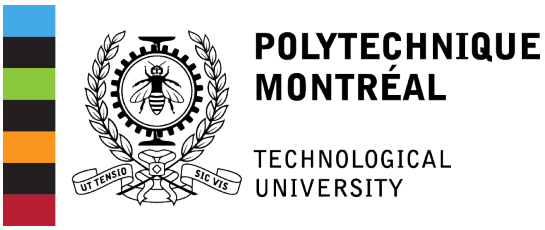

Polytechnique Montreal
Department of Computer Engineering

Lab 1 INF6804 - Winter 2020
Video Segmentation

Daniel Wang & Oumayma Messoussi

Supervisors:

David-Alexandre Beaupre
Soufiane Lamghari

February 2020

Table of Contents

1. Presentation of the two methods

In this first lab of the computer vision course, we had the opportunity to study and explore two techniques for extracting regions of interest from video sequences.

There are two categories of video foreground/background segmentation methods: optical flow algorithms and detection by classification algorithms. One method of each of these categories will be tested and presented in this report.

Below is a brief introduction to the selected methods:

1.1 Farneback Optical Flow

The first method we tested was optical flow. In a nutshell, optical flow is a representation of pixel motion between consecutive frames.

This approach is built on the following assumptions and approximations:

- The pixels belonging to a moving object have constant intensities across the two frames.
- We apply the Taylor series expansion and then divide by dt to get this equation:

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0, \text{ with } u = \frac{\partial x}{\partial t} \text{ and } v = \frac{\partial y}{\partial t}$$

This way, it boils down to solving u and v, but we have one equation and two unknown variables.

We also distinguish between two types of optical flow: Sparse optical flow which estimates the flow vectors for selected features in the images (e.g. corners), whereas dense optical flow returns estimation for the entire image (all the pixels), therefore it is more adequate for segmentation tasks.

For this lab, we will use the Farneback optical flow method [1] proposed by Gunnar Farneback. This methods starts by approximating a pixel's neighborhood using a quadratic polynomial. Next, observations of the polynomial transform under translation allow to derive a method to estimate the displacement. Finally, some refinements are applied to obtain the dense optical flow.

1.2 Mask R-CNN

Convolutional Neural Networks, commonly abbreviated as CNNs, constitute one of the modern deep learning architectures employed for the purpose of object classification. Empirical evidence shows that CNNs surpass humans in their ability to classify images on the ImageNet challenge [2]. In view of such promising results, a series of augmentations have been developed to extend the task domain in which Convolutional Neural Networks function to process images and video. To begin with, Regional Convolutional Neural Networks were created to perform the task of object detection or image segmentation. As the name suggests, R-CNNs generate sub-spaces delineated by a bounding box within an image, which the inventor denotes as region proposals, and then classify the content within the area demarcated. If an object is found in the region proposed, the bounding box is then refined so as to accurately capture the object's location. Previously, a selective search algorithm was used to generate region proposals, but such a method was found to be computationally time-consuming. To remedy this issue, Shaoqing et al. (2015) elected to insert into the detection network a region proposal network (RPN) [3], whose function to create bounding boxes around regions predicted to have high probability of containing an object. Mask-RCNN is the next iteration of object detection neural architectures [4], and is essentially an upgrade consisting of an implementation of three additional features, the first of which being a branch that conducts pixel-wise segmentation of objects in the foreground. In keeping with its name, Mask-RCNN overlays a coloured mask on every detected object in addition to bounding box encasements.

Secondly, He et al. (2017) have integrated feature pyramid networks (FPNs) as a backbone that has greatly ameliorated the accuracy of object detection models. The purpose it serves is to generate region of interest (ROI) features. Lastly, Mask-RCNNs replace the ROI pooling operation with a ROI align operation to alleviate the burden of needing to perform quantization for a number of times whose magnitude is greater than what may be computationally feasible for instance segmentation.

2. Performance hypotheses in specific use cases

2.1 Hypotheses for Baseline condition

It is expected that Optical flow should yield good performance since the baseline data-set provides favourable conditions, specifically of which are the presence of little to noise. However, Mask-RCNN is recognized as the subsequent iteration of Faster-RCNN, and performs better than its predecessor for the task of object detection and instance segmentation. It had produced state of the art results in 2017. For this reason, we hypothesize that Mask-RCNN will outperform Optical flow on the baseline dataset.

