Linköping University | Department of Computer and Information Science
Project, 1.5 ECTS | TSBB17 – Visual Object Recognition and Detection
Autumn Term 2020

Project 2

– Visual Object Tracking

Lukas Borggren
Viktor Gustafsson
Gustav Wahlquist

Supervisor: Gustav Häger
Examiners: Per-Erik Forssén and Michael Felsberg

Linköping University
SE-581- 83 Linköping
013-28 10 00, www.liu.se

Introduction

This project aims to evaluate the MOSSE tracker and how it performs with different maps of features on the OTB-mini dataset. The source code can be found at https://gitlab.liu.se/lukbo262/tsbb17-project2.

Method

Initially, the supplied NCC tracker is evaluated on the dataset, using greyscale representation, to be used as a baseline comparison. Thereafter, the MOSSE tracker is evaluated on greyscale in order to compare it directly to the baseline. Subsequently, a multi-channel version of the MOSSE tracker is evaluated employing a single feature type at a time, namely RGB representation, HOG descriptors, gradient filter banks and deep features. Each feature is additionally evaluated with a scale estimator component. In order to further assess the tracker's performance, a selection of pairs of feature types are evaluated in conjunction.

Naturally, it would be favorable to test all possible combinations of features to obtain the best performing tracker. During the evaluation, it is mainly the deep features that are not evaluated in conjunction with the other features. This is due to the long run time when using these features and the limited timespan of the project.

Dataset and Parameters

The dataset used in this project is a subset of the OTB benchmark called OTB-mini. It consists of 30 image sequences together with a ground truth bounding box for each frame. The sequences include a varying number of frames, but the size of the frames are consistent across the whole dataset.

During evaluation, the learning rate and regularization term are both fixed to 0.01 for the MOSSE tracker. For the two-dimensional Gaussian function, $\mu = 0$ and $\sigma = 2$ is used. These parameters were varied and tested for a few iterations on the MOSSE tracker with different feature configurations before settling on the final values.

Evaluation

The different trackers are evaluated using average intersection over union (IoU). It is computed by first averaging IoU over the frames of each sequence individually and then averaging the results over all sequences. If a bounding box is completely outside a frame, the IoU's for all subsequent frames are set to 0. This measure does not explicitly indicate for how long the tracker is able to track objects reasonably accurate. Therefore, an improvement could have been to use accuracy and robustness as evaluation criterions instead, i.e. to have separate measurements for how accurate the tracking is and how well it sticks to the target over time. However, we deem average IoU to be sufficient for the purpose of this project. Additionally, the run time for the trackers are measured to indicate how efficient they are. The run time is measured as the time to complete the evaluation of all sequences.

Results

NCC and Basic MOSSE

Below are the evaluation results for the NCC tracker, as well as for the MOSSE tracker when using greyscale and RGB.

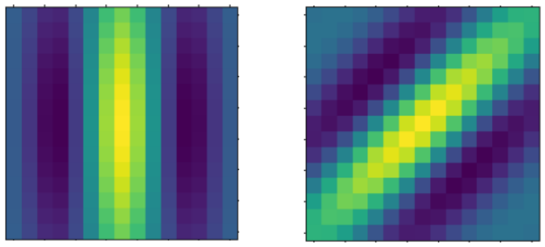
Tracker	Average IoU	Run time (min)
NCC greyscale	0.0437	1.44
MOSSE greyscale	0.2212	1.47
MOSSE greyscale with scale estimation	0.1464	2.19
MOSSE RGB	0.2211	1.49
MOSSE RGB with scale estimation	0.2158	4.38

When comparing the average IoU for NCC and MOSSE greyscale it is evident that MOSSE's performance is superior. It is noteworthy that the performance of MOSSE is almost identical when using greyscale and RGB. The explanation for this could be that both features in some sense comprises the same information about an image, albeit in a more compressed form for greyscale. Furthermore, it seems that the scale estimation does not enhance the tracker's performance in this case. On the contrary, IoU is lower for both greyscale and RGB when scale estimation is used. Henceforth, the tracker being evaluated is MOSSE if not stated otherwise.

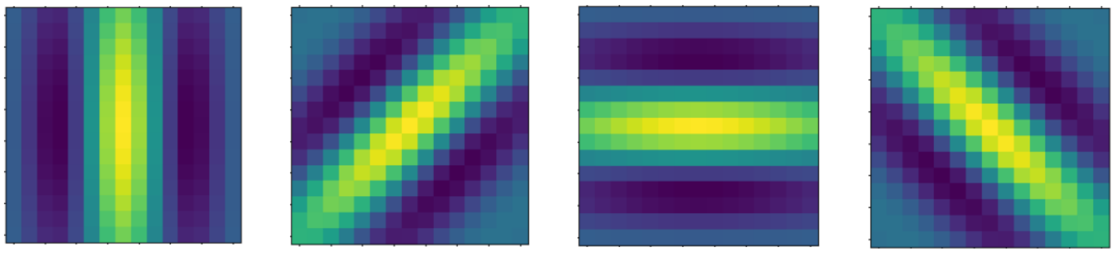
Gradient Filter Banks

In the experiment, three different variations of gradient filter banks where tested. The difference between these variations was both different number of gradient filters and different directions of the filters. Following are the filters for each variation is plotted along with a table with the evaluation results.

