# Graph Cuts Optimization for Multi-Limb Human Segmentation in Depth Maps

Antonio Hernández-Vela[1,2] ahernandez@cvc.uab.cat
Nadezhda Zlateva[3] zlateva@iinf.bas.bg
Alexander Marinov[3] amarinov@iinf.bas.bg
Miguel Reyes[1] mreyes@cvc.uab.cat
Petia Radeva[1,2] petia@maia.ub.es
Dimo Dimov[3] dtdim@iinf.bas.bg
Sergio Escalera[1,2] sergio@maia.ub.es

[1]Computer Vision Center, Edifici O, Campus UAB, 08193 Bellaterra (Cerdanyola), Barcelona, Spain
[2]Dept. MAIA, Universitat de Barcelona, Gran Via 585, 08007 Barcelona, Spain
[3]Inst. of Information and Communication Technologies, BAS, Acad. G. Bonchev St., Block 2, Sofia 1113, Bulgaria

## Abstract

*We present a generic framework for object segmentation using depth maps based on Random Forest and Graph-cuts theory, and apply it to the segmentation of human limbs in depth maps. First, from a set of random depth features, Random Forest is used to infer a set of label probabilities for each data sample. This vector of probabilities is used as unary term in $\alpha - \beta$ swap Graph-cuts algorithm. Moreover, depth of spatio-temporal neighboring data points are used as boundary potentials. Results on a new multi-label human depth data set show high performance in terms of segmentation overlapping of the novel methodology compared to classical approaches.*

## 1. Introduction

Human motion capture is an essential acquisition technology with many applications in computer vision. However, detecting humans in images or videos is a challenging problem due to the high variety of possible configurations of the scenario, such as changes in the point of view, illumination conditions, and background complexity. An extensive research on this topic reveals that there are many recent methodologies addressing this problem [5, 6, 11, 4]. In order to treat human pose in uncontrolled scenarios, an early work used range image for object recognition or modeling [10]. This approach achieved a straighforward solution to the problem of intensity and view changes in RGB images through the representation of 3D structures. The progress and spread of this method came slowly since data acquisition devices were expensive and bulky, with cumbersome communication interfaces when conducting experiments. Recently, Microsoft has recently launched the Kinect, a cheap multisensor device based on structured light technology, capable of capturing visual depth information (RGBD technology, from Red, Green, Blue, and Depth, respectively). The device is so compact and portable that it can easily be installed in any environment to analyze scenarios where humans are present. Following the high popularity of Kinect and its depth capturing abilities, there exists a strong research interest for improving the current methods for human pose and hand gesture recognition. While this could be achieved by inter-frame feature tracking and matching against predefined gesture models, there are sce-

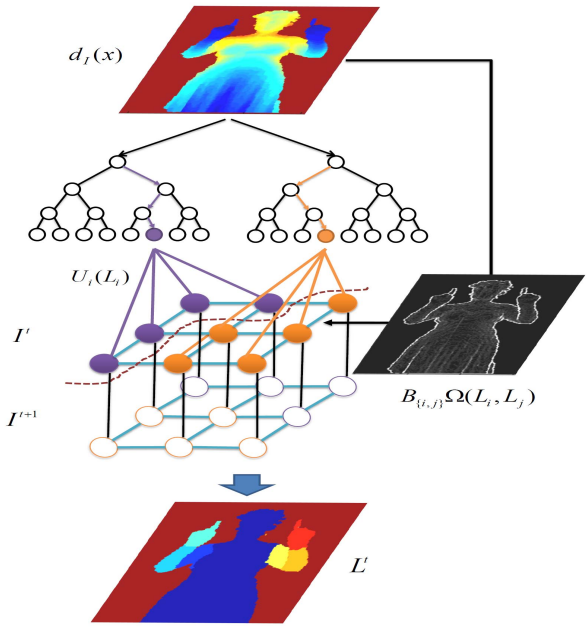


Figure 1. Pipeline of the presented method, including the input depth information, Random forest, spatio-temporal Graph-cuts optimization, and the final segmentation result.

narios where a robust segmentation of the hand and arm regions are needed, e.g. for observing upper limb anomalies or distinguishing between finger configurations while performing a gesture. In that respect, depth information appears quite handy by reducing ambiguities due to illumination, colour and texture diversity. Many researchers have obtained their first results in the field of human motion capture using this technology. In particular, Shotton et al. [12] present one of the greatest advances in the extraction of the human body pose from depth images, an approach that also forms the core of the Kinect human recognition framework. The method is based on inferring pixel label probabilities through Random Forest (RF), using mean shift to estimate human joints, and representing the body in skeletal form. Other recent work uses the skeletal model in conjunction with computer vision techniques to detect complex poses in situations where there are many actors [8].

