OUTLINE FUSION FOR HUMAN SEGMENTATION FROM A STATIC CAMERA

Rui Wang (rw2@ualberta.ca)
Jiaxin Xu (jiaxin7@ualberta.ca)
Vaibhav Rakheja (rakheja@ualberta.ca)

ABSTRACT

Topics related to Human segmentation and tracking in video sequences have become increasingly active due to high demand for real-life applications like video surveillance, action localization and virtual reality simulation. Thanks to the advancement in instance segmentation and object detection, human segmentation has become easier and more feasible. Previous methods used to partition the problem into three sub-questions, which are classification, object detection, and instance segmentation. And there are mainly two kinds of methods that are implemented from the top-down view and bottom-up view, respectively. Top-down methods, like[1][2] is implemented by learning the representation of a specific class and use the known shape characteristic to guide the segmentation process. Bottom-up methods, such as[3][4] mainly contains two steps. First, segment input image into regions of interest and then identify these regions with continuity of gray-level, texture and bounding contours. Human segmentation used to be a complex problem, but now this issue can be addressed by deep learning models. In this study, we performed human segmentation using three methods: In the first method, we employed Mask R-CNN to produce human segmentation mask; while in the second approach, we implemented YOLACT[5] which combines unified, Real-Time Object Detection (YOLO)[6] for object detection and Mark-R-CNN for mask generation; Lastly, we utilized YOLO output and use background subtraction for human mask.

***Index Terms*—** Human Segmentation, Video Processing, Object Recognition, Mask R-CNN, YOLACT

1. INTRODUCTION

The goal of human segmentation is to identify a human in an image or video and separate it from the background. There are two types of videos from which human images can be segmented: they are static cameras where background and foreground are static and moving cameras where the foreground and background changes over time. The methods deal with these two types of cameras differ a lot. In this paper, we mainly focus on human segmentation from static cameras.

Segmentation methods are classified into two categories: semantic segmentation and instance segmentation. Semantic segmentation is a technique that detects, for each pixel, the object category it belongs to without differentiating object instances, whereas instance segmentation not only identifies all objects in an image correctly but also segment individual instances. Our objective for this project is instance segmentation

The challenges associated with human segmentation lies in occlusion and strange poses, as well as real-time applications. In heavily occluded images or videos, many human segmentation models fail to precisely separate individual human object from the background. Additionally, due to the variation of human poses, it is difficult to fully segment a person without missing body parts. Furthermore, deep learning models for human segmentation requires high computation power, some of them are could not serve in real-time, for example, Mask R-CNN only achieves 5 fps [7].

In this report, our team proposed three approaches for human segmentation from a static camera. In the first approach, we applied Mask R-CNN for human segmentation. While, in the second method, we implemented YOLACT which introduces mask prediction on top of YOLO. Lastly, our team combined YOLO and background subtraction for mask prediction.

2. RELATED WORK

Segmentation Method could be classified into two categories: semantic segmentation and instance segmentation. For semantic segmentation, Long et al. proposed Fully Convolutional Network(FCN) [8] in 2016. Later, a semantic segmentation method based on Densely Connected Convolutional Networks(DenseNets)[9] is presented by Jégou et al[10]. They combined FCN and DenseNet to deal with semantic segmentation and achieved a good result on urban scene benchmark datasets. Semantic segmentation plays an important role in human-computer interaction, action localization and robotic field. Instead of segmenting objects in the same class as a whole, instance segmentation identifies and segments individual objects. There are three orders of execution for instance segmentation, which are segmentation-first, instance-first or implement the two processes simultaneously. In 2014, Hariharan and his groups presented a segmentation-first instance segmentation methods, called Simultaneous Detection and Segmentation[11]. They first implement proposal generation with Multi-scale Combinatorial Grouping(MCG)[12] to produce 2000 region candidates each image. Then, extract features using a convolutional network (ConvNet) and classification by Support Vector Machine (SVM). Kirillov et al. proposed a new modelling paradigm for instance-aware semantic segmentation, named InstanceCut[13]. They implemented an instance-agnostic semantic segmentation with standard ConvNet and extract instance-boundaries with a new instance-aware edge detection model. There is an instance-first method, presented by Dai et al.[14]. Their method consists of instance identification, mask estimation and classification formed a cascade structure. Fully convolutional instance-aware semantic segmentation[15] is a special approach very close to simultaneously segmentation, which performs instance masks prediction and classification jointly. The advancement of instance segmentation enables human segmentation.

