ACTIVE LEARNING BASED AUTOMATIC FACE SEGMENTATION FOR KINECT VIDEO

Jixia Zhang[1], *Haibo Wang*[2], *Shaoguo Liu*[1], *Franck Davoine*[3], *Chunhong Pan*[1], *Shiming Xiang*[1]

National Laboratory of Pattern Recognition, CASIA[1], CNRS[3]
The Robotics Research Center, Shandong University[2]
jixiazhang@gmail.com

ABSTRACT

This paper presents a novel segmentation approach for extracting faces from videos. Under an active learning framework, the segmentation is conducted automatically without human interactions. A small portion of pixels are first labeled as face or non-face. Given these labeled samples, a semi-supervised spline regression model is then applied to obtain the face region. Based on the segmentation result, new pixels are selected and labeled. These two steps perform iterately until convergence. The main novelty is that color and depth data are combined to provide the labeling information. Our approach is validated via comparisons with state-of-the-art methods on real videos captured from the commodity Kinect camera.

1. INTRODUCTION

Face segmentation plays an important role in many computer vision applications, such as human computer interaction, video conferencing and video editing [1]. However, it is difficult to extract accurate face regions. The challenges lie in occlusions, diverse illuminations and complex background.

During the past decades, many approaches have been proposed for object segmentation. In earlier times, the approaches are commonly data-driven [2, 3]. Unlike those, recent approaches partition an input image into semantic objects with specified prior information [4, 5]. According to how the prior is given, this kind of approach can be classified into two categories: interaction based method and automatic method. In the former, the prior information is supplied by human [6, 4]. A rectangle containing the object to be segmented is required in [4] while many scribbles about the object and background are needed in [6]. This cost of human interventions limits its application in dynamic conditions [7]. Besides, different interventions lead to different results as shown in Fig. 1. Efficient segmentation is usually achieved after many trials which is also restricted in video applications. Different from that, the automatic method obtains prior information by an offline learned model [5]. For example, a frontal face detector is employed to locate the face and the segmentation is then conducted based on the face location in [5]. However, current frontal face detector works well but designing a robust profile face detector is still unsolved.

Obtaining priors via face detection is restricted for the difficulty of robust face detector. However, skin color, being invariant to pose changes, is a stable information for face. It has been employed in skin detection [9], face tracking [10]. Hence, it sounds reasonable to obtain prior information via skin color detection. Nevertheless, the skin colors vary with lighting changes and many background colors may look similar as skin. Those decrease its robustness for providing reliable clues. With the development of low-cost depth camera, the depth information is easily acquired. Its insensitivity to illuminations makes it compensating the limitations of color cue. That makes fusing depth with color reasonable. It should be noted that the depth image from low-cost camera is usually coarse with possible noises and holes. Thus, segmentation with only depth is apt to be inaccurate. Motivated by that, we propose an automatic face segmentation approach with both color and depth cues. Skin color detection and depth constraint work together to provide prior information. Given the priors, the segmentation is performed by the local spline regression [6] which is embedded in active learning framework to improve the accuracy. Experimental results verify its efficiency and robustness in video segmentation.

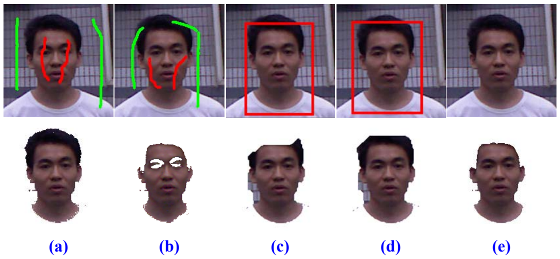

Fig. 1. Face segmentation comparisons: (a) and (b) the results by [6] with different scribbles; The red and green strokes illustrate the face and non-face respectively; (c) and (d) results of GrabCut with different interactive regions [4]; (e) the result of our method.

Our method is distinguished by the following contributions:

- Color and depth cues are fused to acquire semantic priors automatically. The priors are comparative with the information provided by users as in the interactive methods [4, 6]. Besides, it is robust to face pose variations compared to the frontal face detector in [5].
- Active learning is applied to improve the segmentation accuracy if needed. Under difficult conditions, efficient segmentation results are usually achieved after several trials of human intervention in [4, 6].

