FAST HUMAN SEGMENTATION USING COLOR AND DEPTH

Raushan Kumar[1], Rakesh Kumar[1], Viswanath Gopalakrishnan[1], Kiran Nanjunda Iyer[1]

[1]Samsung R&D Institute, Bangalore, India

ABSTRACT

Accurate segmentation of humans from live videos is an important problem to be solved in developing immersive video experience. We propose to extract the human segmentation information from color and depth cues in a video using multiple modeling techniques. The prior information from human skeleton data is also fused along with the depth and color models to obtain the final segmentation inside a graph-cut framework. The proposed method runs real time on live videos using single CPU and is shown to be quantitatively outperforming the methods that directly fuse color and depth data.

Index Terms— User Extraction, Human Segmentation, Color, Depth, Kinect.

1. INTRODUCTION

The advances made in video and networking technologies in the past decade have resulted in a plethora of applications enabling immersive video experiences, and there-by making distance between humans irrelevant to a large extent. Real time segmentation of humans from a live video is most crucial component in many immersive video applications like telepresence solutions, virtual meetings, gaming etc. The availability of color and depth information from Kinect like sensors plays a major role in tasks involving accurate real time foreground segmentation. Our work is aimed at segmenting humans from a live video using color and depth information coming from a Kinect like sensor.

Segmentation of humans from unconstrained background scenes is a challenging task owing to many factors like high variance of foreground (FG) and background (BG) color distributions, similarity of FG and BG color distributions, variations in human poses and chances of occlusions. In the proposed work, we develop a system that uses color and depth data to form multiple hypotheses on probable foreground regions. We input these multiple foreground information coming from color and depth data into a graph-cut based pixel segmentation framework. We improve the segmentation quality using skeleton information (from Kinect) by modelling the depth and then refining the segmentation using this model. We further enhance the quality of the segmentation and smooth the flickering noise, produced because of frame by frame segmentation, using motion based compensation. We achieved a very high accuracy ~97.5% (F0.5 measure) on our data set comprising of person/s in different poses in different background settings. We are also able to run the engine almost real time ~ 38 milliseconds per frame (640x480) using single CPU.

2. RELATED WORK

In recent years with advances in depth sensing technology, focus on use of depth in Video and image analytics has increased. Many recent works in human segmentation have also focused on leveraging the depth data along with the color data information. In [6] Greff et al. compare different background subtraction algorithm using depth cues. Del Blanco et al. [2] discussed a system in which results from MoG-based background subtraction and Bayesian network-based foreground/background (FG/BG) prediction is combined using probability based FG/BG estimation. But the color cues haven't been utilized in the above methods. Work on background subtraction based on color and stereo depth by S. Javed et al [8] present depth extended OR-PCA (Online Robust Principal Component Analysis) and experimentally show that segmentation on color cues can be enhanced using depth cues. In [5] Fernandez-Sanchez et al proposed an adaption of Codebook Model to use depth as an additional channel with RGB. Ling Ge et al [10] presented a unified hierarchical Graph cut framework which uses both color and depth cue to segment foreground based on user interaction. Euclidean distance on color channels and geodesic distance on depth channel is used. To accelerate the algorithm, a hierarchical strategy is used for graph cut from coarse to fine segmentation. This method needs manual input which is not suitable for video segmentation. In [16], S. Chandra et al developed a unified deep CNN (Convolutional Neural Network) framework combining Deeplab Network [9] for semantic segmentation in RGBD images and Alexnet-HHA network [17] for object detection in depth images, for segmenting human limbs. The performance of the proposed method is ~ 8 fps on GPU which can act as deterrent for real time applications. M. Zhao et al in [11] present a GPU based framework for real-time object extraction, in which they first process the depth data to remove noisy/shadow regions. Color and depth based masks are adaptively combined to generate a tri-map which is further used in closed form temporal matting. However this method does not uses any prior information about humans (skeleton) as compared to our proposed method. In this paper we present a 3-stage method for segmenting humans from live video in almost real time, using only single CPU by fusing RGB & Depth cues with skeleton information in a unified framework.