In this paper we present a generic framework for object segmentation using depth maps based on RF and Graph-cuts theory (GC) and apply it to the segmentation of human limbs. The use of GC theory has recently been applied to the problem of image segmentation, obtaining successful results [2]. RF is used to infer a set of probabilities for each data sample, each one indicating the probability of a pixel to belong to a particular label. Then, this vector of probabilities is used as unary term in the $\alpha - \beta$ swap GC algorithm. Moreover, depth of neighbor data points in space and time are used as boundary potentials. As a result, we obtain a robust segmentation of depth images based on the defined energy terms. Our method is evaluated on a 3D data set designed in our lab, obtaining higher segmentation accuracy compared to standard RF and GC approaches.

The rest of the paper is organized as follows: Section 2 presents the novel segmentation framework in depth images based on RF and GC theory. Section 3 presents the results on new multi-label depth video data. Finally, Section 4 concludes the paper.

## 2. Method

The depth-image based approach suggested in [12] interprets the complex pose estimation task as an object classification problem by evaluating each depth pixel affiliation with a body part label, using respective Probability Distribution Functions (PDF). The pose recognition phase is addressed by re-projecting the pixel classification results and inferring the 3D positions of several skeletal joints using the RF and mean-shift algorithms. Our goal is to extend the work of [12] and combine it with a general segmentation optimization procedure to define a robust segmentation of objects in depth images. As a case study, we segment pixels belonging to the following seven body parts[1]: LU/LW/RU/RW for arms, (from Left, Right, Upper and loWer, respectively), LH/RH for hands, and the torso. The pipeline of the segmentation framework is illustrated in Fig. 1.

### 2.1. Random Forest

Considering the human body a priori segmented from the background in a training set of depth images, the procedure for growing a randomized decision tree $t$ is formulated over the same definition of a depth comparison feature as defined in [12],

$$f_\theta(I, \mathbf{x}) = \mathbf{d_I}\left(\mathbf{x} + \frac{\mathbf{u}}{\mathbf{d_I}(\mathbf{x})}\right) - \mathbf{d_I}\left(\mathbf{x} + \frac{\mathbf{v}}{\mathbf{d_I}(\mathbf{x})}\right), \quad (1)$$

where $d_I(\mathbf{x})$ is the depth at pixel $\mathbf{x}$ in image $I$, $I$ is considered subspace of the Euclidean space $E^2$, $\theta = (\mathbf{u}, \mathbf{v})$, and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^\mathbf{2}$ is a pair of offsets, the normalization of which ensures depth invariance. Thus, each $\theta$ determines two new pixels relative to $\mathbf{x}$, the depth difference of which accounts for the value of $f_\theta(I, \mathbf{x})$. Each tree consists of split and leaf nodes (the root is also a split node), as depicted in the upper part of Fig. 1. The training procedure of a given tree $t$ over a unique set of ground truth images (avoid sharing images among trees), runs through the following steps:

1. Define a set $\Phi$ of node splitting criteria $\phi = (\theta, \tau)$ , through the random selection of $\theta = (\mathbf{u}, \mathbf{v})$, and $\tau, \tau \subset \mathbb{R}$ (a set of random splitting thresholds for each $\theta$), with both $\theta$ and $\tau$ lying within some predefined range limits. After training, each split node will be assigned with its optimal $\phi$ value from $\Phi$

2. Define a set $Q$ of training examples $Q = \{(I, \mathbf{x}) | \mathbf{x} \in \mathbf{I}\}$, over the entire set of training images for the tree, where $I$ stands for an image from the training set, $\mathbf{x}$ is a randomly selected pixel in $I$, and the number of pixels $\mathbf{x}$ per image is fixed. Estimate the PDF of $Q$ over the whole set of labels $C$ (in our case $|C| = 7$),

$$P_Q(c) = \frac{h_Q(c)}{|Q|}, c \in C, \quad (2)$$

   where $h_Q(c)$ is the histogram of the examples from $Q$, associated with the label $c \in C$. Each example from $Q$ enters the root node, thus ensuring optimal training of the tree $t$.