There are many approaches to human segmentation. Some earlier research papers focus on extracting distinctive features from an image and the extracted features are used for object classification and segmentation, such as Histograms of Oriented Gradient (HOG) descriptors [16] and Scale Invariant Feature Transform (SIFT) [17]. There is an edge detection method called canny edge detection that outlines the boundary of a human in an image. Later, the pose-based human segmentation approach like [18], [19] and [20] are proposed. In [20] Zhang et al. detects human body key parts, and then build up a segmentation mask on top of the key parts. Up to now, the most accurate human segmentation approach first identifies the region or bounding box around the human image then perform segmentation, one of the examples is Mask R-CNN [7].

Besides segmentation models, our team also investigated on object detection methods. One of the bounding box detection methods we investigate is YOLO[6]. YOLO is a state-of-art, real-time detection method. Compared to conventional object detection methods that sliding classifiers into multiple locations in the test image, YOLO only looks at the test image once then predict possibilities of object existence in the entire image. For the reason that YOLO only processes the input image once, it is superior in speed compared with existing object detection methods.

3. OUR IMPLEMENTATIONS

Our team implemented three models: first of all, we applied Mask R-CNN for human detection and mask prediction; In the second model, we implemented YOLACT which adds mask prediction on top of YOLO; Additionally, we combined YOLO and background modelling which performed background subtraction on YOLO's detection results.

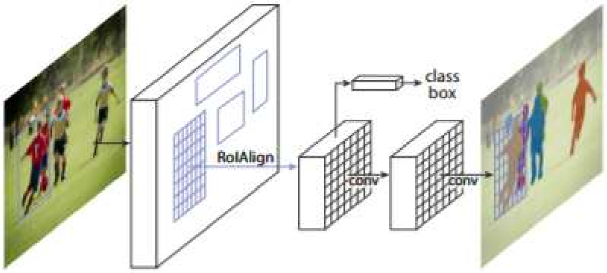


Fig. 1: An overview of Mask R-CNN proposed by Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick, "Mask r-cnn," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969

3.1. Mask R-CNN

Mask R-CNN[7] simultaneously detects and segments mask at high accuracy. An overview of Mask R-CNN is illustrated in Figure [7]. For object detection, the Mask R-CNN uses ResNet 101 to extract features from the images. The feature maps are input to Regional Proposal Networks which predicts if a human is present in the region and outputs the regions that contain humans. To get the universal size of proposal regions, a pooling layer converts all regions to the same shape. Later, these regions are passed through a fully connected network. Class labels and bounding boxes are produced. The object segmentation part of Mask R-CNN further segments the region of interest. It calculates Intersection over Union (IoU) with ground truth box. Once the region of interest is generated, Mask R-CNN uses fully convolutional networks (FCN) to create a mask on the intersection between the region of interest and ground truth image. The FCN avoids vector transformation that loses spatial dimensions. It operates on the input image and creates corresponding output with spatial information. To predict pixel-accurate masks, ROIAlign is applied to align feature maps of the region of interest. The disadvantage of Mask R-CNN is that the segmentation results are heavily dependent on detection performance. For example, the poor performance of detection on overlapped human images will negatively impact segmentation outputs.

3.2. YOLACT

You Only Look At CoefficienTs (YOLACT) is a fully convolutional framework for real-time instance segmentation which can achieve 29.8 mAP on Microsoft COCO data set at 33.5 fps on a single Titan Xp GPU.

YOLACT is much faster than existing competitive instance segmentation frameworks and is capable of generating high-quality instance masks. It takes advantage of FCN as a mask generating branch and adds this mask branch to an existing object detection network. YOLACT breaks instance segmentation into two parallel sub-tasks: (1) generating prototype masks over an entire input image, and (2) producing a set of per-instance mask coefficients. By implementing prototype mask generation and coefficients computation in a parallel mode, YOLACT avoids explicit feature localization step which contributes to generating high-quality masks and also greatly improves segmentation speed. Then, the instance masks are generated by linearly combining prototype masks and coefficients. After going through a sigmoid nonlinearity, the final masks are generated. The whole process of YOLACT follows a simple matrix multiplication and sigmoid:

$$M = \sigma(PC^T) \tag{1}$$

where P is a $w \times h \times k$ matrix of prototype masks and C is a $n \times k$ matrix of mask coefficients. The loss function is composed of three parts: classification loss L_{cls}, box regression loss L_{box} and mask loss L_{mask}. The former two loss sections are the same as SSD counterparts. And the mask loss function is shown as follows:

$$L_{mask} = BCE(M, M_{gt}) \tag{2}$$

where M demotes mask M and M_{gt} denotes ground truth mask. BCE is short for Binary Cross Entropy.