3. HUMAN SEGMENTATION: OUR APROACH

In the proposed system, we overcome the shortcomings of only color or only depth based foreground extraction by fusing both color and depth in a unified framework. Depth data, obtained from commercial depth sensors, can have artifacts like shadow regions at the boundaries of object and noise regions where depth couldn't be measured. Noise regions are more prominent if an object has fuzzy structure (like hair), reflective texture. Also the depth data is

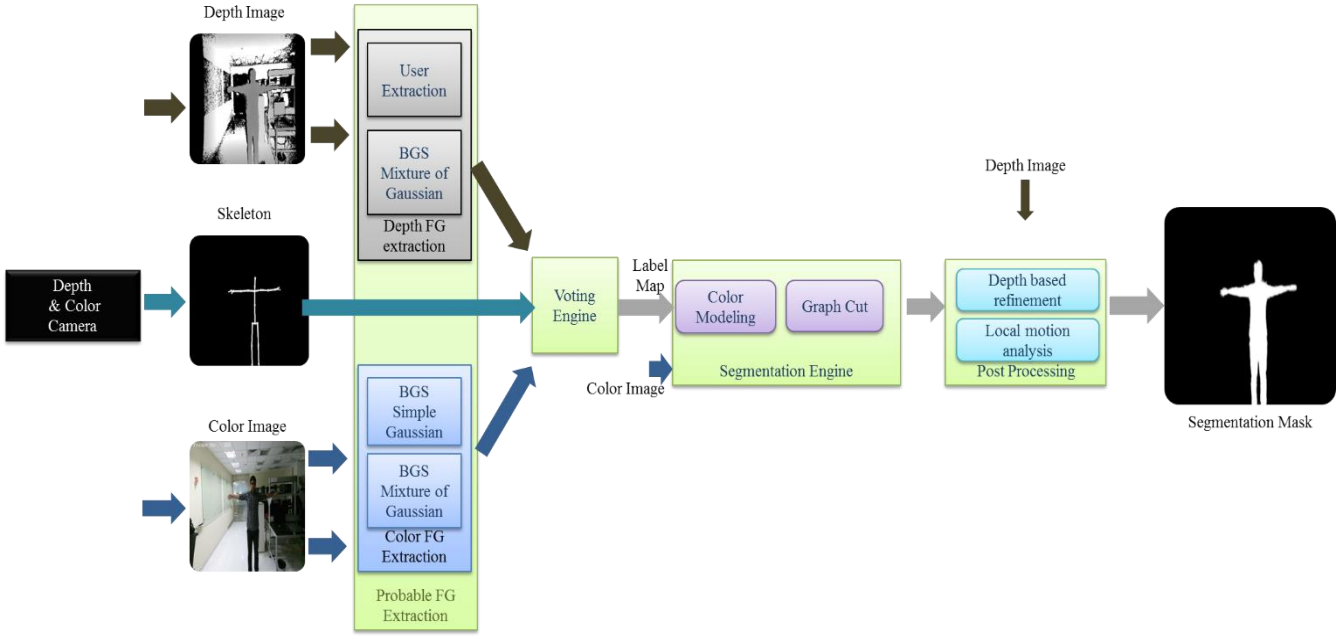


Fig.1. Block Diagram of proposed system for real time human Segmentation

not reliable if the object is transparent in nature. For these reasons, performing segmentation using only on depth data can give undesirable results. Similarly, segmentation using only color data can give inaccurate results if the FG and BG regions have some color similarity. Depth data used along with color data will provide complementary information useful for FG segmentation. We differentiate the proposed method from previous methods that fuse color and depth data by providing multiple hypotheses for probable FG regions and using the human skeleton information in the unified framework.

In the proposed system, segmentation is performed in three stages - Probable foreground extraction stage, Multi-hypothesis based segmentation stage & Post processing stage. In Probable foreground extraction stage four foreground masks are generated – two on each color and depth using foreground-background segregation techniques. This foreground-background information together with the skeleton information (from Kinect) is used to assign foreground probabilities for each pixel. In Multi-hypothesis based segmentation stage the current frame's color data are modelled as FG and BG using the assigned probabilities. Using these models, energy minimization is done in a graph cut framework as discussed in [4]. Due to shrinking bias, the segmentation mask may have short comings –small BG regions surrounded by FG (e.g. area between arms and waist when they are close) can get segmented as FG. Also, any other object which is in motion might get segmented as foreground. These shortcomings of the segmentation stage are perfected in the post processing stage. Depth data of the segmentation are analyzed and depth based refinement is done to remove small connected backgrounds having different depth. Also a Skeleton based masking is done to filter out the non-humans (if any) from the final mask. Since performing a frame by frame segmentation will produce temporal incoherency on the segmentation boundaries, localized motion based compensation is introduced to smooth the transition of boundaries across frames. A top level block diagram of our system is shown in Fig.1.