2.2 Hypothesis for Nighttime condition

Variations in brightness and color in images are regarded as undesirable noise that impair the function of computer vision algorithms. Deep learning methods hold a tendency to adhere too rigidly to the patterns or features in the training data, and hence easily become sensitive to noise. Similar to the blizzard weather condition, the hypothesis is that Mask-RCNN will not perform as well under the nighttime condition.

However, it is worth noting that certain video processing methods carry assumptions which alter the definition of what constitutes noise. With regard to optical flow in this case, there is a presumption that the brightness of the points being tracked is constant. Thus, a change in the lighting condition will not impact the performance of optical flow so long as the brightness across every pixel in the video remains invariant. The hypothesis to be drawn here is that Optical Flow will outperform Mask R-CNN when image brightness deviates from standard levels.

2.3 Hypothesis for Blizzard weather condition

Something that ought to be mentioned is that deep learning is not without its shortcomings. From a theoretical viewpoint, the foundational principles of machine learning on which Mask-RCNN is largely based brings to light some of the limitations suffered by modern deep learning methods. For instance, the generalization gap is a longstanding problem that has yet to be fully remedied given contemporary advancements in the field. To explain briefly, the generalization gap describes the discrepancy in model performance between executions on the data previously used for training and unobserved data. Such a concept provides a good reason explaining the occurrence of misclassification errors when testing the deep learning model on non-identical data whose qualities differ to an extent from that on which the model was trained, or on data polluted by noise. Thus, object occlusions effected by environmental conditions such as bad weather constitute noise that may drastically decrease reduce the accuracy of the deep learning model. As such, we predict that Mask-RCNN will perform worse under the blizzard weather condition.

As for the other method, Optical Flow only works effectively subject to certain constraints, namely being brightness constancy and minuteness in the degree of object displacement or motion. Flurries of snow in the video brought about by blizzards render the conditions impossible for optical flow to perform well, due to a violation of the brightness constancy condition effectuated by the presence of large fluctuations in the brightness of the video.

Determining whether Mask-RCNN or Optical Flow is the best method for this particular use case is a matter of comparing the level of noise sensitivity between the aforementioned methods. We conjecture that Mask-RCNN is less sensitive to noise formed by brightness variation in consideration of the fact that deep learning does not depend upon the condition of brightness constancy to the same extent as Optical Flow to function properly. Therefore, the hypothesis to be held is that Mask-RCNN will perform better in this case.

3. Description of experiments, datasets and evaluation criteria

For the purpose of this lab, we used the Change Detection 2014 dataset [5]. This dataset provides frames of video sequences spanning different conditions such as baseline videos in normal conditions, as well as difficult weather, camera jitter, low framerates, night views, etc. The ground truth segmentation masks are also provided which shall be used for evaluation.

We selected the following three conditions to test our two selected methods:

3.1 Baseline

This category contains sequences filmed in normal conditions to serve as a baseline for results comparison. We selected a pedestrian video sequence with 1099 frames.

3.2 Night videos

We selected a sequence of images taken at night time of cars on a highway. The sequence frames are captured under lower brightness settings.

3.3 Blizzard

The third and last sequence we used for testing the two methods belongs to the subset subject to the blizzard weather condition from the CD-Net data-set. This is another sequence where we have a view of cars on a highway whose environment is chiefly characterized by heavy snowfall.

3.4 Evaluation Criteria

The two principle criteria by which we evaluate performance include both a qualitative and quantitative metric. For the first measure, the qualitative metric concerns itself with a visual comparison of a few images between those which are segmented by optical flow and those which are segmented by Mask-RCNN. Secondly, we calculate the intersection over union (IoU) of the segmentations for all frames in the dataset to obtain a quantitative score given in the form of a ratio, and record the minimum, average, and maxiumum IoU.

4. Description of the implementations used

In this section, we describe the implementations of the two methods selected. To accomplish this, an exposition of the implementation details, frame pre-processing, and results generation shall be given.

