Using Optical Flows and a CNN for Visual Object Tracking

Vivek Roy [1], David Russell [2], Satyaki Chakrobarti [3], Martial Hebert [3]
[1] Jadavpur University, [2] Clarkson University, [3] Carnegie Mellon University

Visual Object Tracking

Is: Predicting an object's location in the next video frame, given the current location

Has: Applications in robotics, surveillance, and autonomous driving

Motivation

CNNs are widely used in object tracking but current approaches rely on mostly implicit feature matching.

We provide optical flow as input so the network learns from explicit geometric information about the scene.

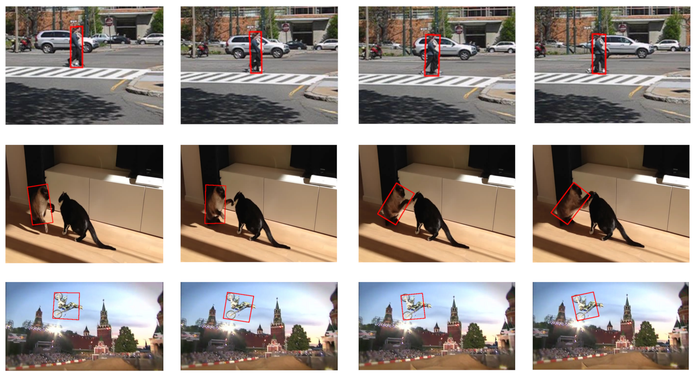

Fig. 1. Consecutive frames from the VOT challenge dataset. The goal is to predict the red bounding boxes, given only its position in the first frame.

Optical Flow

Estimation of object motion betwen two images

We compute flow between consecutive video frames

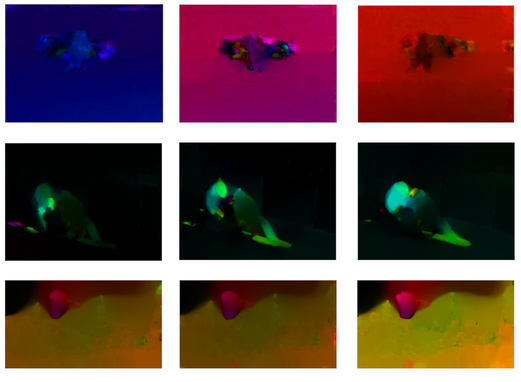

Fig. 2. Optical flows between the frames shown in Fig. 1.

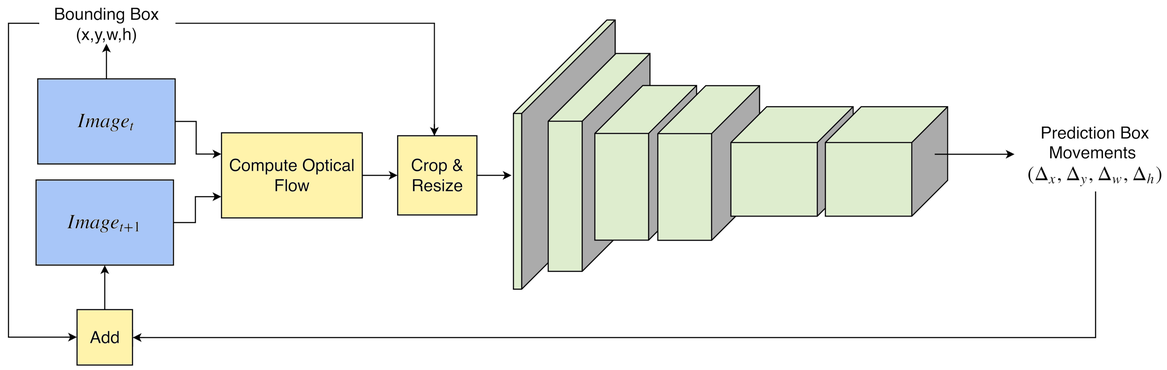

Fig. 3. System Architecture. Given the current bounding box for the object, we seek to predict a regression identifying the bounding box in the next frame.

Approach

Compute optical flow between current and next image

Crop flow to the current object location and resize

Predict the bounding box change with a CNN

Our network is a variant of VGG11

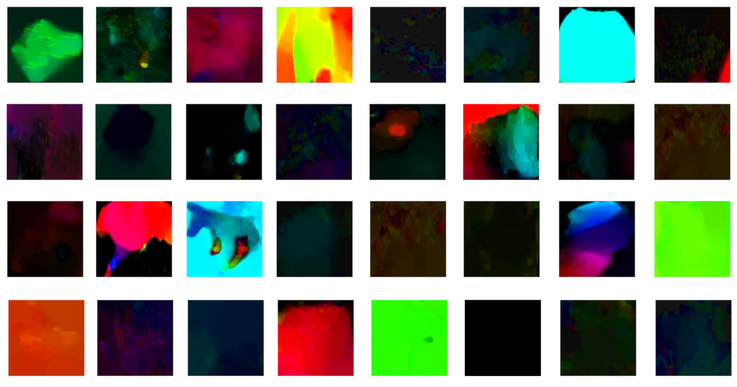

Fig. 4. The network learns to predict bounding box regressions from optical flow. The flow is cropped based on the ground truth in frame t, and the loss is evaluated with respect the displacement between the bounding boxes in t and t+1.

Training

Trained using image pairs from ImageNetVID

Ground truth crops and ground truth deltas

SmoothL1 loss function

Adam Optimizer, ReLU activation

Learning rate divided by 10 if validaton loss is greater than the loss from previous epoch

Conclusions & Future Directions

The training deltas are predominantly small values and thus the network is biased towards predicting small deltas.

We plan to conduct evaluations on the VOT2017 Dataset.

We would like to integrate long term motion models.

Additionally, we would like to train on synthetic data which models challenging situations.

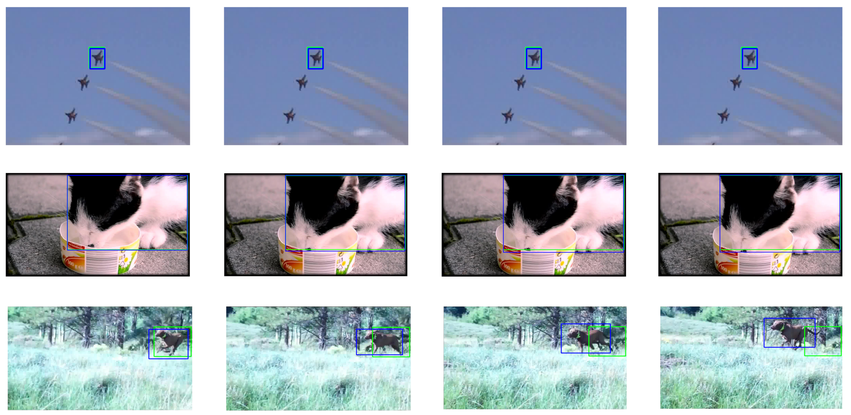

Fig. 5. Results on the validation set of ImageNetVID. Blue boxes represent the ground truth and the green boxes represent the predictions from the proposed approach.

Acknowledgments

We would like to thank the the National Science Foundation (NSF) and the Federation of Indian Chambers of Commerce & Industry (FICCI) for their financial support of this project.

Thank you Dr. Martial Hebert for this opportunity and Satyaki Chakrobarti for your guidance.

We are also incredibly grateful to Rachel Burcin, Dr. John Dolan, and Ziqi Guo for their tireless contribution to the Robotics Institute Summer Scholars program.

THE ROBOTICS INSTITUTE · Jadavpur University · Clarkson UNIVERSITY · NSF · FICCI