3.1. Probable foreground extraction

3.1.1. Color based probable foreground extraction

Probable FG extraction is done by modeling the background with a single multivariate Gaussian model as discussed in [19]. The model is updated at a learning rate alpha α. For each pixel, if the Mahalanobis distance between the pixel and its Gaussian is higher than a threshold, it is classified as FG. Since a single Gaussian is used, long term history of the background is modelled well. The background is also analyzed using multi-modal Gaussian (MoG) [1], [20] & [21] which helps to model the short term history more accurately. As discussed in [20], the background is modelled as multiple Gaussians with adaptive weights and the Gaussians are ranked on weight to variance ratio of the model. The parameters of the Gaussians are updated dynamically as well as the number of Gaussians is constantly adapted. Using the above methods, two FG masks are obtained on color cues.

3.1.1. Depth based probable foreground extraction

For depth based foreground extraction, the background is analyzed to define initial seeds for the FG and the seed regions are expanded using BFS based region growing algorithm to extract the FG. The seed pixel extraction is done in 3 steps -Floor removal, Bad-pixel analysis and Seed extraction. Floor Modelling is based on the hypothesis that the depth values for the floor will have a gradient in vertical direction, whereas the objects perpendicular to the floor will not have any gradient. Initial floor points are gathered for the bottom half of the frame based on above hypothesis. Finally, a 3-D plane is fit on the floor pixels using RANSAC [3] algorithm. The floor pixels are not considered as seed pixels. As Kinect depth data can have noise, Bad Pixel Analysis is done by observing pixel's history. If a pixel has zero depth for more than pre-set duration, then it is classified as bad pixel. These are also excluded from the user seed points. For seed pixels selection, the depth of each pixel is modeled and is compared against current depth at that pixel. The depth value of a pixel in model is updated if the corresponding current depth of the pixel is more than the depth in the model. Otherwise, if the pixel's current depth is less than that of its model

by some threshold, it is considered as seed pixel. Using these extracted seed pixels, region growing is performed to extract the FG mask. For multiple disconnected regions, the projections of the regions on XOZ (Horizontal) and XOY (Vertical) planes are analyzed and the regions are merged if the Hausdorff distance between these projections is less than some threshold, thus providing a connected single mask for individual FG objects. Similar to MoG based background subtraction for color; a multimodal Gaussian based background subtraction is done for depth data as well. Here the learning of background is restricted only to the pixels classified as background by the depth based user extraction. This MoG on depth data provides improvement over the depth based user mask.

3.2. Multi-hypothesis based segmentation

Using FG masks obtained in the probable foreground extraction stage, each pixel is assigned label *L,* by the voting engine, as –

$$L = \begin{cases} Definite\ FG, & s = 4 \\ Probable\ FG, & s = 3 \\ Unknown, & s = 2 \\ Probable\ BG, & s = 1 \\ Definite\ BG, & s = 0 \end{cases}$$