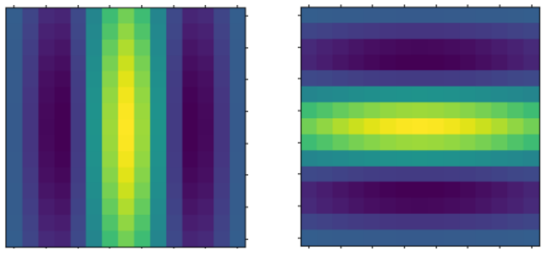
V1:



V2:



V3:



Filter Bank Variation	Average IoU	Run time (min)
V1	0.2394	3.93
V2	0.2936	8.63
V3	0.296	4.36
V3 with scale estimation	0.2245	9.62

From the result above we can see that V3 is the best performing variation, having both a slightly higher average IoU and being twice as fast than as the second-best V2. The fact that V3 performs better than V2 while having fewer filters indicates that the filters that are in V2 but not in V3 do not add more useful information for the tracker. Again, scale estimation decreases the performance and increases the time efficiency, which is not preferable.

HOG Descriptors

To later enable combinations of HOG descriptors and other features, it is desirable if these follow the same data-structure conventions, namely being represented in stacks of two-dimensional NumPy arrays. Therefore, the HOG descriptors are chosen to have (1,1) cells per block, as this implies that the number of blocks will equal the number of cells. Consequently, the dimensions are allowed to be squeezed out to a different shape, since the corresponding block shape is 1.

```
hog_hist = hog_hist[:, :, 0, 0, :]
```

The above code snippet results in a HOG vector representation: [cell_x, cell_y, orientation]. Accordingly, each orientation will be represented as a separate channel in the tracker. These features are then upscaled, in accordance with the below code, so that the cell_x- and cell_y-dimensions have the same shape as the original images. This entails that the HOG descriptors can be combined with other feature types when subsequently evaluating MOSSE with pairs of features.

```
upsample_fd = cv2.resize(hog_hist, (image.shape[1], image.shape[0]))
```

Below are the evaluation results for some HOG descriptors with different parameter configurations. The features reported here are normalized, as this gave better performance than without normalization. Some configurations perform better than when using greyscale or RGB, but most of them are inferior to the gradient filter bank features.

Parameter Configuration	Average IoU	Run time (min)
4 direction bins 40x40 cells	0.1966	12.15
8 direction bins 40x40 cells	0.2311	12.03
16 direction bins 40x40 cells	0.2482	15.45
16 direction bins 20x20 cells	0.1630	12.67
16 direction bins 40x40 cells With scale estimation	0.2176	25.31

Deep Features

The implementation of deep features utilized in this project up-samples the outputs from the network layers to the same size as the images, so that the bounding box will be placed in the right indexes of the deep features.

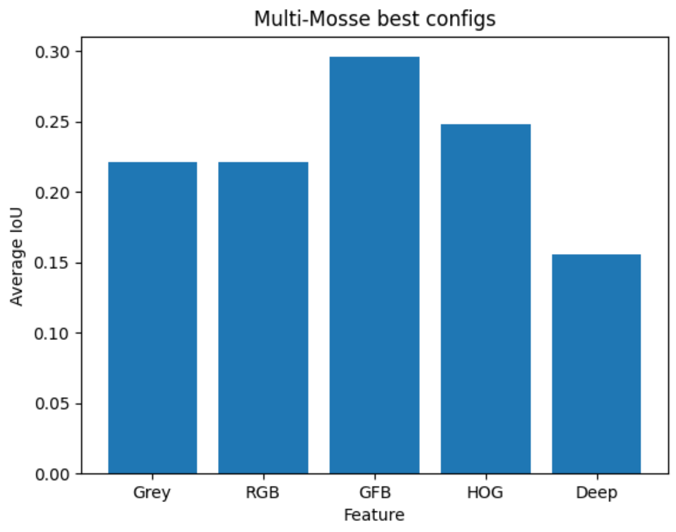
A few variations of deep features are tested, retrieved from the outputs of different layers in the network. The forward function in AlexNetFeatures class is adjusted to include a "layers"-parameter, which defines from which layers the features should be retrieved from. For the purpose of simplicity, one layer is considered to consist of one convolution, one activation and one pooling (in some case non-present).