The paper is organized as follows. The main segmentation framework is described in Section 2. How the active learning approach works is talked in Section 3. In Section 4, experimental results and analysis are presented. Finally, discussion and future work are illustrated in Section 5.

This work was supported in part by the National Basic Research Program of China under Grant 2012CB316304, and the National Natural Science Foundation of China under Grants 61175025, 61272331, 61005036 and 61203279.

2. THE SEGMENTATION FRAMEWORK

In this section, the framework of the proposed segmentation method is first illustrated, which applies active learning to the local spline regression based segmentation (LSR-Seg). To make the paper self-contained, the main idea of LSR-Seg is then described. How the active learning is utilized is presented in next section.

2.1. Main Framework

Given a color image and its corresponding depth information, the proposed method extracts face region without user intervention. This is achieved automatically through applying active learning to LSR-Seg. The flowchart is illustrated in Fig. 2. LSR-Seg classifies all the pixels into face and non-face given the labels of partial points. Active learning is then utilized to improve its performance. This is conducted by sampling query points based on the segmentation results and labeling them by an oracle. These labeled points are employed for the next LSR-Seg. That process iterates until no segmentation improvement is achieved. Here, the query points are labeled automatically using both color and depth information.

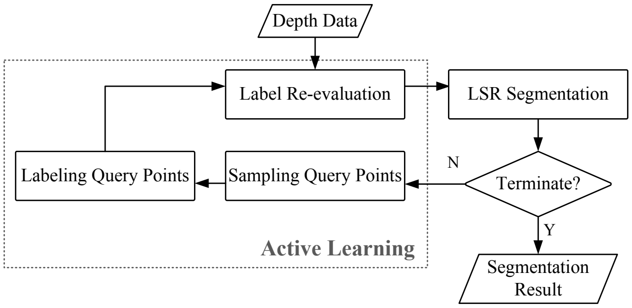

Fig. 2. The flowchart of active learning based segmentation.

To formulate the problem mathematically, the input color image is denoted as I and the depth information as $\mathcal{D}$. The color feature for one pixel is represented as $\mathbf{x} \in \mathbb{R}^3$ and its label is denoted as $y \in \{1, -1\}$ (1 for face point and -1 for non-face point). The color features for all the pixels are collected into $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ and their corresponding depth features are collected in $\mathcal{D} = \{d_1, d_2, \dots, d_n\}$, where n is the number of pixels. Thus, the problem is to estimate the labels for all pixels $\mathcal{Y} = \{y_1, y_2, \dots, y_n\}$ with $\mathcal{X}$ and $\mathcal{D}$:

$$\{\mathcal{X}, \mathcal{D}\} \rightarrow \mathcal{Y}. \tag{1}$$

Before conducting LSR-Seg, the labels for a small portion of $\mathcal{X}$ must be available. Denote the labels for these pixels as $\mathcal{F}_l = \{f_{l_1}, f_{l_2}, \dots, f_{l_m}\}, 1 < m < n, 1 \leq l_j \leq n$ which are automatically obtained in this paper.

2.2. Local Spline Regression based Segmentation

LSR-Seg [6] works in a semi-supervised way by estimating $\mathcal{Y}$ given $\mathcal{X}$ and $\mathcal{F}_l$: $\{\mathcal{X}, \mathcal{F}_l\} \rightarrow \mathcal{Y}$. To achieve that, a local spline regression function is utilized to map one pixel's neighboring features to their labels and the regularized losses between these pixels' true labels and their corresponding mapped labels are minimized. Fitting a spline function for each pixel and minimizing all the local losses leads to a global loss which can be written in matrix format as $\mathbf{y}^T\mathbf{M}\mathbf{y}$ with $\mathbf{y} = [y_1, y_2, \dots, y_n]^T$. More details can be found in [6].