3. At the currently being processed node (starting from the root), split the (sub)set $Q$, entering this node, into two subsets $Q_L$ and $Q_R$, obeying Eq. (1):

[1] Note that the method can be applied to segment any number of labels of any object contained in a depth image.

$$\begin{aligned} Q_L(\phi) &= \{(I,\mathbf{x})\,|f_\theta\,(I,\mathbf{x}) < \tau\}, \phi = (\theta,\tau) \\ Q_R(\phi) &= Q \setminus Q_L, \end{aligned} \tag{3}$$

and estimate the PDF of $Q_L$, $P_{Q_L}(c)$, as defined in Eq. (2). Compute the PDF of $Q_R$, which may be speeded up by the following formulae,

$$P_{Q_R}(c) = \frac{|Q|}{|Q_R|}P_Q(c) - \frac{|Q_L|}{|Q_R|}P_{Q_L}(c), \tag{4}$$

$$c \in C, Q_R = Q_R(\phi), Q_L = Q_L(\phi). \tag{5}$$

4. Estimate the best splitting criterion $\phi^*$ for the current node, so that the information gain $G_Q(\phi^*)$ of partitioning set $Q$, entering the node, into left and right subsets to be maximum:

$$G_Q(\phi) = H(Q) - \frac{|Q_L(\phi)|}{|Q|}H(Q_L(\phi)) - \frac{|Q_R(\phi)|}{|Q|}H(Q_R(\phi)), \tag{6}$$

where $\phi = (\theta,\tau) \in \Phi$, and $H(Q) = -\sum_{c\in C} P_Q(c)\ln(P_Q(c))$ represents Shannon's entropy for the input (sub)set $Q$ and its splits ($Q_L$ and $Q_R$) over the set of labels $C$. It is more or less obvious that $G_Q(\phi) > 0$, $\phi \in \Phi$, but it is difficult to make a more analytical statement for the behaviour of $G_Q(\phi)$. That is why we also use the full search approach to evaluate $\phi^* = \arg\max_{\phi\in\Phi} G_Q(\phi)$.

5. Recursively repeat step 3 and 4 over $Q_L\,(\phi^*)$ and $Q_R\,(\phi^*)$ for the left and right node children respectively, until some preset stop conditions are met. The node where the stop condition occurred is treated as a leaf node, where, instead of $\phi^*$, the respective PDF for the subset $Q$ reaching the node, is stored (see Eq. (2)).

Once trained, each image pixel for recognition, i.e. an example $(I,\mathbf{x})$, is run through the tree, starting from the root and ending at a leaf node, taking a path that depends solely on the value of $f_\theta(I,\mathbf{x}) < \tau$, using the splitting criterion $\phi = (\theta,\tau)$, stored at the current tree node. The pixel acquires the PDF kept at the reached leaf node. The inferred pixel probability distribution within the forest is estimated by averaging the PDFs over all trees in the forest as follows,

$$P\,(c|I,\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} P_t\,(c|I,\mathbf{x})\,, c \in C \tag{7}$$

where $P_t\,(c|I,\mathbf{x})$ is the PDF stored at the leaf, reached by the pixel for classification $(I,\mathbf{x})$ and traced through the tree $t$, $t \in T$.

### 2.2. Spatio-Temporal Graph-cut optimization

GC [2] is an energy minimization framework which has been considerably applied in image segmentation –both binary and multi-label–, with highly successful results. In this work, we extend the GC theory to be used in depth images and optimize the results obtained from the RF approach in order to deal with automatic spatio-temporal multi-label segmentation.

Given $I = \{I^1,...,I^s,...,I^S\}$ the set of frames of the video sequence, and $\mathcal{X} = (\mathbf{x_1},...,\mathbf{x_i},...,\mathbf{x}_{|\mathcal{P}|})$ the set of pixels of $I$, let us define $\mathcal{P} = (1,...,i,...|\mathcal{P}|)$ the set of indexes of $I$; $\mathcal{N}$ the set of unordered pairs $\{i,j\}$ of neighboring pixels in space and time, under a defined neighborhood system –typically 6- or 26-connectivity–, and $L = (L_1,...,L_i,...,L_{|\mathcal{P}|})$ a vector whose components $L_i$ specify the labels assigned to pixels $i \in \mathcal{P}$. This framework defines an energy function $E(L)$ that combines local and contextual information, and whose minimum value corresponds to the optimal solution of the problem –in our case, the optimal segmentation:

$$E(L) = U(L) + \lambda B(L) \tag{8}$$

The first term of the energy function is called the "unary potential". This potential encodes the local likelihood of the data by assigning individual penalties to each pixel for each one of the defined labels $U(L) = \sum_{i\in\mathcal{P}} U_i(L_i)$. The second term or "boundary potential" encodes contextual information by introducing penalties to each pair of neighboring pixels as follows $B(L) = \sum_{\{i,j\}\in\mathcal{N}} B_{\{i,j\}}\ \Omega(L_i,L_j)$, where $\Omega(L_i,L_j)$ a function that introduces prior costs between each possible pair of neighboring labels. Finally, $\lambda \in \mathbb{R}^+$ is a weight that specifies the relative importance of the boundary term against the unary term.