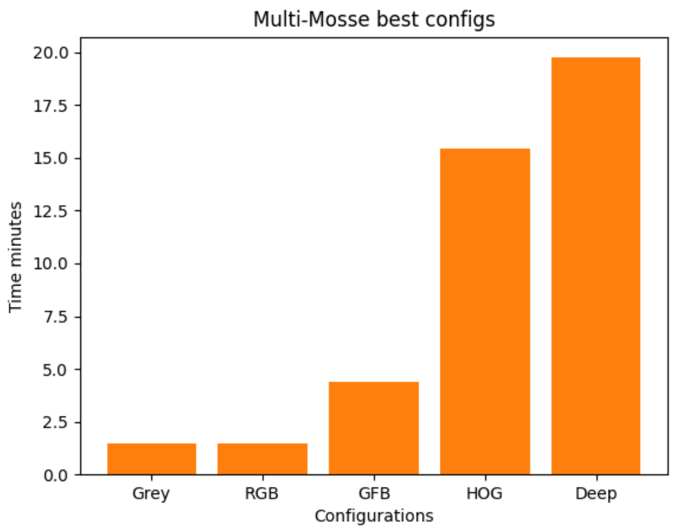
Network Layer	Average IoU	Run time (min)
1	0.1560	19.75
2	0.1252	54.78

A decision was made to not continue investigating the deep features past this point, neither by trying more layers nor by trying scale estimation. The reason being that tracking using the deep features takes a significantly larger amount of time because of the many channels used, and that the results from the first two layers are not particularly promising.

Comparison of Single Features

Below is a visual comparison of the tracking performance for the best parameter configuration, for each individual feature.

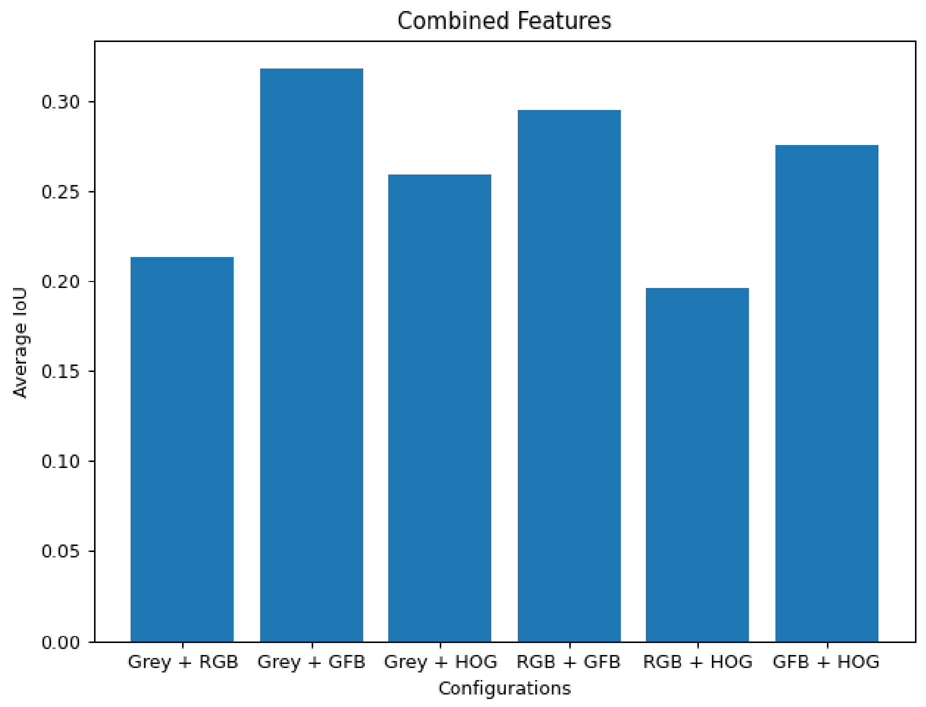




Gradient filter bank clearly has the best performance, although one could argue that greyscale or RGB might be preferred in some cases, if the performance loss may be compensated by a faster tracker.

Pairwise Combination of Features

Testing to combine pairs of features, the results below are obtained. For each feature type, the best parameter configuration from the earlier, individual evaluation is used.



Compared to the previous evaluation, when the features were tested individually, the result does not differ significantly. Generally, the average IoU is again between 0.2 and 0.3 making it difficult to discern exactly what effect the different combinations have on the performance. However, it may be noted that the addition of gradient filter bank boosts the performance for all features that it is combined with. The best IoU is obtained from the tracker using greyscale and gradient filter banks, with an average IoU of 0.3180 and a run time of 4.69 min.

