g. corners of the rectangle). Using a shape prior, we penalize these regions by assigning them a low probability. The shape prior map is learned according to Algorithm 3 and example is illustrated in Figure 2(a). It basically corresponds to the expectation of each pixel to belong to the foreground, based on the available training data.

(a) (b) (c) (d)

Figure 1: Different edge potentials: Canny edges (a), color distance (b), 3D Euclidean distance. (c), surface normals (d). **(Best viewed in colour)**.

(a) (b)

Figure 2: Unary potentials: Shape prior map (a), decision tree output(b).

Decision Trees Ensemble - As prior probability, $p_1(x_j)$ is completely independent of the measured RGBD data of a specific image. A data-dependent initial estimate is represented by $p_2(x_j)$, which corresponds to the probabilistic output of a pixel-wise classification based on the Projection-Based Random Forest (ProB-RF) framework introduced in (Hänsch, 2014). The ProB-RF classifier is used to assign each pixel a posterior probability to belong to either foreground or background based on many simple binary features extracted implicitly by the decision trees themselves. Figure 2(b) shows the estimated classification map of an exemplary scene. Since this estimation is only based on local, appearance-based information it cannot provide highly accurate and reliable results. However, this first pixel-wise probability estimate serves as an additional cue to the shape prior and is now used in the global optimization framework of CRFs.

3.4.2 Edge Potentials

Edge potentials capture the similarity between pixels lying within a local neighbourhood (also known as *Markov blanket*). Taking into consideration the richness of RGBD information, two points sharing the same label should be assigned a cost greater than zero. Specifically,

$$\psi_{jk}(\mathbf{y},x) = \begin{cases} \alpha_{jk} & \text{if } y_j = y_k \\ 0 & \text{otherwise,} \end{cases}$$

Different pairwise relations were evaluated within the RGBD domain:

Canny Edges - Canny edge extractor is a very known operator for extracting strong edges in an image. Within this framework, Canny edges are used for finding the boundaries between areas and objects, assigning a value of 1 for neighbourhood pixels that do not lie on a Canny edge and 0 otherwise.

Color Distance - Points which are part of the same neighbourhood should have similar colors. Specifically,

$$\alpha_{jk} = \exp\left(-\frac{\|c_j - c_k\|}{\sigma_c}\right) \quad (8)$$

where c_j, c_k correspond to the RGB values of points j and k respectively and σ_c is a bandwidth parameter whose value is set through cross validation.

3D Euclidean Distance - 3D points which are very close to each other are more likely to share the same label. This relationship is expressed by:

$$\alpha_{jk} = \exp\left(-\frac{|(p_j - p_k)^T n_k|}{\sigma_n} - \frac{\|p_j - p_k\|^2}{\sigma_d}\right) \quad (9)$$

where p_j, p_k correspond to the 3D values of points j and k respectively, n_k is the surface normal at point p_k and σ_n, σ_d are bandwidth parameters whose values are defined by cross validation.

Surface Normals - 3D points lying on the same part of the body should have similar normal orientations. Concretely,

$$\alpha_{jk} = \exp\left(-\frac{\theta}{\sigma_\theta}\right) \quad (10)$$

where θ is the angle between two neighbourhood normals and σ_θ is a bandwidth parameter whose value is specified by cross validation.

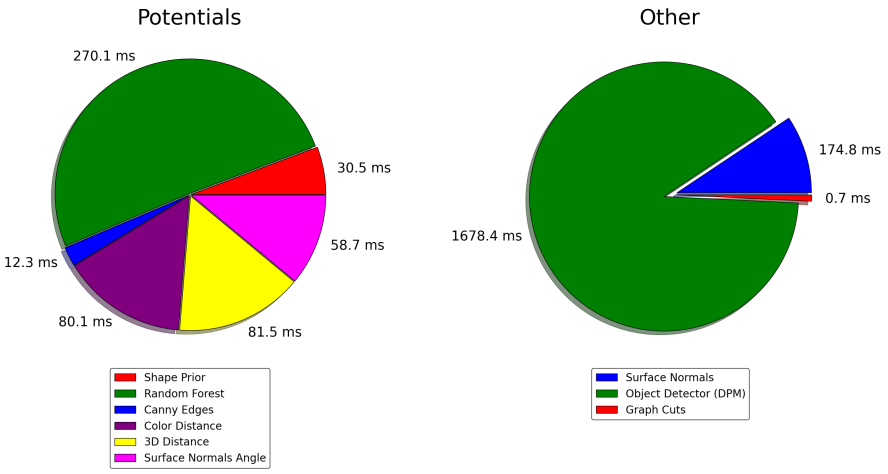

Figure 3: Average computation time for every node potential, edge potential, surface normals, object detector and graph cuts. **(Best viewed in colour)**.