Our tests were mainly performed using the popular computer vision library OpenCV for the optical flow method and TensorFlow for Mask R-CNN.

4.1 Farneback optical flow

Our implementation of the Farneback method was largely inspired by the official optical flow tutorial provided on the OpenCV website [6].

For each one of our three selected subsets, we iterate through all the frames with *glob.glog()*. For every frame, we execute OpenCV's implementation of the Farneback method. Then, we compute the direction and magnitude and visualize them respectively by hue and HSV color value.

For evaluation, we apply the following series of operations to both the output of optical flow and the correspoding ground truth segmentation (inspired from [7]): First, we apply *cv2.Canny()* to get edges in images and then pass its output to the *findContours()* method to get the contours of different objects. Afterwards, we generate the bounding rectangle which we both display on the image and also use to calculate IoU for every contour. Lastly, the average IoU is calculated and then the average IoU per subset is also shown for every frame.

4.2 Mask R-CNN

The code for Mask R-CNN was cloned from the Matterport software library on github (https://github.com/matterport/Mask_RCNN). Training of the neural network was done using the popular MS COCO data-set. The model was then executed on the CDnet dataset subject to the base line, night video, and blizzard conditions.

4.3 Evaluation metric: IoU

To evaluate optical flow and Mask R-CNN, we used PyImageSearch's implementation of intersection over union [8] to evaluate our models (with a slight modification). The defined function *bb_intersection_over_union()* takes as input the (x,y) coordinates of the top-left and bottom-right corners of the bounding boxes.

5. Experimentation results

For our experiments, we compare the average, min and max of the computed IoU scores for each condition. The scores are obtained from processing 300 images from every subset. With respect to the implementation of Mask R-CNN, we were unable to process the entire set of frames in each category due to limitations in computing power on Google Colab. For the bad weather condition, we selected image frames ranging from 003400 to 003699, and for the baseline frames we selected images from 000400 to 000699. A reduction in the number of frames processed from 300 to 100 was obligatory, as unfortunately, constraints on the output size on Google Colab prevented us from processing more than one hundred images.

5.1 Baseline results

Table 5.1: IoU results for baseline subset

Method\Score	Average IoU	Min IoU	Max IoU
Farneback optical flow	0.616913	0.0	0.885826
Mask R-CNN	0.848363627386053	0.0	0.9805825242718447

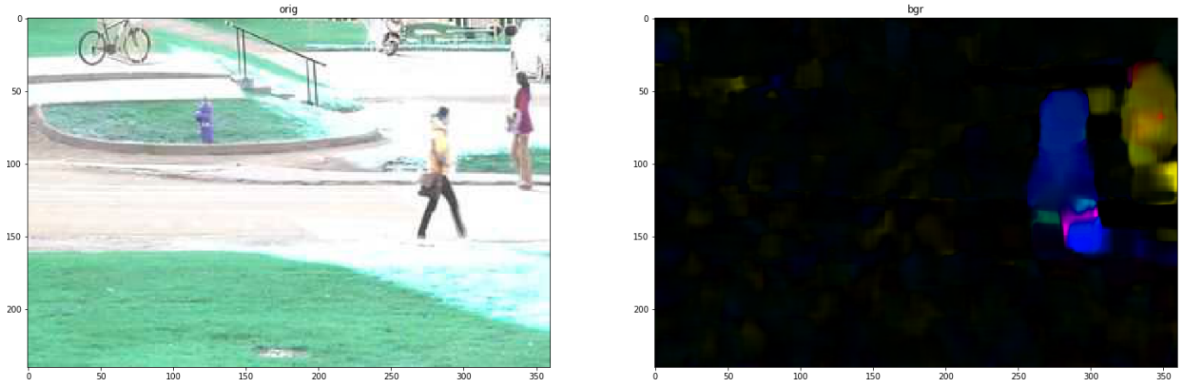

Figure 5.1: Optical flow result for baseline image in000418

Figure 5.2: Mask R-CNN result for baseline image in000418

5.2 Bad weather results

Table 5.2: IoU results for bad weather subset

Method\Score	Average IoU	Min IoU	Max IoU
Farneback optical flow	0.696491	0.0	0.969397
Mask R-CNN	0.7923620393284657	0.0	1.0