Once the energy function is defined, a graph $\mathcal{G} =< \mathcal{V},\mathcal{E} >$ is built following the neighborhood system used in the boundary potential $B(L)$. From a practical point of view, and considering that computer memory resources are limited, we adopt a sliding-window approach. More specifically, we define a fixed size volume window $V$ like the one depicted in the bottom of Fig. 1. The sliding-window approach starts segmenting the first $|V|$ frames, and covers all the video sequence volume, with a one-frame stride. This means that all the frames except the first and the last one are segmented at least twice, and $|V|$ times at most. In order to select the final hypothesis for each frame, we use the energy value resulting from the minimization algorithm at each execution. Therefore, the execution with the lowest energy value is the one we trust as the best hypothesis. Once the graph is built with the energy function values, two main algorithms can be applied in order to find not the minimum energy, but a suboptimal approximation of it: $\alpha-\beta$ swap and $\alpha$-expansion [3]. While the first one is less restrictive

| edge | weight (cost) | for |
|---|---|---|
| $t_i^{\alpha}$ | $U_i(\alpha) + \sum_{\substack{j \in \mathcal{N}_i \\ L_j \notin \{\alpha,\beta\}}} B(\alpha, L_j)$ | $L_i \in \{\alpha, \beta\}$ |
| $t_i^{\beta}$ | $U_i(\beta) + \sum_{\substack{j \in \mathcal{N}_i \\ L_j \notin \{\alpha,\beta\}}} B(\beta, L_j)$ | $L_i \in \{\alpha, \beta\}$ |
| $e_{\{i,j\}}$ | $B(\alpha, \beta)$ | $\{i,j\} \in \mathcal{N}$ <br> $L_i, L_j \in \{\alpha, \beta\}$ |

Table 1. Weights of edges in $\mathcal{E}$.

and can be applied in a broader range of energy functions, the second one has been proven to obtain better results, as long as the energy function fulfills some conditions [3]. In the case of $\alpha - \beta$ swap, the boundary term $B_{\{i,j\}}$ must be *semi-metric*, which means that the conditions in Eq. (9) and (10) must be fulfilled,

$$B(L_i, L_j) = B(L_j, L_i) \geq 0 \quad (9)$$

$$B(L_i, L_j) = 0 \leftrightarrow L_i = L_j \quad (10)$$

$$B(L_i, L_j) \leq B(L_i, L_n) + B(L_n, L_j), \quad (11)$$

for any $L_i, L_j, L_n \in L$, being $B(L_i, L_j) = B_{\{i,j\}}$ $\Omega(L_i, L_j)$. Additionally, if we want to apply $\alpha$-expansion, the condition in Eq. (11) must also be fulfilled. In that case, the boundary term $B_{\{i,j\}}$ is said to be *metric*.

In our case, Eq. (11) is not true for all nodes in $\mathcal{G}$, and so, we use $\alpha - \beta$ swap in our segmentation methodology for depth maps. This way, the set of nodes $\mathcal{V}$ contains a node for each pixel in $I$, plus two terminal nodes: $\alpha$ and $\beta$. Similarly, $\mathcal{E}$ is composed by two kinds of edges: terminal links $t_i^{\alpha}$ and $t_i^{\beta}$, and neighbor links $e_{\{i,j\}}$. The values assigned to the edges of $\mathcal{G}$ are then assigned following Table 1. The following subsections define the specific energy function potentials that we designed for our problem.

**Unary potential**. The unary potential encodes the local likelihood for each pixel to belong to each one of the labels $L_i$ of our problem. In our case, we have used the log-likelihood of the probabilities returned by the RF for the computation of the unary potential $U_i(L_i) = -\ln(P(c|I, x))$, obtaining a unary cost potential for each class $c_i$ –corresponding to label $L_i$ in GC. This step is shown at the top of Figure 1, where the output probabilities of the leafs of the RF trees are used to compute the unary potentials $U_i(L_i)$ at the input edges of the GC graph.

**Boundary potential**. In the case of the boundary potential, we use the following formulation,

$$B_{\{i,j\}} = \frac{1}{\mathsf{dist}(i,j)} e^{-\beta ||\mathbf{x}_i - \mathbf{x}_j||^2}$$

, where $\beta = \left(2\langle(\mathbf{x}_i - \mathbf{x}_j)^2\rangle\right)^{-1}$, and $\mathsf{dist}(i,j)$ computes the Euclidean distance between the cartesian coordinates of pixels $\mathbf{x}_i$ and $\mathbf{x}_j$. The information about the pixels $\mathbf{x}_i$ and $\mathbf{x}_j$ that we use in the exponential function is just the depth value, although in the experimental section we test other additional approaches. Finally, we defined two different $\Omega(L_i, L_j)$ functions in order to introduce some prior costs between different labels. On one hand, we considered the trivial case where all different labels have the same cost,

$$\Omega_1(L_i, L_j) = \begin{cases} 0 & \text{for} \quad L_i = L_j \\ 1 & \text{for} \quad L_i \neq L_j \end{cases} \quad (12)$$