3.3. YOLO and Background Modeling

YOLO is a state-of-art, real-time detection method. YOLO only looks at the test image once then predicts possibilities of object existence in the entire image. The model is constructed upon 24 convolutional layers and 2 fully connected layers. The input image into the model is divided into S*S grids. Each grid is responsible for predicting B bounding boxes. The bounding box contains information, including object class, and y coordinates, and the height and width of the bounding box[6]. The test image is input into convolutional layers with maxpooling layers to generate probabilities of human presence, bounding box coordinates and bounding box size. YOLO uses Non-Maximal Suppression (NMS) to remove bounding boxes with a low probability. The box with probability lower than a defined NMS threshold will be eliminated, indicating it unlikely contains an object. After removing low probability bounding boxes, the box with the highest probability values will be selected. Other boxes with high similarity as the best bounding box are taken out based on Intersection Over Union (IOU) value. It ensures the final result will only include one best-predicted bounding box for each class. It is superior in speed compared with existing object detection methods.

Our team proposed to apply background and foreground subtraction on the output image from YOLO. With YOLO generating bounding boxes for humans, background modelling can directly apply on the part of the image which is circumscribed by YOLO's bounding box. By doing it, segmentation only happens on humans even there are other objects present in the image. Background modelling utilized detection results from YOLO and generated human masks using YOLO's output. There are two techniques that we used for background modelling: frame differencing and Mixture of Gaussians (MOG). For frame differencing, we first estimated the background frame which was assumed to be the previous frame, then subtracted the current frame from the background frame, and applied Otsu Thresholding to the absolute difference between the current frame and the background frame to obtain a foreground mask. The frame differencing is calculated using Equation 3. The drawback of frame differencing is that it is very sensitive to the threshold. The second technique is MOG. The MOG method is more complex than frame differencing. It tracks multiple Gaussian distributions simultaneously assuming foreground and background follow Gaussian distributions. Each pixel in the image, $X_1 ... X_t$, is modelled by y a mixture of K Gaussian distributions. The probability of the current pixel is included in Equation 4.

$$|Frame_i - Frame_{i-1}| > Th \tag{3}$$

where Th is the threshold.

$$P(X_t) = \sum_{n=1}^{K} \omega_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}) \tag{4}$$

where K is the number of distributions,$\omega_{i,t}$ is an estimate of the weight of the ith Gaussian in the mixture at time t,$\mu_{i,t}$ is the mean value of the ith Gaussian in the mixture at time t, $\Sigma_{i,t}$ is the covariance matrix of the ith Gaussian in the mixture at time t, and eta is a Gaussian probability density function in Equation 5. The premise of the MOG is that red, green, and blue pixel values are independent and have the same variances. The model parameters are updated using Expectation Maximization algorithms. Eventually, the MOG will find optimal model parameters that best separate background and foreground [21].

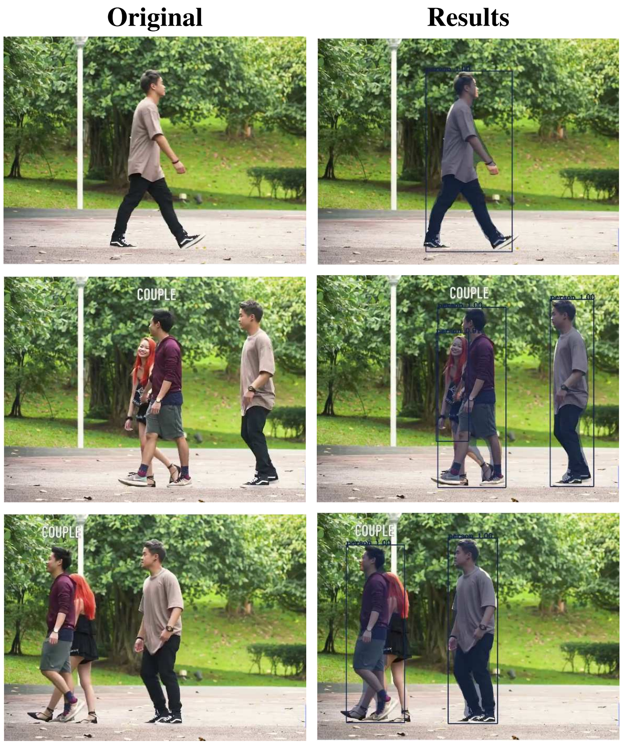
$$\eta(X_t, \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_t - \mu_t)^T \Sigma^{-1} (X_t - \mu_t)} \tag{5}$$

