Flow R-CNN: Flow-enhanced object detection

Athanasios Psaltis, Anastasios Dimou, Federico Alvarez and Petros Daras

Outline

- Introduction
- Motivation
- Proposed method
- Experimental evaluation
- Conclusions
- Future work

Introduction

- Address the problem of multi-task object detection:
  - Fundamental task for the human visual system
  - Human brain uses multiple object properties to achieve the required recognition performance
    - object shape, structure, color and texture
  - Only appearance-, shape-related features have been employed until now in multi-target learning methods

Motivation

- Vast majority of objects are not stationary
- The motion characteristics of an object treated as a signature
- Exploiting the motion characteristics of an object can improve our object recognition capabilities
- Involves a number of strongly interconnected modalities
  - The shape of an object has been shown to be correlated with its motion characteristics
- Predicting the flow of an object from a single frame

Proposed approach: main contributions

- A neuroscience-inspired scheme to improve object detection by introducing an additional pseudo-temporal stream (branch) for motion prediction from still images

- An object-level flow field is incorporated in the object recognition process
  - by penalizing the global loss computation with an optical flow loss factor
  - motion prediction at RoI level

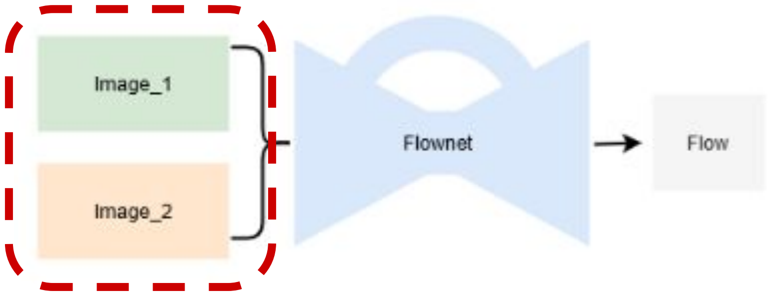
Background: Neuroscientific inspired methodology

- "Human brain predicts the path of a moving object (visual motion), to adapt human behavior to surrounding objects moving in real-time"

- Given a single static image:
  - the brain's ventral stream (what) interprets the instantaneous semantic content,
  - and at the sametime the dorsal stream (where) predicts what is going to happen based on scene spatial configuration

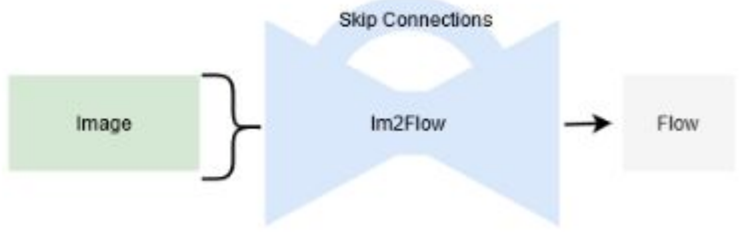
Background: Object-based motion analysis

- Information included in a pair of successive images is first spatially compressed in a contractive part of the CNN and then refined in an expanding part
- An encoder-decoder CNN equipped with a novel optical flow encoding scheme that is able to translate a single static image into an accurate flow field