$$s = \sum_{i=1}^{4} Mask_i$$

Where $Mask_i = 1$, if the pixel is labelled FG, otherwise 0. Skeleton, obtained by joining the skeleton points (from Kinect), can always be assumed to be on person and hence are labelled as *Definite FG* strengthening our model for FG. The generated label map and the color image are used in graph cut frame work to obtain globally optimum segmentation as is discussed in [4]. The labels assigned to the pixels are used to model the data as foreground and background with respect to the color image. The definite FG and probable FG pixels are used to model the foreground color as a mixture of Gaussians with k components. Similarly, definite BG and probable BG pixels are used to model the background as a mixture of Gaussians with k components. A graph $\zeta = (\upsilon,\varepsilon)$ is defined with a set of vertices (υ) and edges (ε). Each edge $e \in \varepsilon$ is assigned a non-negative weight ω_e. Each node, $v \in \upsilon$, represents a pixel in the image. Additionally two special vertices – terminal nodes (for FG and BG) are introduced. An 8-connected graph is considered in which the edges represents the relationship between the neighboring nodes. The edges between the pixels (normal vertices) – N-links, represents the connectivity between the two pixels based on their intensity differences. The terminal nodes are connected to all pixels. The edges between a terminal node and a pixel (normal nodes) – T-links, represent the affinity of that pixels towards terminal node (FG or BG). The T-links are computed as a sum of probabilities defined by the PDF of Gaussians. We estimate the segmentation using a min-cut on the constructed graph. In our formulation we assign very high T-link weights to pixels belonging to definite FG and definite BG, thus ensuring they belong to FG and BG respectively (as estimated in probable foreground extraction stage) after the cut. Since the probable FG and probable BG pixels were used to model the color data, hence T-link's edge weight for those pixels gets higher weight (for FG and BG nodes correspondingly) as compared to unknown pixels. Modelling the FG and BG as mixtures of Gaussians is done using k-means clustering which is computationally expensive. In a mostly static camera view, most of the background remains same. To take advantage of this we introduce a novel temporal reuse of color model. The label-map for the current frame is compared against the label-map of the previous frame and only for pixels having different label, as compared to the previous frame, the models are updated as shown in Fig.2. The temporal re-use of color models reduces the model computation time and also allows the information from previous frames to propagate. For further improving upon the time complexity, segmentation is performed on smaller resolution image. But up-sampling of results directly to higher resolution lead to jitter at the boundary of segmentation. Hence, another iteration of graph cut is performed on smaller overlapping blocks along the boundaries of the mask. The graph cut algorithm using such overlapping structures along the boundaries provides a smooth contour.

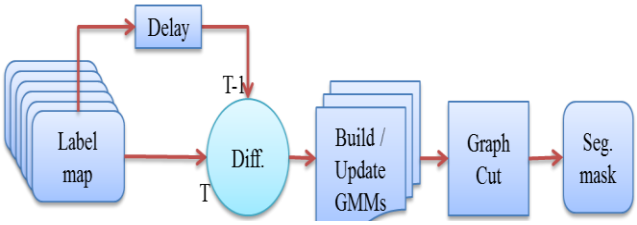


Fig.2. Graph cut with temporal re-use of color models

3.3. Post-processing

3.3.1. Depth based refinement

Since graph cut attempts to find a minimum cut, it can be biased towards producing smaller contour – shrinking bias [14]. The FG and BG can have small areas where the color difference is very less between them, leading to small background areas labelled as FG. To mitigate these issues, mostly arising due to similarity of color, depth based refinement is introduced. The depth values for the multi-hypothesis based segmentation stage's result are modelled as a Gaussian. Since the depth values for a person (from Kinect) is not inherently multimodal in nature, it can be modeled as a single Gaussian. The probability of each of the result is calculated using the PDF of the Gaussian. If the probability is more than a certain threshold, the pixels are retained in the results (Fig.3). For multiple users/persons we track each individual, using skeleton information, and model the depth data of each of them separately.

3.3.2 Motion based Compensation

The result of frame-by-frame segmentation looks jittery at the boundaries of the mask between frames. To present a smooth boundary transition across frames, a local-motion based compensation (LMBC) mechanism is introduced where in a

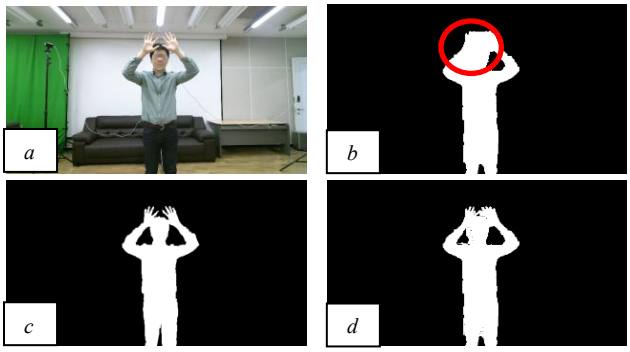


Fig.3. Depth based refinement - a. Color image, b. Multi-hypothesis based segmentation stage with error, c. Ground truth, d. Post processed result

weighted average of the current segmentation and previous segmentation is used as final mask. The weight of the previous frame depends on its difference from the current frame. But in case of high localized motion, a global weighted average will produce drag kind of effect. Hence the mask is divided into 25 blocks and on each block, based on the frame difference, change ratio, r is computed. If this ratio is less than a threshold, averaging based on the previous mask is done. Otherwise, the current foreground mask is retained.