4. EXPERIMENT AND COMPARISON

The data set we are using is change detection dataset from [22] . Our team generated results from all three models of the proposed method. The results are included in Figure 2 3 and further explained in following sections.

4.1. Mask R-CNN

The output from Mask R-CNN is illustrated in Figure 2. It shows that Mask R-CNN predicts bounding boxes and binary


Original

Results

Fig. 2: Mask R-CNN Results

masks on the individual human object. The results show that the Mask R-CNN is not able to fully outline details of human body parts: missing part of the head (first row in 2 or missing part of feet (second row in 2.The Mask R-CNN also has difficulties with occluded objects: part of occluded human is not fully segmented as shown in the third row in 2. Although the Mask R-CNN has some drawbacks, it detects a person accurately and generates a human mask correspondingly.

4.2. YOLACT

The testing results of YOLACT is shown in Figure 3. YOLACT provides both bounding box and segmentation results of input targets. In this experiment, we test YLACT with different lighting conditions, and record the performance of YOLACT when dealing different human density and human pose. When comparing the second row of Figure 3, we can see that the bounding box of YOLACT can localize target object correctly under different lighting conditions, human density and the results are not effected by camera angles. But when two persons stay very close or overlap with each other, there appears to be detection error as shown in the first row of Figure 3. The results illustrate that there are two persons in the image but YOLACT displays three bounding boxes. And in the second row of Figure 3, the mask failed to cover part of the standing man's legs. The overall performance of YOLACT is good, but there's still space for upgrading.

4.3. YOLO and Background Modeling

The segmentation results from YOLO and background modelling are illustrated in Figure 3. In Figure 3, FD refers to segmentation results from frame differencing, and MOG refers to results from the MOG model. Background modelling performs better with a single person compared to multiple persons present in an image. Also, the model generates better results for persons closer to the camera compared to persons farther away from the camera. The MOG is a more complex model compared to frame differencing and is expected to achieve better results. However, in Figure 3, frame differencing provides a better segmentation mask with fewer missing body parts than the MOG. The possible explanation is that MOG implies the background and foreground distribution are Gaussian which are not always true [23]. For example, indoor scenes are much closer to the Laplace model than Gaussian distribution[24]. The other reason is that initializing Gaussians is very important. However, the MOG model used in this project is randomly initialized. Therefore, frame differencing has better performance than the MOG model.

5. CONCLUSION

We implement three different methods to deal with the human segmentation subject. Overall, Mask R-CNN ranks at the top in terms of accuracy but failed to serve in real-time applica-