Aside from the global loss, the loss on the labeled points is also considered. Hence, the labels for all pixels are obtained by $\min_{\mathbf{y}}(\mathbf{y}^T\mathbf{M}\mathbf{y} + \gamma\sum_{j=1}^{m}(f_{l_j} - y_{l_j})^2)$, which can be rewritten as:

$$\min_{\mathbf{y}}(\mathbf{y}^T\mathbf{M}\mathbf{y} + \gamma(\mathbf{f} - \mathbf{y})^T\mathbf{D}(\mathbf{f} - \mathbf{y})), \tag{2}$$

where $\mathbf{f} = [f_1, f_2, \dots, f_n]^T$ and $\mathbf{D}$ is a diagonal matrix with 1 for labeled pixels and 0 for unlabeled pixels. Each element of $\mathbf{f}$ is assigned as 1 or -1 if it is labeled, and 0 otherwise. The first term in Eqn. 2 constrains the estimated labels being smooth between neighboring pixels. The second term makes the estimated labels of the m labeled points approximating their previous assigned values. γ is a trade-off parameter.

The LSR-Seg has been proven efficient in interactive image segmentation with human supplied scribbles. However, the scribbles are critical to its performance with different scribbles leading to possibly distinct results as shown in Fig. 1. To obtain accurate segmentation, it often needs to supply appropriate scribbles, which is not applicable in video segmentation. In this paper, it is conducted for face segmentation with automatic prior information instead of human scribbles.

3. ACTIVE LEARNING BASED FACE SEGMENTATION

In this section, we first formulate the active learning based query point sampling process. How these query points are automatically labeled is then illustrated.

3.1. Active Learning based Sampling

In machine learning, query points are actively sampled with a suitable strategy to improve the model's performance [11]. Different scenario calls for different sampling strategy. In the context of video segmentation, the sampled points provide labeling information for LSR-Seg. As demonstrated in Fig. 1(a), insufficient labeling of the hair region leads to it being treated as face. Fig. 1(b) shows that the lack of labels of skin pixels around the eye region yields face holes in the eye and eyebrow regions. Thus, the query points, to be automatically selected, should cover these critical positions for discriminating face and non-face.

Before segmentation, no prior is known about face and non-face. Thus, they are assumed to be equally distributed in the image. Accordingly, a dense and random sampling strategy is utilized, which randomly selects up to 10% pixels from the segmentation window as query points. The dense and random sampling makes the critical positions covered with a high possibility. The information could be propagated even one pixel is sampled and labeled for each critical region. These selected points will then be automatically labeled. The sampling and labeling process is iterated to improve the segmentation performance. The repeat of the process makes those critical and discriminative information gradually included.

3.2. Automatic Labeling

The selected query points are required to be labeled for the latter semi-supervised LSR-Seg. Since the segmentation part does not consider labeling noise, the labeling information will be propagated via the data graph no matter it is correct or not. Thus, the sampled query points are needed to be either labeled correctly or not labeled. Here, they are automatically labeled utilizing both color and depth information. A skin color detector is adopted to identify possible skin and nonskin pixels. To reduce the possible labeling outliers of skin detection, the depth information is employed.

The skin color detection is accomplished via a linear regression tree, due to its good generalization ability and discriminative ability [12]. At each internal node of the tree, linear regression between feature $\mathbf{x}$ and label y is performed and a split is conducted with the regressed value. Each leaf node encodes the likelihood for skin $P(\mathbf{x}|y=1)$ and nonskin $P(\mathbf{x}|y=-1)$. The split functions at internal nodes and likelihoods at all leaves are learned offline from a training data set. At online stage, each query point denoted as $\mathbf{x}$ is dropped down the tree until reaching one leaf node and is labeled as

$$d(\mathbf{x}) = \begin{cases} 1, & \dfrac{P(\mathbf{x}|y=1)}{P(\mathbf{x}|y=-1)} > \alpha \\ -1, & \dfrac{P(\mathbf{x}|y=-1)}{P(\mathbf{x}|y=1)} > \alpha \\ 0, & otherwise \end{cases}, \quad (3)$$

where $d(\mathbf{x}) = 1$ and -1 represent that $\mathbf{x}$ is a skin and nonskin point, respectively. 0 means unlabeled, α controls different prior for skin and nonskin points.