Require a pair of images as input

(a) FlowNet

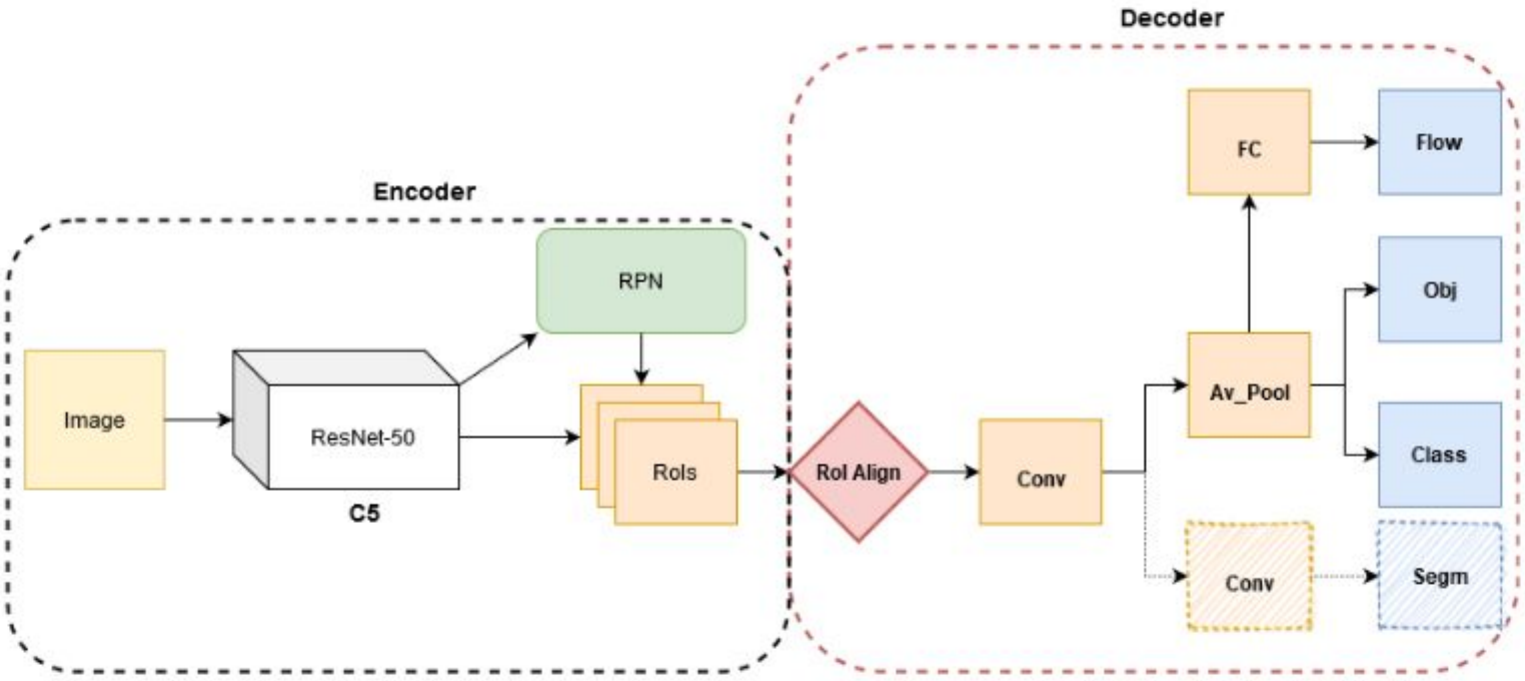


(b) Im2Flow

Background: Mask R-CNN

- Region-based: Mask R-CNN as baseline
  - an RPN mechanism in the first stage in order to propose candidate RoIs
  - locates the relevant areas of the feature map by utilizing a RoI-Align layer
  - The extracted features are further processed in parallel to perform classification, bounding box regression and instance-level semantic segmentation

Overall Flow R-CNN architecture



A composite region-based object detection model, the backbone of the network is used for the image encoding, while the object-level flow estimation branch is used to infer the optical flow field

Proposed architecture (Flow R-CNN)

- The proposed approach mimics the visual perception procedures that take place in the human brain, following an appropriate deep neuro-physiologically grounded architecture

- Flow R-CNN exhibits the following advantageous characteristics:
  - it enhances the two-stage detector by introducing an additional pseudo-temporal stream, and
  - it incorporates the aforementioned stream in a multi-task learning process

Experimental evaluation

For the evaluation, the following datasets were used:

- `KITTI`,
- `V-KITTI',
- `Visdrone'
- `Cityscapes'
- `Berkeley Deep Drive'
- `UDacity'

Experimental results

Incorporating the flow stream into the learning process of an R-CNN architecture may have a positive impacting in the detection and recognition of moving objects

Table 1: Comparative results on KITTI dataset

	Easy Mask	Flow	Moderate Mask	Flow	Hard Mask	Flow
Car	0.893	0.905	0.843	0.849	0.733	0.736
Pedestrian	0.804	0.812	0.672	0.677	0.619	0.622
Cyclist	0.739	0.746	0.635	0.638	0.554	0.556
mAP	0.812	0.821	0.717	0.721	0.635	0.638

Table 2: Comparative results on V-KITTI dataset

Class	Mask R-CNN	Flow R-CNN
Car	0.932	0.958
Van	0.917	0.940
mAP	0.924	0.949

Table 3: Comparative results on Visdrone dataset

Class	Mask R-CNN	Flow R-CNN
Pedestrian	0.205	0.223
People	0.071	0.064
Bicycle	0.029	0.033
Car	0.406	0.428
Van	0.208	0.232
Truck	0.148	0.181
Tricycle	0.132	0.148
Awn	0.091	0.085
Bus	0.216	0.253
Motor	0.153	0.151
mAP	0.166	0.180

Table 4: Comparative results on Cityscapes dataset

Class	Mask R-CNN	Flow R-CNN
Person	0.345	0.364
Rider	0.271	0.307
Car	0.488	0.505
Truck	0.296	0.306
Bus	0.401	0.387
Train	0.302	0.252
Motorcycle	0.237	0.256
Bicycle	0.182	0.204
mAP	0.315	0.323

Experimental results

Achieves improved performance in all datasets using deeper ResNet architectures

Table 5: Comparative results on BDD dataset

Class	Mask R-CNN	Flow R-CNN
Bike	0.383	0.391
Bus	0.481	0.489
Car	0.732	0.746
Motor	0.194	0.198
Person	0.531	0.537
Rider	0.349	0.352
Traffic-light	0.479	0.473
Traffic-sign	0.558	0.547
Truck	0.506	0.514
mAP	0.421	0.424

Table 6: Comparative results on Udacity dataset

Class	Mask R-CNN	Flow R-CNN
Bike	0.625	0.629
Bus	0.949	0.951
Car	0.724	0.736
Motorbike	0.738	0.736
Person	0.747	0.752
Traffic-light	0.502	0.498
Traffic-sign	0.701	0.696
mAP	0.712	0.714

Table 7: Comparative results on six datasets using different backbone architectures

Backbone	KITTI	V-KITTI	Visdrone	Cityscapes	BDD	Udacity
ResNet-50	0.724	0.949	0.180	0.323	0.424	0.714
ResNet-101	0.731	0.956	0.185	0.329	0.430	0.720
ResNet-50-FPN	0.735	0.961	0.194	0.334	0.432	0.725
ResNet-101-FPN	0.742	0.967	0.207	0.340	0.438	0.731

Experimental results

Mask R-CNN

Flow R-CNN

Conclusions

- A methodology for incorporation of pseudo-temporal information in Region-based CNN object detection schemes
- Pseudo-temporal stream was effectively incorporated into the learning process
- Experimentally shown to achieve improved performance in the six currently broadest and most challenging publicly available semantic urban scene understanding datasets

Future Work

- Investigation of re-adjusting the proposed pseudo-temporal branch utilizing a more sophisticated optical flow estimation methodology.

Thank You!