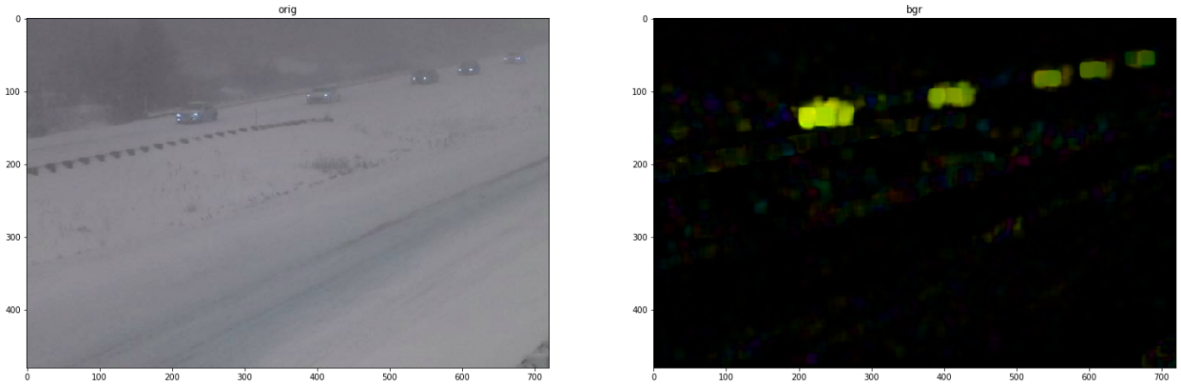

Figure 5.3: Optical flow result for bad weather image in001024

Figure 5.4: Mask R-CNN result for bad weather image in001024

5.3 Night videos results

Table 5.3: IoU results for night videos subset

Method\Score	Average IoU	Min IoU	Max IoU
Farneback optical flow	0.141303	0.0	0.751795
Mask R-CNN	0.49261198815819807	0.0	0.930952380952381