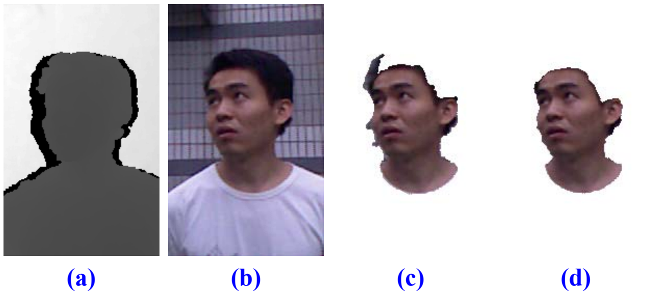

Fig. 3. (a): depth image. (b): RGB image. (c): the head region after depth constraining. (d): final result by the proposed result.

The depth information is utilized since it is invariant to illumination changes. Besides, the skin and its analogous counterpart in background may encode different depth values. Given the skin detection result, the coarse face region is obtained with depth for constraining. All the edges are first detected from the depth image and the depth image is labeled into several regions with that. The region with most query skin points is treated as the region containing human. Then the head region is obtained with the aid of color information. An over-segmentation process is conducted on the RGB image leading to many over-segmentation regions [2, 13]. Inside the human region, all the over-segmentation regions with the ratio of skin detection points larger than a threshold are merged and treated as the head region. To make the constraint reliable, dilation and erosion are performed on the head region to constrain the skin and nonskin query points. For clarity, two things should be noted. First, the depth image is coarse with many holes around the object boundaries, thus the edges are not accurate. Second, the over-segmentation result is also not so faithful. Based on those, the resulting head region is accurate as shown in Fig. 3. Accurate segmentation calls for later process with the coarse region as a constraint. After LSR-Seg, accurate face is obtained which can be seen from Fig. 3.

4. EXPERIMENTS

In this section, a set of experiments are conducted to verify the efficiency of our proposed face segmentation method. The experimental settings are first illustrated, then the results and analysis are reported.

4.1. Experimental Settings

The proposed face segmentation method has been tested on videos captured with Kinect which provides us both RGB and depth information with a pair of calibrated color and depth sensors. We collect videos under different scenes. The first is in a laboratory scenario under a simple lighting condition but with a confusing background (s1). The second one was grabbed outdoor with a building wall as background (s2). Another two difficult scenes are also tested.

To validate our proposed method, we also evaluate the performance of GrabCut [4] and LSR-Seg [6] with only the color feature. Since depth is also considered in our approach, an interactive method denoted as GrabCutD [14] with both color and depth is also compared. For our method, only a rectangle containing face is given in the first frame since there is no need to segment face in the whole frame. For the skin detection, several images are randomly captured and labeled to train the detector. On each training image, 3000 skin pixels and 5000 nonskin pixels are randomly selected out to train the skin detector. For all videos, γ is experimentally set to 10000 and α is set to 1.5. The parameter k in [14] manipulating the influence of two cues is set to 80% according to our experiences.

4.2. Results and Analysis

Given the initial face region at first frame, the segmentation method is conducted automatically. The position in the current frame is obtained from the segmentation results of last frame. LSR-Seg, GrabCut and GrabCutD are all run with given scribbles or rectangles.

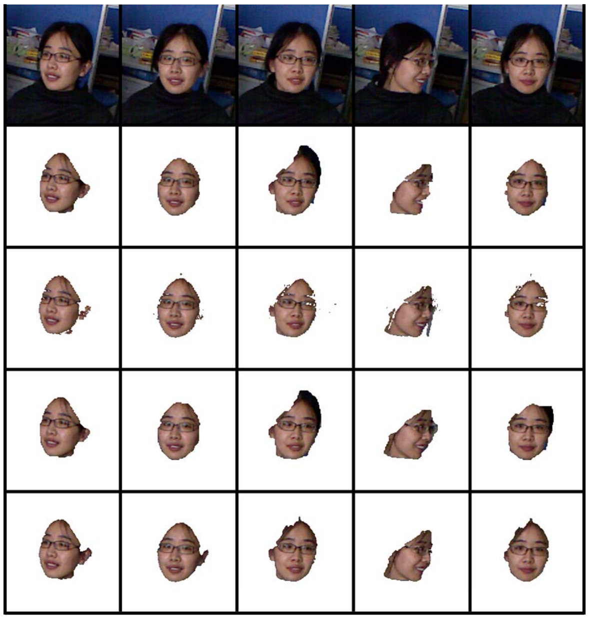

Fig. 4. Results for the indoor video. The first row is the segmentation windows. From the second to the fifth rows are the results of GrabCut, LSR-Seg, GrabCutD, and our method. From left to right are frames 51, 65, 80, 112, 134.