$$r = \frac{\Delta n}{n_F}$$

$$w_t = \begin{cases} \frac{1 - w_o}{r_o} * r + w_o; & r \le r_o \\ 1; & r > r_o \end{cases}$$

$$M_t = w_t * M_t + (1 - w_t) * M_{t-1}$$

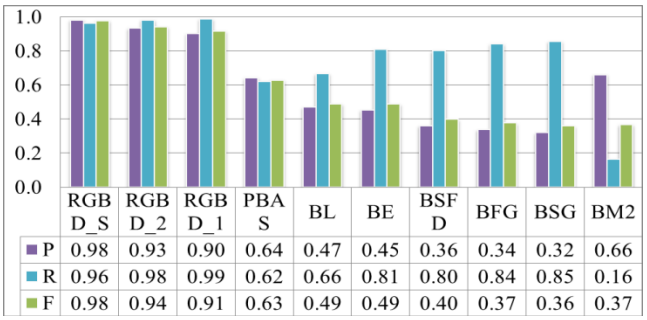
Where, Δn is the number of pixels changed in current mask M_t, having n_F foreground pixels, with respect to previous maskM_{t-1}, $r_o = 0.08$ is the threshold and w_t is the weight assigned to previous frames and $w_o = 0.03$ is a constant.

4. EXPERIMENTS & RESULTS

We recorded 8 videos of ~500 frames each, using Kinect V2, with different number of people (up to 3), in different poses & with different backgrounds. We randomly selected 15 frames from each of these videos for accuracy calculations. We used F-Measure for calculating overlap as given by the equation

$$F_\beta = \frac{(1 + \beta^2) \times recall \times precision}{recall + \beta^2 \times precision}$$

Emphasizing on precision, we used $F_{0.5}$ score (β= 0.5) giving more weightage to precision than recall. We compared our results (RGBD_S) with other methods - our method without skeleton information, using 2 color based mask and 2 depth based mask for energy minimization (RGBD_2), direct fusion of color and depth using 1 color based mask and 1 depth based mask (RGBD_1). We also compared our results with color based



	RGBD_S	RGBD_2	RGBD_1	PBAS	BL	BE	BSFD	BFG	BSG	BM2
P	0.98	0.93	0.90	0.64	0.47	0.45	0.36	0.34	0.32	0.66
R	0.96	0.98	0.99	0.62	0.66	0.81	0.80	0.84	0.85	0.16
F	0.98	0.94	0.91	0.63	0.49	0.49	0.40	0.37	0.36	0.37

Fig.4. Results comparison

Background subtraction methods discussed in [18] – PBAS[8], LOBSTER (BL)[15], EIGEN (BE)[13], Static frame difference (BSFD), Fuzzy Gaussian (BFG)[12], Simple Gaussian(BSG)[19], and Mixture of Gaussian(BM2)[1]. The result is summarized in Fig.4 & Fig.5. It is observed that the proposed method out performs state of the art color based background subtraction methods. Also our proposed method outperforms color and depth's fusion based methods significantly. The time for segmentation of 640 x 480 resolution image is ~ **38ms** on 3.4 GHz Intel Core i7 - 4770 CPU (using two threads and no GPU).

5. CONCLUSION

This paper presented a novel multi-hypothesis based segmentation technique to extract humans in real time using depth, color and skeleton information. Use of complementary information from different hypotheses helped to mitigate the shortcomings of working individually on color or depth data. Use of top level information about skeleton has served two fold purposes - re-acquiring body parts which could have been lost even in methods using both color & depth, and removing unwanted backgrounds that might have been labeled as foreground incorrectly. A novel local-motion based compensation was also introduced to minimize jitter/fluctuations at the boundaries of the segmentation across frames. The proposed system performs very high quality segmentation **($F_{0.5}$ score ~ 0.975)** in almost real time **(~38ms per frame)** without using GPU.

Fig.5. a. Color Image, b. Depth, c .Ground Truth, d. Depth and Color Fused (RGBD_2) e. RGBD_S(our method)

6. REFERENCES

[1] Bouwmans, Thierry, Fida El Baf, and Bertrand Vachon. "Background modeling using mixture of Gaussians for foreground detection-a survey." Recent Patents on Computer Science 1.3 (2008): 219-237.