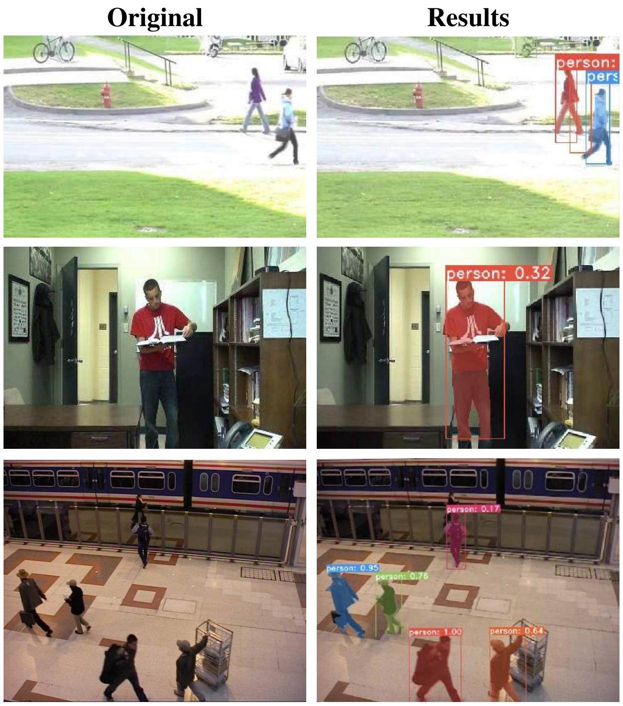


Fig. 3: YOLACT Results

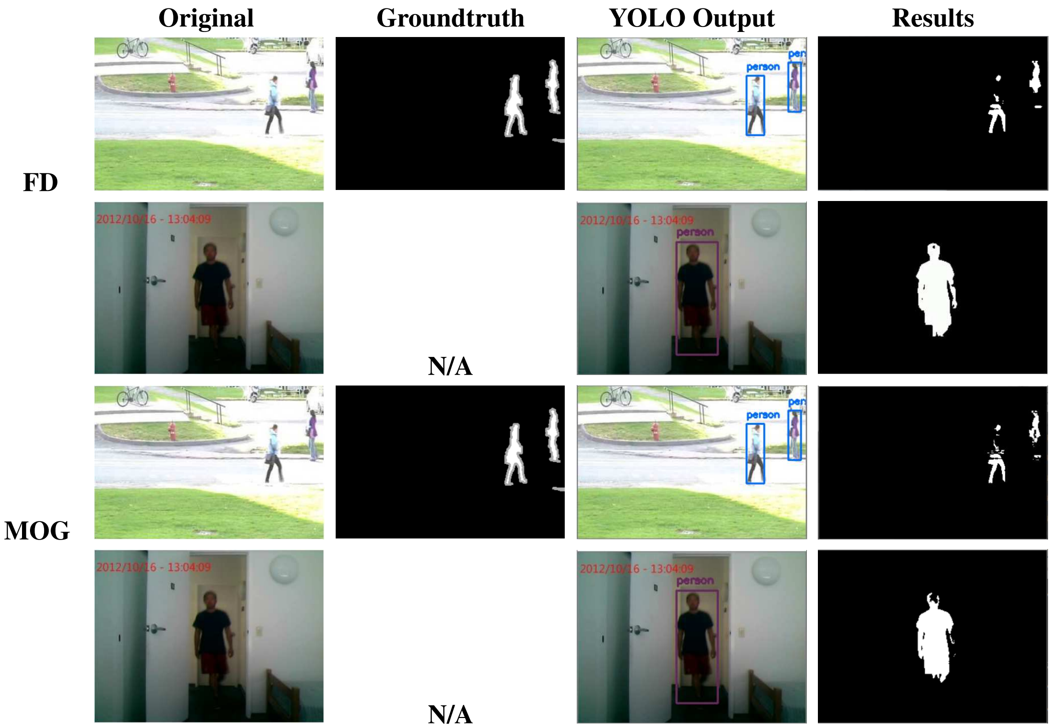
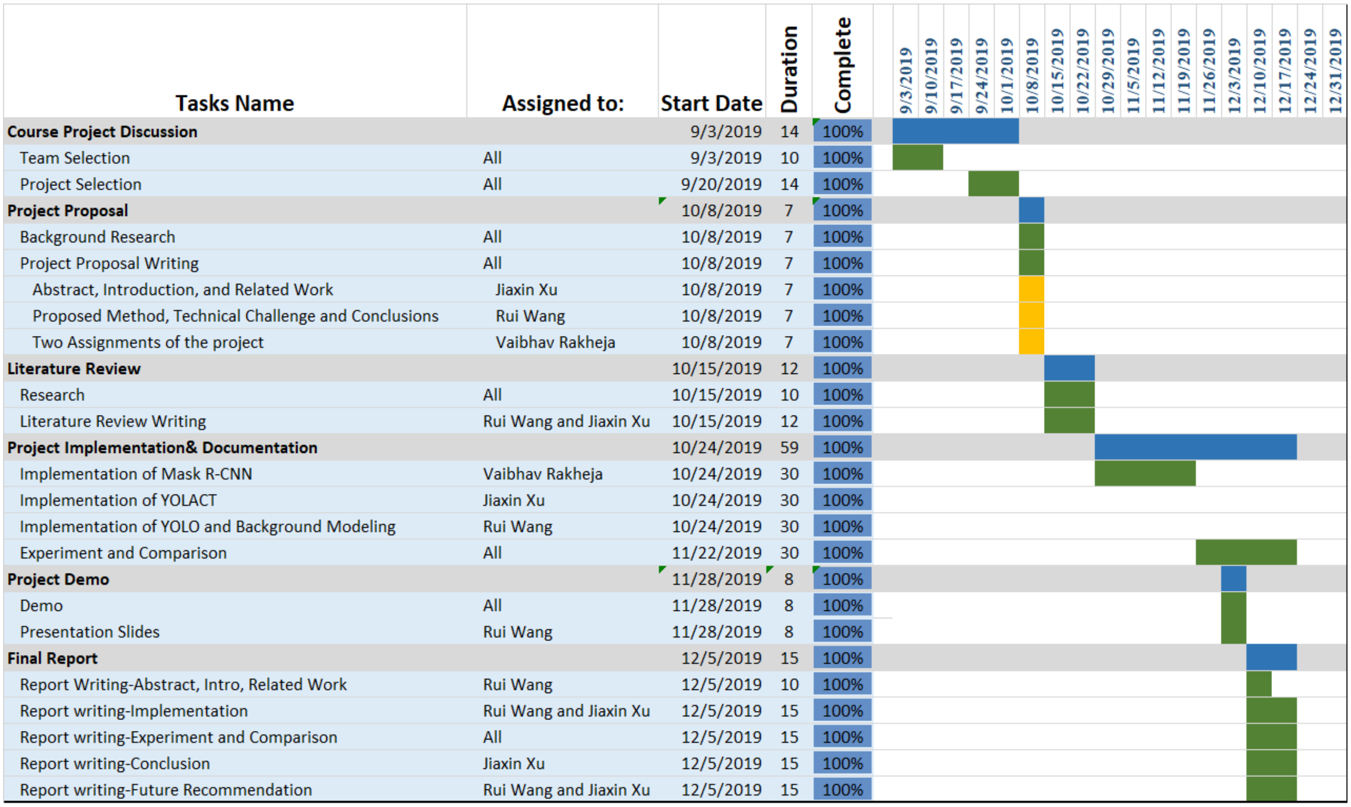