On the other hand, we introduced some spatial coherence between the different labels, taking into account the kinematic constraints of the human body limbs,

$$\Omega_2(L_i, L_j) = \begin{cases} 0 & \text{for} \quad L_i = L_j \\ 10 & \text{for} \quad L_i = \text{LU}, L_j = \text{RU} \\ & \quad\;\; L_i = \text{LH}, L_j = \text{RH} \\ 5 & \text{for} \quad L_i = \text{LW}, L_j = \text{RH} \\ & \quad\;\; L_i = \text{RW}, L_j = \text{LH} \\ 1 & \text{otherwise} \end{cases} \quad (13)$$

With this definition of the inter-label costs, we are making it difficult for the optimization algorithm to find a segmentation in which there exists a frontier between the right and left upper-arms, right and left hands, or in the lower measure, between left hand and right lower-arm, and vice-versa. Therefore, we are assuming that poses in which the two hands are touching are not probable[2].

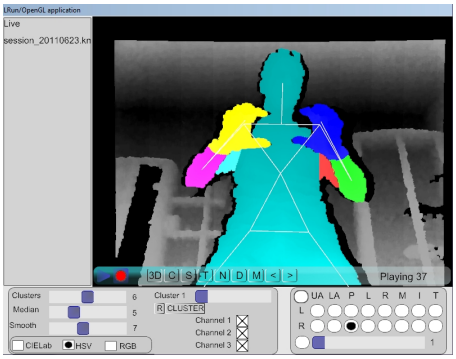


Figure 2. Interface for semi-automatic ground-truth generation.

## 3. Experiments and results

This section starts with a brief description of the considered data and the different methods, parameters, and validation protocol of the evaluation.

**Data:** We defined a new data set of several sessions where the actors are performing different gestures with their hands in front of the Kinect camera – only the upper body is considered. Each frame is composed by one 24 bit RGB image of size 640×480 pixels, one 12 bit depth buffer of

[2]This label coherence cost should be estimated for each particular problem domain. In our particular data set of poses, the values of 1, 5, and 10 were experimentally computed.

| | Torso | LU arm | LW arm | L hand | RU arm | RW arm | R hand | Avg. per class |
|---|---|---|---|---|---|---|---|---|
| 100 $f_\theta$, $O_{max}$ = 30, $L_{max}$ = 20 | 92.90 | 73.29 | 71.42 | 57.75 | 74.25 | 76.26 | 59.38 | 72.18 |
| 100 $f_\theta$, $O_{max}$ = 60, $L_{max}$ = 20 | 94.17 | 79.83 | 77.69 | **77.10** | 81.04 | 82.65 | 80.17 | 81.81 |
| 80 $f_\theta$, $O_{max}$ = 60, $L_{max}$ = 20 | **94.22** | 79.08 | 76.46 | 74.19 | 81.24 | 83.26 | 79.05 | 81.07 |
| 60 $f_\theta$, $O_{max}$ = 60, $L_{max}$20 | 94.09 | 78.86 | 75.86 | 73.49 | 79.43 | 82.60 | 78.08 | 80.34 |
| 100 $f_\theta$, $O_{max}$ = 60, $L_{max}$ = 15 | 94.06 | 79.81 | 78.69 | 76.59 | 81.18 | 83.10 | **80.23** | **81.95** |
| 100 $f_\theta$, $O_{max}$ = 60, $L_{max}$ = 10 | 91.83 | **81.47** | **78.98** | 72.30 | **83.00** | **83.74** | 76.85 | 81.17 |
| 60 $f_\theta$ + 20 $g_\theta$, $O_{max}$ = 60, $L_{max}$ = 20 | 94.04 | 77.73 | 74.93 | 71.97 | 77.62 | 81.22 | 76.64 | 79.17 |

Table 2. Average per class accuracy in % calculated over the test samples in a 5-fold cross validation. $f_\theta$ represents features of the depth comparison type from Eq.1, while $g_\theta$ - the gradient comparison feature from Eq. (14). $O_{max}$ is the upper limit for the module of the $\mathbf{u}$ and $\mathbf{v}$ offsets , and parameter $L_{max}$ stands for the maximal depth level of the decision trees.