4 EXPERIMENTAL RESULTS

4.1 Quantitative Analysis

We evaluate our method in an indoor environment, formed for the purpose of carrying out detection, segmentation and generally recognition tasks within the internal part of a simulated train wagon. A Kinect sensor is mounted on an aluminium construction, looking into the complete FOV of the scene. We create a dynamic environment, capable of generating different quality point clouds, depending on the amount of reflected areas/lightning conditions present in the scene. Current state of our work performs validation only on human instances but we plan to extend the evaluation also on different deformable objects which could potentially appear in an indoor environment.

Edge potentials defined on depth measurements require high precision between points lying in the local neighbourhood. Although Kinect sensor uses low distortion lenses with faintly apparent displacement errors around the corners/edges of the images, we perform a calibration of the infrared and RGB cameras for improving the quality/accuracy of the 3D points placed on these regions.

A total of 25 sequences were generated, every sequence containing 200 frames. From all 5000 images, 3000 images over 15 sequences were used for training and the rest for testing. The same training set is used for learning the weights of the structured SVM, shape prior and decision tree ensemble.

To the best of our knowledge, there is no publicly available RGBD dataset providing label images with ground truth detection boxes for the task of human instance detection and segmentation. Generating ground truth label images is a very time consuming process as it requires a lot of manual work by the user. For eliminating the effort, reference images were generated using the approach of (Shotton et al., 2013), a well known human pose estimation algorithm which was also commercialised for Kinect games.

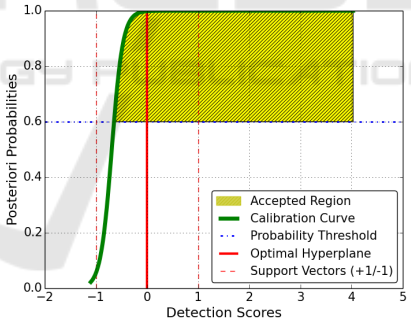

Figure 4: Learned calibration curve from the PASCAL VOC dataset. (**Best viewed in colour**).

We learn the calibration curve for converting detector scores into conditional probabilities from the publicly available PASCAL VOC dataset (Everingham et al., 2015). The resulting curve is shown in Figure 4. As it is expected, the calibration curve takes the form of a sigmoid function, capable of switching between scores and probabilities. For a detection rectangle to be assigned to the background class, a probability threshold of 0.6 was given.

The average computational times recorded for a complete scene are presented in Figure 3. It is observed that node and edge potentials require minimal effort within the pipeline while most time is needed

Figure 5: These figures outline the improvement in quality of our segmentation approach using the ground truth box rather than the detection box provided by (Dubout and Fleuret, 2013). From top to bottom: segmentation result using the detection box computed by (Dubout and Fleuret, 2013); segmentation result using the bounding box extracted by the ground truth label images; ground truth label images. (**Best viewed in colour**).

Table 1: Results in a tabular form; Every row represents a different metric evaluator; Every column corresponds to a different edge potential; Top table presents segmentation results produced by the ground truth detection box; Bottom table presents segmentation results from (Dubout and Fleuret, 2013).

GROUND TRUTH BOUNDING BOX				
Metric \ Edge Pot.	**Canny Edges**	**Color Distance**	**3D Euclidean Distance**	**Surface Normals**
Hamming Loss	5508.970 ± 1626.130	5566.37 ± 1667.510	5464.600 ± 1590.560	5504.360 ± 1896.560
Norm. Hamming Loss	0.758 ± 0.063	0.755 ± 0.069	0.760 ± 0.064	0.758 ± 0.079
PASCAL Seg. Acc.	0.799 ± 0.050	0.797 ± 0.056	0.801 ± 0.052	0.798 ± 0.071
DETECTION RECTANGLE (Dubout and Fleuret, 2013)				
Hamming Loss	8012.790 ± 2578.160	7907.000 ± 2308.790	7911.910 ± 2205.740	7936.520 ± 2178.710
Norm. Hamming Loss	0.646 ± 0.113	0.651 ± 0.104	0.651 ± 0.102	0.650 ± 0.098
PASCAL Seg. Acc.	0.705 ± 0.093	0.712 ± 0.076	0.710 ± 0.075	0.710 ± 0.074

by the object detector. For a VGA image resolution, our implementation takes $\approx 1.5s$ per frame. Graph cuts require the least effort (0.7ms) as they can be solved in polynomial time. Experiments were performed on a DELL M4800 Workstation, i7-4800MQ CPU at 2.70GHz processor and 16GB RAM. The complete pipeline is designed in a multithreaded fashion, parallelising all computations.