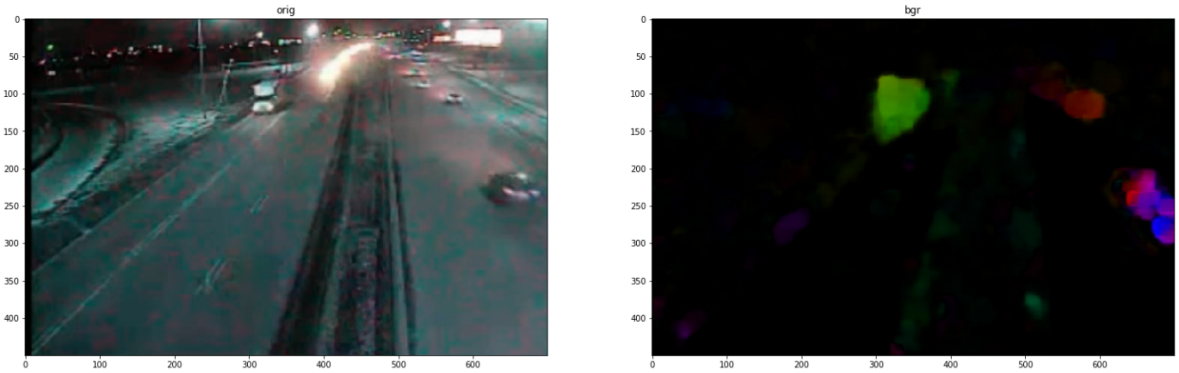

Figure 5.5: Optical flow result for night video image in000011

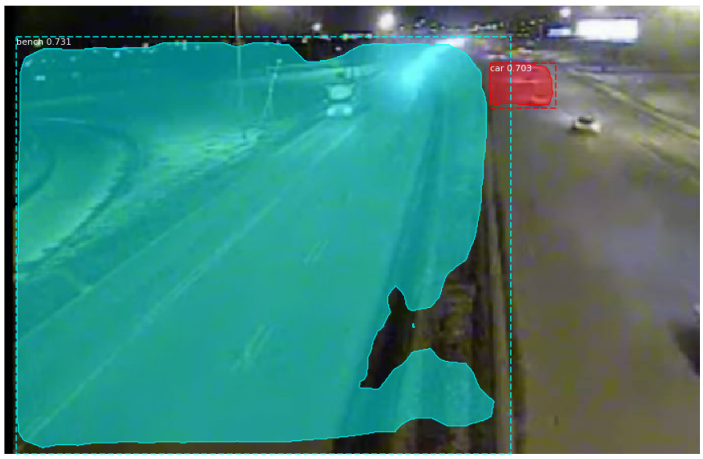

Figure 5.6: Mask R-CNN result for night video image in000011

6. Discussion on results and prior hypotheses

6.1 Baseline results

As expected, Mask-RCNN demonstrated near impeccable segmentation performance on the baseline image frames, affirming the hypothesis proposed in regard to performance under normal conditions. Contrary to our theoretical supposition that optical flow would perform relatively well, our tests showed that this method failed to correctly detect objects in motion. The average IoU is rather low which might be explained by a slight lacking in brightness constancy across frames.

6.2 Bad weather results

Mask-RCNN was able to detect and segment some of the cars in the foreground, but nevertheless there were instances in which Mask-RCNN failed to detect the majority of the cars displayed. This lends credence to our initial hypothesis which postulated that noise in the image would negatively affect performance of our neural network. Optical flow's fairly good performance in this condition may be explained by the fact that, in spite of a failure to meet the brightness constancy requirement brought to effect by noise in the form of heavy snowfall, the frames are filmed in a low light cloudy day with a relatively low resolution such that any existing fluctuations are smoothed. Therefore, we get approximately an IoU ratio of 67% between the prediction and ground truth bounding boxes. Ultimately, the visual results demonstrate that optical flow is better at discerning foreground objects in videos containing snowfall, but from a quantitative viewpoint, Optical Flow is not as accurate as Mask-RCNN at object detection by a margin of 10 percent in light of the difference in IoU ratio. This leads to an indefinite conclusion pertaining to the matter of determining the superior method. Mask-RCNN yields greater accuracy in the overlay of segmentation masks, yet suffers from an inability to consistently detect multiple objects. For the other method, optical flow succeeds in detecting the totality of objects located in the foreground, but the quality of the segmentations still possesses room for improvement with regard to precision.

6.3 Night videos results

Upon altering the brightness of the video to settings, in some cases Mask-RCNN not only failed to detect the objects in the foreground, but also erroneously placed segmentation masks over areas where foreground objects are non-existent. However, Mask-RCNN was able to detect and superimpose a segmentation mask over one or two cars with fairly good accuracy in a number of instances. With regard to the other method, the results produced by optical flow were not particularly impressive. Optical flow was able to locate the position of some of the foreground objects but failed to accurately capture the shape of the car in the foreground. After completing our empirical investigation, the conclusion to be drawn here is that brightness levels on the lower end of the spectrum render video conditions unfavourable for segmentation through the utilisation of Mask-RCNN. It seems that optical flow performed somewhat better than Mask-RCNN at foreground background segmentation, but lacks in accuracy and precision, as evidenced by an abysmal average IoU ratio. In conclusion, the evidence generated by our experiment with the introduction of a the nighttime environmental variable affirms to some extent our proposed hypothesis in respect of nighttime video settings.

6.4 Suggestions for Test Improvement

The contrast in brightness between the baseline condition and the nighttime condition may be a bit too extreme for optimal experimentation. Ideally, the tests should be designed so that, through observation, one may determine the threshold of intensity over which performance begins

to worsen. Hence, it would be prudent to test for a wide array of intensity levels, and perhaps also for internal variations in brightness.

Bibliography

[1] Farnebck G. Two-Frame Motion Estimation Based on Polynomial Expansion. In: Bigun J, Gustavsson T, editors. Image Analysis. Springer Berlin Heidelberg; 2003. pp. 363370. . 1

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[5] CDNet 2014 dataset. 4

[6] OpenCV Optical Flow Tutorial. 5

[7] Creating Bounding boxes and circles for contours. 5

[8] Intersection over union. 5