Fig. 4: Background Modeling Results



Tasks Name	Assigned to:	Start Date	Duration	Complete	9/3/2019	9/10/2019	9/17/2019	9/24/2019	10/1/2019	10/8/2019	10/15/2019	10/22/2019	10/29/2019	11/5/2019	11/12/2019	11/19/2019	11/26/2019	12/3/2019	12/10/2019	12/17/2019	12/24/2019	12/31/2019
Course Project Discussion		9/3/2019	14	100%	🟦	🟦	🟦	🟦	🟦													
Team Selection	All	9/3/2019	10	100%	🟩	🟩																
Project Selection	All	9/20/2019	14	100%				🟩	🟩													
Project Proposal		10/8/2019	7	100%						🟦												
Background Research	All	10/8/2019	7	100%						🟩												
Project Proposal Writing	All	10/8/2019	7	100%						🟩												
Abstract, Introduction, and Related Work	Jiaxin Xu	10/8/2019	7	100%						🟨												
Proposed Method, Technical Challenge and Conclusions	Rui Wang	10/8/2019	7	100%						🟨												
Two Assignments of the project	Vaibhav Rakheja	10/8/2019	7	100%						🟨												
Literature Review		10/15/2019	12	100%							🟦	🟦										
Research	All	10/15/2019	10	100%							🟩	🟩										
Literature Review Writing	Rui Wang and Jiaxin Xu	10/15/2019	12	100%							🟩	🟩										
Project Implementation& Documentation		10/24/2019	59	100%									🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦		
Implementation of Mask R-CNN	Vaibhav Rakheja	10/24/2019	30	100%									🟩	🟩	🟩	🟩						
Implementation of YOLACT	Jiaxin Xu	10/24/2019	30	100%																		
Implementation of YOLO and Background Modeling	Rui Wang	10/24/2019	30	100%																		
Experiment and Comparison	All	11/22/2019	30	100%													🟩	🟩	🟩	🟩		
Project Demo		11/28/2019	8	100%														🟦				
Demo	All	11/28/2019	8	100%														🟩				
Presentation Slides	Rui Wang	11/28/2019	8	100%														🟩				
Final Report		12/5/2019	15	100%															🟦	🟦		
Report Writing-Abstract, Intro, Related Work	Rui Wang	12/5/2019	10	100%															🟩			
Report writing-Implementation	Rui Wang and Jiaxin Xu	12/5/2019	15	100%															🟩	🟩		
Report writing-Experiment and Comparison	All	12/5/2019	15	100%															🟩	🟩		
Report writing-Conclusion	Jiaxin Xu	12/5/2019	15	100%															🟩	🟩		
Report writing-Future Recommendation	Rui Wang and Jiaxin Xu	12/5/2019	15	100%															🟩	🟩		