4.2 Qualitative Analysis

Segmentation approach was assessed using three different metrics: Hamming loss, normalized Hamming loss (Teichman et al., 2013) and the *intersection over union* loss proposed in (Everingham et al., 2015). Similarly, the detector was evaluated using a formulation introduced by (Everingham et al., 2015). The normalized Hamming loss is considered a hard penalization metric compare to the other evaluators, as it

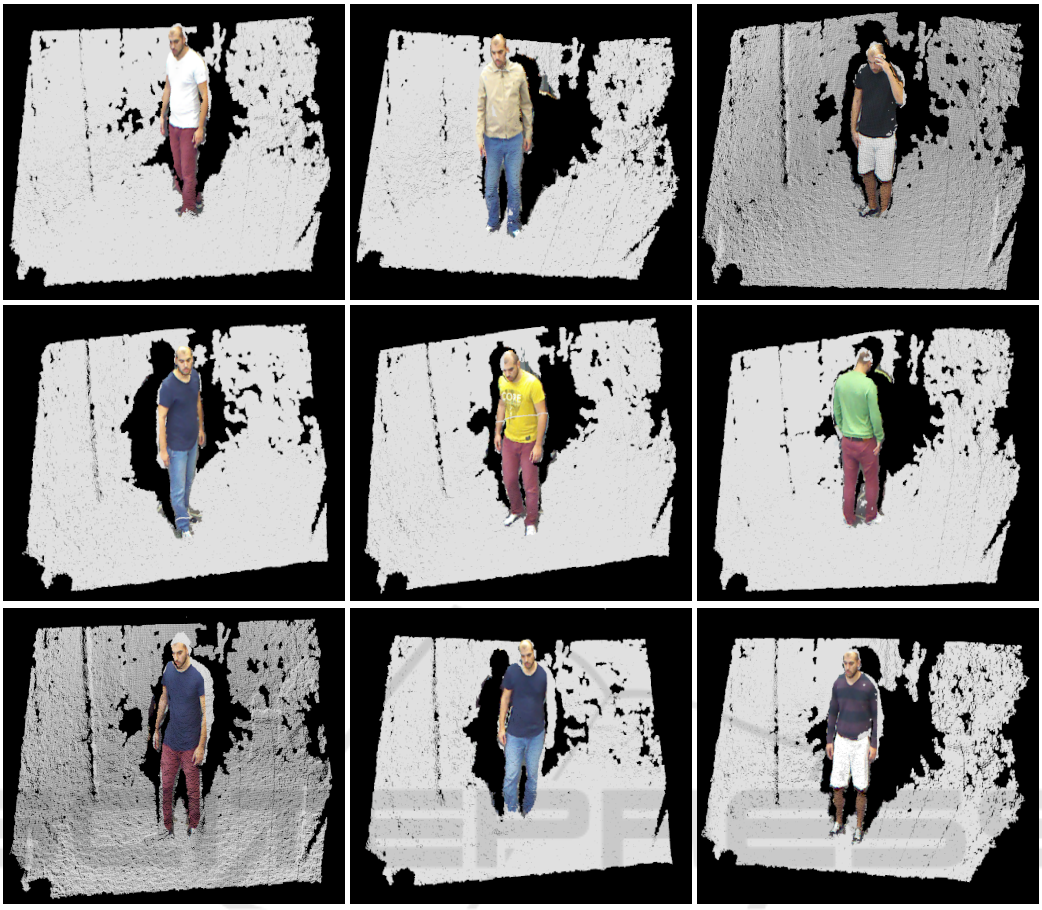

Figure 6: Results of human instance segmentations in RGBD space. (**Best viewed in colour**).

gives a zero loss if the number of incorrectly labeled pixels is equal or exceeds the number of pixels corresponding to foreground in the label image.

We provide a comparison evaluation between the different metrics (see Table 1) computed over 2000 test images. Is it evident that all metrics computed by the ground truth bounding box show an overall improvement in the segmentation accuracy, outperforming the results produced by the detection box of (Dubout and Fleuret, 2013). Furthermore, comparing the metrics computed by the different edge potentials, it is easy to perceive the insignificance between the values. This can be explained as follows: as we are not using any hard constrains such as a manual seed frame (Teichman et al., 2013) to force the s-t min cut towards a desired shape, we use the edge potentials which have a node potential larger than a predefined probability threshold. Thus, only the edge potential values which lie at the borders of the object should effect the cut.