[2] Del-Blanco CR, Mantecón T, Camplani M, Jaureguizar F, Salgado L, García N. "Foreground Segmentation in Depth Imagery Using Depth and Spatial Dynamic Models for Video Surveillance Applications". Sensors. 2014; 14(2):1961-1987.

[3] Derpanis, Konstantinos G. "Overview of the RANSAC Algorithm." Image Rochester NY 4 (2010): 2-3.

[4] Felzenszwalb, Pedro F., and Daniel P. Huttenlocher. "Efficient graph-based image segmentation."International Journal of Computer Vision 59.2 (2004): 167-181.

[5] Fernandez-Sanchez, E.J.; Diaz, J.; Ros, E. "Background Subtraction Based on Color and Depth Using Active Sensors". Sensors2013, 13, 8895-8915.

[6] Greff, Klaus, Brandão, André, Krauß, Stephan, Stricker, Didier and Clua, Esteban. 2012, "A Comparison between Background Subtraction Algorithms using a Consumer Depth Camera." Paper presented at the meeting of the VISAPP (1), 2012.

[7] Hofmann, Martin, Philipp Tiefenbacher, and Gerhard Rigoll. "Background segmentation with feedback: The pixel-based adaptive segmenter." 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2012.

[8] Javed, S.; Bouwmans, T.; Soon Ki Jung, "Depth extended online RPCA with spatiotemporal constraints for robust background subtraction," in Frontiers of Computer Vision (FCV), 2015 21st Korea-Japan Joint Workshop on , vol., no., pp.1-6, 28-30 Jan. 20155

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

[10] Ling Ge, Ran Ju, TOngwei Ren, Gangshan Wu; "Interactive RGBD Image Segmentation using hierarchical Graph Cut and Geodesic Distance", Advances in Multimedia Information Processing, PCM 2015, Volume 9314 of the series Lecture Notes in Computer Science pp 114-124, http://dx.doi.org/10.1007/978-3-319-24075-6_12

[11] M. Zhao, C. W. Fu, J. Cai and T. J. Cham, "Real-Time and Temporal-Coherent Foreground Extraction With Commodity RGBD Camera," in IEEE Journal of Selected Topics in Signal Processing, vol. 9, no. 3, pp. 449-461, April 2015. doi: 10.1109/JSTSP.2014.2382476

http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6990481&isnumber=7063282

[12] M. Sigari, N. Mozayani, H. Pourreza. "Fuzzy Running Average and Fuzzy Background Subtraction: Concepts and Application", International Journal of Computer Science and Network Security, Vol. 8, No. 2, pp. 138-143, 2008.

[13] N. M. Oliver, B. Rosario and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 831-843, Aug 2000. doi: 10.1109/34.868684

http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=868684&isnumber=18808

[14] Xu, Ning, Narendra Ahuja, and Ravi Bansal. "Object segmentation using graph cuts based active contours." Computer Vision and Image Understanding 107.3 (2007): 210-224.

[15] P. L. St-Charles and G. A. Bilodeau, "Improving background subtraction using Local Binary Similarity Patterns," Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on, Steamboat Springs, CO, 2014, pp.509-515doi:10.1109/WACV.2014.6836059

http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6836059&isnumber=6835728

[16] S. Chandra, S. Tsogkas and I. Kokkinos, "Accurate Human-Limb Segmentation in RGB-D Images for Intelligent Mobility Assistance Robots," 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), Santiago, 2015, pp. 436-442. doi: 10.1109/ICCVW.2015.64

http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7406413&isnumber=7406349

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

[18] Sobral, Andrews; Bouwmans, Thierry. "BGS Library: An OpenCV C++ Background Subtraction Library", IX Workshop de Visão Computacional (WVC'2013), Rio de Janeiro, Brazil, 2013.

https://github.com/andrewssobral/bgslibrary.

[19] Y. Benezeth, P.M. Jodoin, B. Emile, H. Laurent, C. Rosenberger. "Review and Evaluation of Commonly-Implemented Background Subtraction Algorithms". Proc. International Conference on Pattern Recognition, pp. 1-4, 2008.

[20] Z. Zivkovic and F. V. D. Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction", Pattern Recognition Letters, Volume 27 Issue 7, Pages 773-780, May 2006

[21] Zivkovic, Zoran. "Improved adaptive Gaussian mixture model for background subtraction." Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. Vol. 2. IEEE, 2004.