| | Torso | LU arm | LW arm | L hand | RU arm | RW arm | R hand | Avg. per class |
|---|---|---|---|---|---|---|---|---|
| RF results | 94.06 | **79.81** | 78.69 | 76.59 | 81.18 | 83.10 | 80.23 | 81.95 |
| TC, Depth, $\Omega_2\,(L_i, L_j)$ | 98.44 | 78.93 | **84.38** | 88.32 | **82.57** | **88.85** | 93.86 | **87.91** |
| Fbf, Depth, $\Omega_1\,(L_i, L_j)$ | 98.86 | 75.05 | 82.87 | 91.45 | 77.57 | 87.35 | 93.96 | 86.73 |
| Fbf, Depth, $\Omega_2\,(L_i, L_j)$ | 98.86 | 75.03 | 83.36 | **92.41** | 77.54 | 87.67 | **94.20** | 87.01 |
| Fbf, RGB+Depth, $\Omega_1\,(L_i, L_j)$ | **99.02** | 72.02 | 81.86 | 90.29 | 76.56 | 86.84 | 92.14 | 85.53 |
| Fbf, RGB+Depth, $\Omega_2\,(L_i, L_j)$ | **99.02** | 72.03 | 81.95 | 91.19 | 76.53 | 87.12 | 92.12 | 85.71 |

Table 3. Average per class accuracy in% obtained when applying the different GC approaches –TC: Temporally coherent, Fbf: Frame-by-Frame– , and the best results from the RF approach, in the first row.

the same dimension, and a skeletal graph describing important joints of the upper human body. In order to label every pixel we designed an editing tool to facilitate labelling in a semi-supervised manner. Each frame is accompanied with a label buffer of the same dimension as the captured images. The label buffer is automatically initialized through a rough label estimation algorithm. The pixels bounded by the cylinders between the enclosing joints of the shoulder to elbow are labelled as upper arm (LU/RU). By analogy the pixels inside the cylinder between the elbow and the joint of the hand are labelled as lower arm (LW,RW). The palm is labelled by the pixels bounded by a sphere centered in the joint of the hand (LH,RH). The RGB, depth, and skeletal data are directly obtained via the OpenNI library [1]. Finally each frame is manually edited to correct the roughly estimated labels by the initialization algorithm. The ground truth contains 2 actors in 3 sessions gathering 500 frames in total. An example of the developed interface is shown in Figure 2 [3]. We also made an extra experiment for finger segmentation defining 6 labels per hand - one label for each finger and one for the palm. In this case, we used coloured gloves and 63 frames were generated.

**Methods and validation:** Inspired by the reported test parameters and accuracy results in [12], our experiments rest on the following setup: we perform a 5-fold cross-validation over the available 500 frames by training a random forest of $T = 3$ trees. Therefore, we use 130 unique training images per tree with 1000 uniformly distributed pixels per image. We limit the maximum depth level $L_{max}$ for all trees to 20, and use 100 candidate features $\theta$ and 20 candidate thresholds $\tau$ per feature to build the splitting criteria set $\Phi$. The remaining 100 images form the test set. We also compare the results with another set of features: a mixture of depth Eq. (1) and gradient Eq. (14) features,

$$g_\theta\left(I, \mathbf{x}\right) = \measuredangle\left(\nabla d_I\left(\mathbf{x} + \frac{\mathbf{u}}{\mathbf{d_I}\left(\mathbf{x}\right)}\right) - \nabla d_I\left(\mathbf{x} + \frac{\mathbf{v}}{\mathbf{d_I}\left(\mathbf{x}\right)}\right)\right), \tag{14}$$

where $\nabla d_I\left(\mathbf{x}\right)$ is the gradient vector of depth at pixel $\mathbf{x}$, and the feature $g_\theta\left(I, \mathbf{x}\right)$ represents the angle between the two gradient vectors at offsets $\mathbf{u}$ and $\mathbf{v}$ from $\mathbf{x}$. When applying GC, the $\lambda$ parameter was set to 50 for all the performed experiments, the nodes of the graph are 10-connected –8 spatial neighbors + 2 temporoal neighbors–, and the size of the sliding window is set to $|V| = 5$. In order to do a more complete comparison of the results, we perform an additional GC experiment removing the temporal coherence. In this frame-by-frame approach, the $\alpha$-expansion algorithm is used. Moreover, in this second experiment we also compare the use of different pixel information for the computation of the boundary term. Apart from just depth information, we also test just RGB information, as in the standard GrabCut algorithm [9], and both RGB and depth together, resulting in a 4-D RGBD vector. Finally, we also apply the Friedman test [7] in order to look for statistical significance of the performed experiments.