Discussion

When tracking sequences with the ground truth bounding box close to the edge of the frame, the tracker runs a bigger risk of ending up outside of the image. This results in that the IoU for the frames left in the sequence is set to 0. This means that trackers' performances might be sensitive to sequences where the tracking target approaches the edge of the image. One way to counteract this would have been to re-insert the bounding box into the frame when the bounding box begins to drift outside of the image. Another shortcoming of the evaluation process used is that the results might be overly dependent on the initial ground-truth bounding box for each sequence. As the first frame is used to initialize the tracker and MOSSE filter, in combination with a relatively small learning rate, the bounding box has a large impact on the response for many subsequent frames. To counteract this, multiple evaluations could have been made, initializing the tracker on multiple frames and averaging the obtained scores.

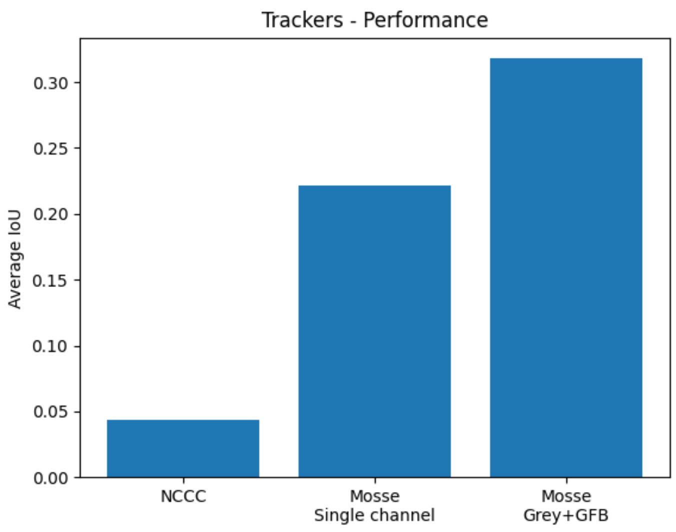
Parameters such as learning rate, μ and σ was, as mentioned in the method section, fixed throughout the experiments. The choice for these values were not based on empirical evidence. It would have been interesting to try different values for these parameters in combination with different features to see if optimizing the parameters would lead to large performance improvements.

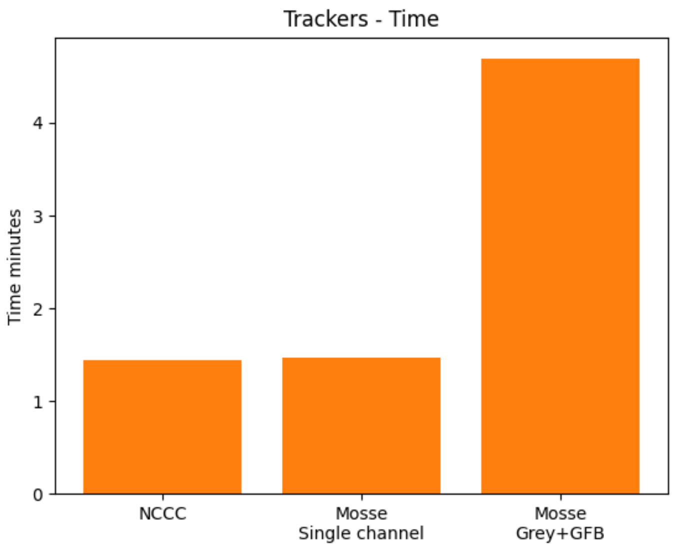
The performance of trackers using a scale estimation component are surprisingly worse in general. Anecdotally, there were observed cases where the tracker resized to unreasonable scales when the target was lost, making it seemingly difficult for the tracker to re-track the target. This might be one reason for the worse performance. Another, simpler explanation, might be that the implementation of the scale estimation is not entirely correct. Yet another reason for the worsened performance might be that the implementation relies on the assumption that the scale of an object in a sequence changes much slower than its position in the frame. This implies that the position of an object is first determined using the bounding-box size from the previous frame and thereafter the different scales are assessed, instead of the other way around. However, if the above-mentioned assumption holds true for the sequences in the OTB-mini dataset is hard to determine.

Contrary to initial expectations, the performance of trackers using deep features was relatively modest and worse than many of the other features tested. Again, this might be due to a sub-optimal implementation. For instance, the images were not resized before being propagated through the network which might result in bad performance. In retrospect it would have been interesting to resize the images to the same size as the images used to train AlexNet to see if how it affects the tracking performance.

Conclusions

Plotted below are the best configurations of the different trackers, NCC and their corresponding performance and run time.


Trackers - Performance
Average IoU
NCCC
Mosse
Single channel
Mosse
Grey+GFB


Trackers - Time
Time minutes
NCCC
Mosse
Single channel
Mosse
Grey+GFB