The results on the first indoor sequence are displayed in Fig. 4. All the methods could extract face region under different poses. However, different performances are reached. GrabCut and GrabCutD shows similar performances with possible outliers in the hair region. They both learn Gaussian Mixture Model [8] with the human given rectangle and measure how possible each pixel belongs to face.

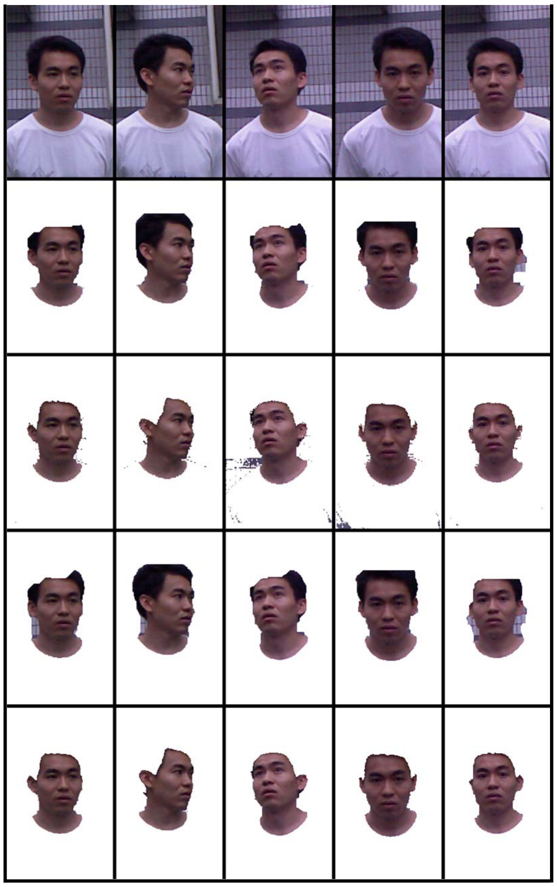

Fig. 5. Results on the outdoor video. The first row shows the segmentation windows. From the second to the fifth rows are the results of GrabCut, LSR-Seg, GrabCutD, and our method. From left to right are frames 32, 89, 104, 141, 151.

The learned model is not so discriminative leading to the outliers. GrabCutD obtains coarser boundaries than GrabCut. That is due to that the coarse depth information decreases its accuracy around face borders. LSR-Seg is prone to reach holes in the face region. The scribbles not covering all the critical positions induce those holes. Comparatively, the extracted regions by our method are smooth in the boundaries and robust to the hair and eye parts. The smooth boundary comes from the efficiency of local spline regression based segmentation. Complete and accurate results especially around hair and eyes benefit from effective prior information. Critical prior information with the LSR-Seg leads to our final performance.

How these methods perform on the recorded outdoor sequence can be seen in Fig. 5. The main difficulties of this scene are the large pose changes and the background wall with analogous color as skin. Akin as the indoor scene, the segmented faces of GrabCut and GrabCutD are inaccurate and GrabCutD works worse than GrabCut. LSR-Seg works better than them due to the more efficient human interaction. However, many outliers exist in the results. Our method performs approximately as LSR-Seg but containing less outliers.

We also evaluate these methods on two other different scenes with difficult illumination conditions and partial occlusion. Due to the space limit, only several snapshots are displayed in Fig. 6. Our method achieves more accurate and robust results than GrabCut and GrabCutD. Comparative performances as LSR-Seg are reached validating its efficiency for automatic segmentation.

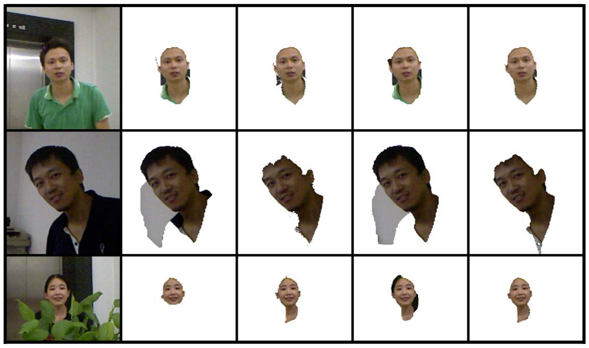

Fig. 6. Results on the two other scenes. From left to right are: segmentation window, GrabCut results, LSR-Seg results, GrabCutD results and our results.