### 3.1. Random forest results

Table 2 shows the estimated average classification accuracy for each of the considered labels. Without claiming

[3] http://www.cvc.uab.es/~ahernandez/data.html

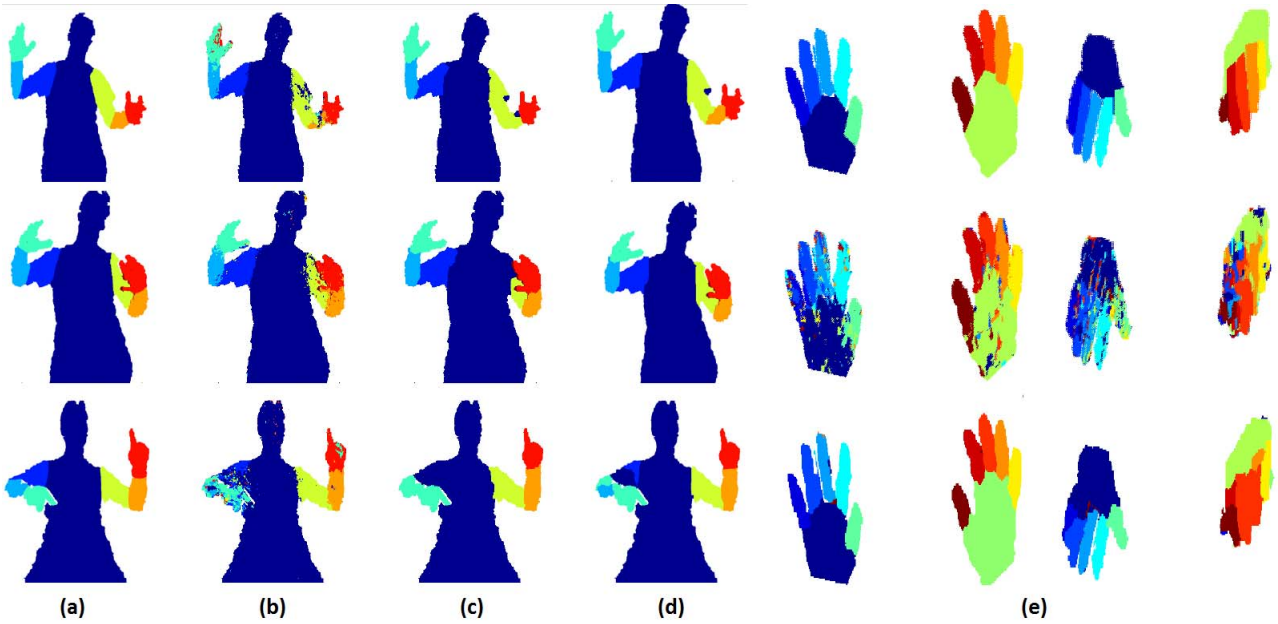


Figure 3. Qualitative results; Ground Truth (a), RF inferred results (b), frame-by-frame GC results (c), and Temporally-coherent GC results (d). (e) Hand segmentation experiment. First row shows the ground-truth for two examples. Second row shows the RF classification results. Third row shows the final $\alpha$-expansion GC segmentation results.

exhaustiveness of our experiments, the results from Table 2 allow us to make the following analysis: The maximum offset $O_{max}$ has the greatest impact on the accuracy results at the hands regions, which are with the smallest area in our body part definition. Doubling the size of $O_{max}$ leads to an increase in the accuracy of 20% for the hands and 6% for the other body parts. In other words, $O_{max}$ increases the feature diversity and the global ability to represent spatial detail. The number of candidate features $Q$ would not have such a tremendous impact on the accuracy as the $O_{max}$ parameter, though a higher number helps identifying the most discriminative features. We also tested the impact of $L_{max}$. Trimming the trees to depth 15 has a very little impact, showing an improvement of 0.1% on the average accuracy that may weakly be attributed to better classification at the lower arm regions. Trimming to depth 10 shows a 4% decrease in the accuracy at the hands, i.e. the tree is not trained well enough. Our final test includes comparison over combination of both features $f_\theta$ and $g_\theta$. Since the depth data provided by Kinect is noisy, we apply a Gaussian smoothing filter before calculating the image gradients and the feature from Eq. (14). We chose the gradient feature since it complements the relations of depth features with information about the orientation of local surfaces. However, in our test we did not find significant improvement in the performance of the RF approach when including this kind of features.

In order to show the generalization capability of the proposed approach, we carried out another case study, consisting of segmenting the finger regions. The results applying the same validation as in the previous case, show the best performance for the following setup: 1 tree of depth 15, 500 pixels per image, 100 candidate features $Q$, 20 thresholds $\tau$ per feature, and $O_{max}$ = 45. The estimated average per class accuracy was 58.5% mostly due to the small number of training images. Fig. 3(e) displays a couple of test images comparing the ground truth and the inferred labels.