We produce several human instance segmentations over 10 sequences, presenting only a fraction of them in Figure 5. All edge potentials discussed in Section 3.4.2 generated similar segmentation results making it hard to notice any visual differences. The main differences between the edge potentials are numerically given in Table 1. As a result, we are interested in showing how the quality of our approach is effected by different detection boxes. Looking at columns 1-3, it is clear that using the ground truth bounding box, a more precise segmentation is produced. However, columns 4 and 5 show cases were the segmentation delivers poorer results. This is a consequence of assigning high weights to the source node in the graph cut algorithm combined with strong edge potentials in that region. Furthermore, we believe that extreme poses may effect the quality of the segmentation but this is currently under investigation.

The accuracy of the detection box was checked against the ground truth bounding box over 10 sequences using the quality metric introduced by (Everingham et al., 2015), achieving an overlapping accuracy of 72.3%.

Finally, all bandwidth parameter values related to the edge potentials were set to: $\sigma_c = 0.3$, $\sigma_n = 0.5$, $\sigma_d = 0.5$, $\sigma_\theta = 0.2$.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we propose an approach for detecting and segmenting human instances in a point cloud, based on an accelerated version of the deformable part model algorithm and a pairwise CRF energy function defined over different RGBD features. Experiments showed that the quality of the segmentation depends highly on the detection box provided from the detection algorithm. Also, metric results between the different edge potentials did not provide a significant difference between them.

Current work in progress is in the direction of improving the unary potentials, incorporating depth based features for the decision tree ensemble but also generating a score map taking into account the scores returned by the detector.

In the future, we are planning to investigate the extension of the proposed energy function for incorporating higher order potentials (defined over a set of pixels) using appearance or depth information. We believe that adding shape constraints will deliver better segmentation results compared to the ones modelling only up to pairwise relations. Furthermore, we are also interested in looking into additional solutions for improving the quality of the detection boxes. Last but not least, our proposed algorithm will be tested and evaluated on different objects for verifying its robustness, using pairwise but also higher order potentials in the energy function.

REFERENCES

Boykov, Y. and Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(38):1124–1137.

Dubout, C. and Fleuret, F. (2013). Deformable part models with individual part scaling. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 28.1–28.10.

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(38):98–136.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645.

Gupta, S., Girshick, R., Arbeláez, P., and Malik, J. (2014). Learning rich features from RGB-D images for object detection and segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Hänsch, R. (2014). *Generic object categorization in PolSAR images - and beyond*. PhD thesis, Technische Universität Berlin, Germany.

Hariharan, B., Arbeláez, P., Girshick, R., and Malik, J. (2014). Simultaneous detection and segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Joachims, T., Finley, T., and Yu, C.-N. J. (2009). Cutting-plane training of structural svms. *Mach. Learn.*, 77(1):27–59.

Ladicky, L., Sturgess, P., Alahari, K., Russell, C., and Torr, P. H. S. (2010). What, where and how many? combining object detectors and crfs. In Daniilidis, K., Maragos, P., and Paragios, N., editors, *ECCV*, volume 6314 of *Lecture Notes in Computer Science*, pages 424–437. Springer.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lai, K., Bo, L., Ren, X., and Fox, D. (2012). Detection-based object labeling in 3d scenes. In *IEEE International Conference on on Robotics and Automation*.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

Shotton, J., Girshick, R. B., Fitzgibbon, A. W., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., and Blake, A. (2013). Efficient human pose estimation from single depth images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2821–2840.

Shu, G., Dehghan, A., and Shah, M. (2013). Improving an object detector and extracting regions using superpixels. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 3721–3727, Washington, DC, USA. IEEE Computer Society.

Szummer, M., Kohli, P., and Hoiem, D. (2008). Learning crfs using graph cuts. In *European Conference on Computer Vision*.

Teichman, A., Lussier, J. T., and Thrun, S. (2013). Learning to segment and track in rgbd. *IEEE T. Automation Science and Engineering*, pages 841–852.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, 6:1453–1484.

Vibhav Vineet, Jonathan Warrell, L. L. and Torr, P. (2011). Human instance segmentation from video using detector-based conditional random fields. In *Proceedings of the British Machine Vision Conference*, pages 80.1–80.11. BMVA Press.