Table 1. The F-score for two scenes

	frame	71	79	85	105	137
s1	GrabCut	0.9771	0.958	0.8963	0.9318	0.9471
	LSR-Seg	0.9725	0.971	0.9607	0.9509	0.9556
	GrabCutD	0.9767	0.961	0.8986	0.9529	0.9174
	Our	**0.9814**	**0.989**	**0.9803**	**0.9720**	**0.9688**
s2	frame	17	36	156	175	188
	GrabCut	0.9435	0.8271	0.9367	0.8920	0.9294
	LSR-Seg	0.9837	**0.9852**	0.9537	0.9744	**0.989**
	GrabCutD	0.9157	0.8061	0.9237	0.8739	0.911
	Our	**0.9902**	0.9805	**0.9913**	**0.983**	0.984

To quantitatively evaluate our method, we randomly selecte several frames from the test videos and manually labeled the groundtruth. The F-scores for the four methods are measured:

$$F = 2 * \frac{precision * recall}{precision + recall}. \tag{4}$$

The $precision$ and $recall$ are the precision and recall of the segmentation result, respectively. The score values of these methods are listed in Tab. 1. Our method reaches to higher score than GrabCut and GrabCutD which can also be observed from the above figures. It achieves almost the same results with LSR-Seg illustrating its efficiency. It should be noted that a slight advantage of F-score corresponds to a big difference on the segmented face region.

5. CONCLUSION

An automatic face segmentation method was proposed. It relied on color and depth information with no human intervention. Comparative experiments with several methods demonstrated its efficiency in video segmentation. Accurate face regions were obtained under different poses, scales and clustered background. These owe to the fusion of the two cues for providing semantics about face. Future work will address its intensive use on more challenging scenarios.

6. REFERENCES

[1] H. Li and K.N. Ngan, "Automatic video segmentation and tracking for content-based multimedia services," *IEEE Commun. Mag.*, vol. 45, pp. 27–33, January 2007.

[2] P. Meer D. Comanicu, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, pp. 603–619, May 2002.

[3] Jianbo Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, pp. 888–905, August 2000.

[4] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake, "Grabcut–interactive foreground extraction using iterated graph cuts," in *Proc. ACM SIGGRAPH*, 2004, pp. 309–314.

[5] Hongliang Li, King N. Ngan, and Qiang Liu, "Faceseg: automatic face segmentation for real-time video," *IEEE Trans. Multimedia*, vol. 11, pp. 77–88, January 2009.

[6] Shiming Xiang, Feiping Nie, and Changshui Zhang, "Semi-supervised classification via local spline regression," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, 2010.

[7] Xue Bai and G. Sapiro, "A geodesic framework for fast interactive image and video segmentation and matting," in *Proc. IEEE Int'l Conf. Computer Vision*, 2007, pp. 1–8.

[8] Michael J. Jones and James M. Rehg, "Statistical color models with application to skin detection," *Int'l J. Computer Vision*, vol. 46, pp. 81–96, January 2002.

[9] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A survey on pixel-based skin color detection techniques," in *PROC. GRAPHICON*. IEEE, 2003, pp. 85–92.

[10] Ying Ren and Chin Seng Chua, "Bilateral learning for color-based tracking," *Image and Vision Computing*, vol. 26, no. 11, pp. 1530 – 1539, 2008.

[11] Burr Settles, "Active learning literature survey," *Computer Sciences Technical Report*, 2010.

[12] Jixia Zhang, Haibo Wang, Franck Davoine, and Chunhong Pan, "Skin detection via linear regression tree," in *Proc. IAPR Int'l Conf. Pattern Recognition*, 2012, pp. 1711–1714.

[13] B. Georgescu P. Meer, "Edge detection with embedded confidence," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, pp. 1351–1365, December 2001.

[14] Z. Tomori, R. Gargalik, and I. Hrmo, "Active segmentation in 3d using kinect sensor," in *Proc. Int'l Conf. Computer Graphics, Visualization and Computer Vision*, 2012.