Fig. 5: Timelines and Individual Responsibility

tion. YOLACT can implement real-time human segmentation by adding an FCN branch on top of RestNet101 backbone. YOLACT successfully remaining the merits of these two frameworks, thus can produce high-quality masks as well as speeding up the segmentation process. We also tried combining YOLO and background subtraction. Although we managed to leverage the speed of YOLO, the accuracy of segmentation still need improvements.

6. FUTURE RECOMMENDATIONS

There are several problems we faced with during our implementation: 1. The outline of human lacks of accuracy. 2. The occlusion problem.

In terms of the above issues, we will modify our models in the following potential directions. We want to utilize human skeleton information to enhance the accuracy of human outline detection. Considering the three models we implemented were not human specialized, we failed to utilize the skeleton structure of the human category. By learning the human skeleton structure, models are more likely to recognize different parts of the human body, thus can deal with occlusion to a certain extent. For improvement in background modelling, based on our observation from this project, background modelling is not able to fully extract features of an image. In the future, we will apply background and foreground subtraction for initial background and foreground generation, then input into deep learning models, such as CNN, to obtain the final foreground mask.

7. TIMELINES AND INDIVIDUAL RESPONSIBILITY

Our project timelines and individual responsibility are included in Figure 4.

8. REFERENCES

[1] Alan L Yuille, Peter W Hallinan, and David S Cohen, “Feature extraction from faces using deformable templates,” vol. 8, no. 2, pp. 99–111,

[2] Eran Borenstein and Shimon Ullman, “Class-specific, top-down segmentation,” Springer.

[3] João Carreira, Fuxin Li, and Cristian Sminchisescu, “Object recognition by sequential figure-ground ranking,” vol. 98, no. 3, pp. 243–262,

[4] Pablo Arbeláez, Bharath Hariharan, Chunhui Gu, Saurabh Gupta, Lubomir Bourdev, and Jitendra Malik, “Semantic segmentation using regions and parts,”

[5] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee, "Yolact: Real-time instance segmentation,"

[6] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection,"

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn,"

[8] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation,"

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks,"

[10] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation,"

[11] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik, "Simultaneous detection and segmentation," Springer.

[12] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik, "Multiscale combinatorial grouping,"

[13] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother, "Instancecut: from edges to instances with multicut,"

[14] Jifeng Dai, Kaiming He, and Jian Sun, "Instance-aware semantic segmentation via multi-task network cascades,"

[15] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei, "Fully convolutional instance-aware semantic segmentation,"

[16] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection,"

[17] David G Lowe, "Distinctive image features from scale-invariant keypoints," vol. 60, no. 2, pp. 91–110,

[18] Meghna Singh, Anup Basu, and Mrinal Kr Mandal, "Human activity recognition based on silhouette directionality," vol. 18, no. 9, pp. 1280–1292,

[19] Meghna Singh, Mrinal Mandal, and Anup Basu, "Pose recognition using the radon transform," IEEE.

[20] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu, "Pose2seg: Detection free human instance segmentation,"

[21] Chris Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking. in proceedings," vol. 2, pp. 246–252

[22] Nil Goyette, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, and Prakash Ishwar, "Changedetection. net: A new change detection benchmark dataset," IEEE.

[23] Thierry Bouwmans, Fida El Baf, and Bertrand Vachon, "Background modeling using mixture of gaussians for foreground detection-a survey," vol. 1, no. 3, pp. 219–237,

[24] Hansung Kim, Ryuuki Sakamoto, Itaru Kitahara, Tomoji Toriyama, and Kiyoshi Kogure, "Robust foreground extraction technique using background subtraction with multiple thresholds," vol. 46, no. 9, pp. 097004,