### 3.2. Spatio-Temporal Graph-cuts results

The results we obtained when applying the GC proposal over the probabilities returned by the RF are detailed in Table 3. We can see how these results improve RF, and also the one obtained in the frame-by-frame approach. We can see that we obtain the best results when using only depth information for the computation of the boundary potential. In our case study, adding RGB to the depth information reduces the generalization of the boundary potential. In Fig. 3(a)-(d) we can see some qualitative results of the segmentations. Another interesting result is the improvement because of the influence of the prior costs given by the different $\Omega(L_i, L_j)$ functions. Having a look at the qualitative results in Fig. 3(a)-(d), one can firstly see how the spatial coherence introduced by the basic frame-by-frame Graph-cuts approach –Fig. 3 (c)–, allows to recover more consistent regions than the ones obtained with just the RF probabilities. Moreover, when introducing temporal coherence

–Fig. 3 (d)–, the classification of certain labels like the ones corresponding to the arms is improved.[4]

Finally, in order to reject the null hypothesis that the measured ranks differ from the mean rank, and that the ranks are affected by randomness in the results, we use the Friedman test [7]. The rankings are obtained estimating each relative rank $r_i^j$ for each test $i$ and each segmentation strategy $j$, and computing the mean ranking $R$ for each strategy as $R_j = \frac{1}{N}\sum_{i=1}^{N} r_i^j$, where $N$ is the total number of performed tests. The Friedman statistic value is then computed as

$$X_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right].$$

In our case, with $k = 6$ segmentation strategies to compare and ranks $R = [6, 2.5, 2.12, 4.12, 3.75, 2.12]$ in the order showed in Table 3, $X_F^2 = 20.06$. Lastly, we compute the statistic proposed by Iman and Davenport,

$$F_F = \frac{(N-1)X_F^2}{N(k+1) - X_F^2},$$

and obtain $F_F = 7.04$. With six methods and ten experiments, $F_F$ is distributed according to the $F$ distribution with five and 35 degrees of freedom. The critical value of $F(5, 35)$ for $0.05$ is $2.45$. As the value of $F_F$ is higher than $2.45$ we can reject the null hypothesis, and thus, looking at the best mean performance in Table 3, we can conclude that the spatio-temporal GC proposal is the first choice in the presented experiments.

In the second experiment, labelling pixels from hands –in a frame-by-frame fashion–, we achieve an average per class accuracy of 70.9%, which supposes even a greater improvement than in the case of human limbs. Fig. 3(e) also shows some qualitative results of the GC approach.

## 4. Conclusion

We proposed a generic framework for object segmentation using depth maps based on Random Forest and Graph-cuts theory in order to benefit from the use of spatial and temporal coherence, and applied it to the segmentation of human limbs. Random Forest estimated the probability of each depth sample point to belong to a set of possible object labels, while Graph-cuts was used to optimize, both locally, spatially and temporally, the RF probabilities. Results on two novel data sets showed high performance segmenting several body parts in depth images compared to classical approaches.

[4]Look the supplemental material for video samples.


## Acknowledgements

This work has been partially supported by the following projects: TIN2009-14404-C02, CONSOLIDER-INGENIO CSD 2007-00018, and BG051PO001-3.3.04/13 by the Operative Programme "Human Resources Development" of the European Social Fund. The work of Antonio is supported by an FPU fellowship from the Spanish government.


## References


[1] Openni. http://www.openni.org. 5

[2] Y. Boykov and G. Funka-Lea. Graph cuts and efficient n-d image segmentation. *IJCV*, 70:109–131, 2006. 2, 3

[3] Y. Boykov, O.Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23:1222–1239, 2001. 3, 4

[4] G. Cheung, T. Kanade, J.-Y. Bouguet, and M. H. A. Real time system for robust 3d voxel reconstruction of human motions. 2:714–720, 2000. CVPR. 1

[5] N. Dalal and B. Triggs. Histogram of oriented gradients for human detection. volume 2, pages 886–893, 2005. 1

[6] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. *European Conference on Computer Vision*, pages 7–13, 2006. 1

[7] J. Demsar. Statistical comparisons of classifiers over multiple data sets. *JMLR*, 7:1–30, 2006. 5, 7

[8] Y. Liu, C. Stoll, J. Gall, and H.-P. Seidel. Markerless motion capture of interacting characters using multi-view image segmentation. *CVPR*, 14(1):1249–1256, 2011. 2

[9] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH 2004 Papers*, pages 309–314, 2004. 5

[10] B. Sabata, F. Arman, and J. K. Aggarwal. Segmentation of 3d range images using pyramidal data structures. *CVGIP: Image Understanding*, 57(3):373–387, 1993. 1

[11] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis. Human detection using partial least squares. pages 24–31, 2009. International Conference on Computer Vision. 1

[12] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. *CVPR*, pages 1297–1304, 2011. 2, 5