OBJECT DETECTION IN A VIDEO

Group Code
PM03

Mentor
Dr. Prerana Mukherjee

Group Members
Chinthala Sai Pranay Raju(S20170010036)
S B Koushik(S20170010131)
D Sai Karthik(S20170010037)

Introduction

- Object detection is one of the fundamental problems in Computer Vision.
- There has been a long history of detecting objects in static images, but now we see people shifting their interest into videos.

- However, cameras on robots, surveillance systems, vehicles, wearable devices, etc., receive videos instead of static images.
- Thus, for these systems to recognize the key objects and their interactions, it is critical that they be equipped with accurate video object detectors.

Problem Statement

- Video perception is an important aspect of every autonomous machine which uses cameras to perceive environment.
- But, due to the different biases and challenges of video (e.g., motion blur, low-resolution, compression ,etc..), a static object detector on video frames doesn’t work well.
- Videos also provide rich temporal and motion information which should be utilized by the detector.
- Also there might be dependencies between frames of videos, which play a crucial role and must be taken into account.
- Therefore by aggregating information across time and taking challenges into account we would design a good video object detector.

Literature

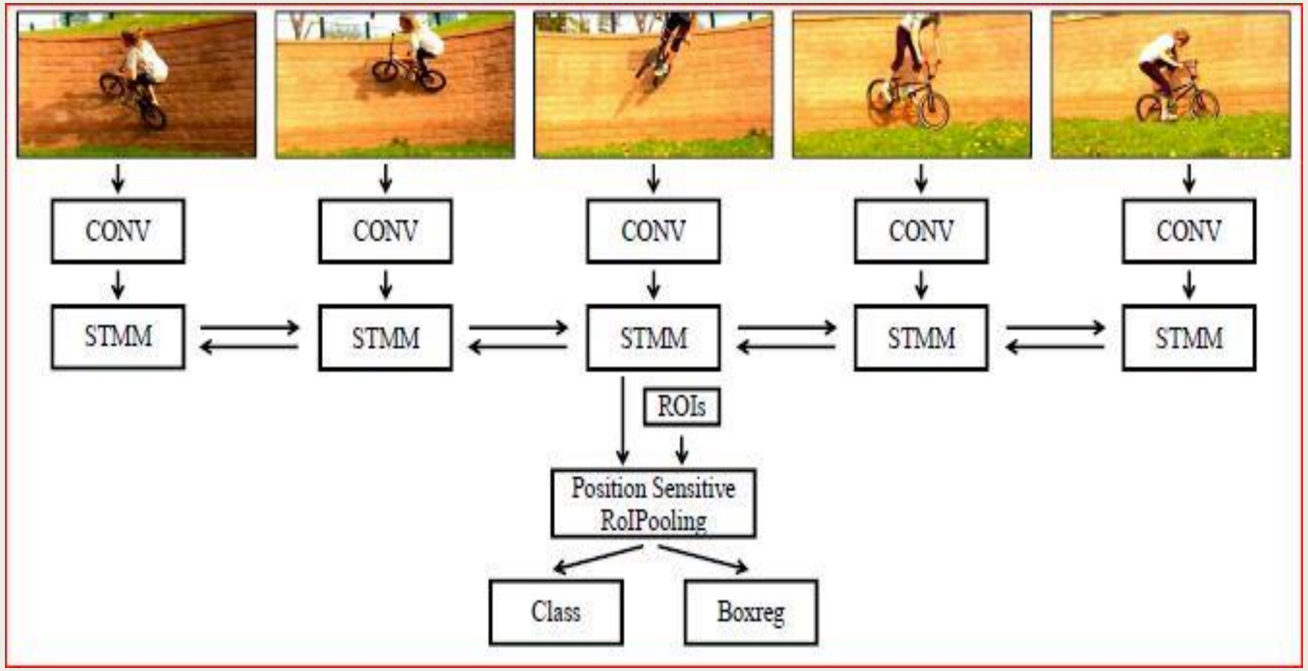
- The recent works of Zhu learn to combine features of different frames with a feed-forward network for improved detection accuracy. Our method differs in that it produces a spatial-temporal memory that can carry on information across long and variable number of frames
- whereas the methods in [4,5] can only aggregate information over a small and fixed number of frames.
- Although the approach of Kang et al. [3] uses memory to aggregate temporal information, it uses a vector representation. Since spatial information is lost, it computes a separate memory vector for each region tube (sequence of proposals) which can make the approach very slow. In contrast, our approach only needs to compute a single frame-level spatial memory, whose computation is independent of the number of proposals.

Objectives

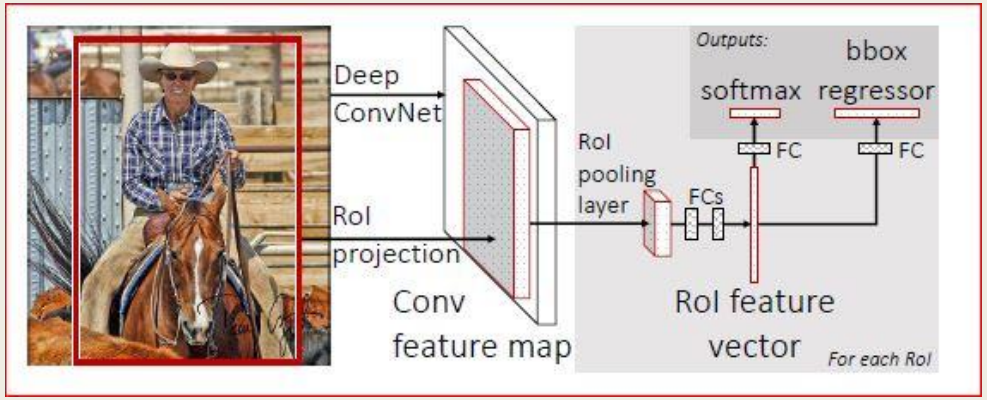
- To design a model which could detect and classify an object using temporal and spatial information.
- To design a model to maintain the alignment of memory along different frames preventing hallucinations.

Methodology [1]

- We use a newly defined memory module called STMM to transfer information among different frames. Here we use ConvGRU's.
- Architecture



RFCN

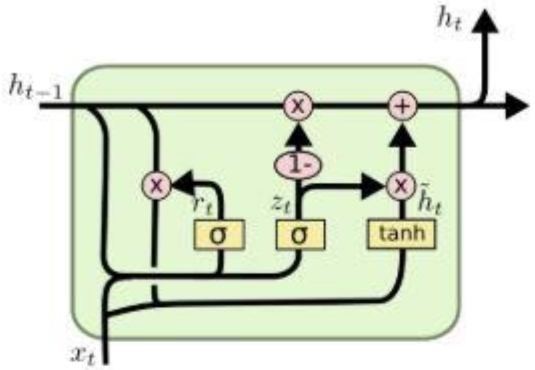


Methodology(Contd..)

- The STMM gets updated using the formulae mentioned below (Similar to GRU)

$$
\begin{aligned}
z_t &= \mathrm{BN}^*(\mathrm{ReLU}(W_z * F_t + U_z * M_{t-1})),\\
r_t &= \mathrm{BN}^*(\mathrm{ReLU}(W_r * F_t + U_r * M_{t-1})),\\
\tilde{M}_t &= \mathrm{ReLU}(W * F_t + U * (M_{t-1} \odot r_t)),\\
M_t &= (1 - z_t) \odot M_{t-1} + z_t \odot \tilde{M}_t,
\end{aligned}
$$

GRU Internal Structure [2]



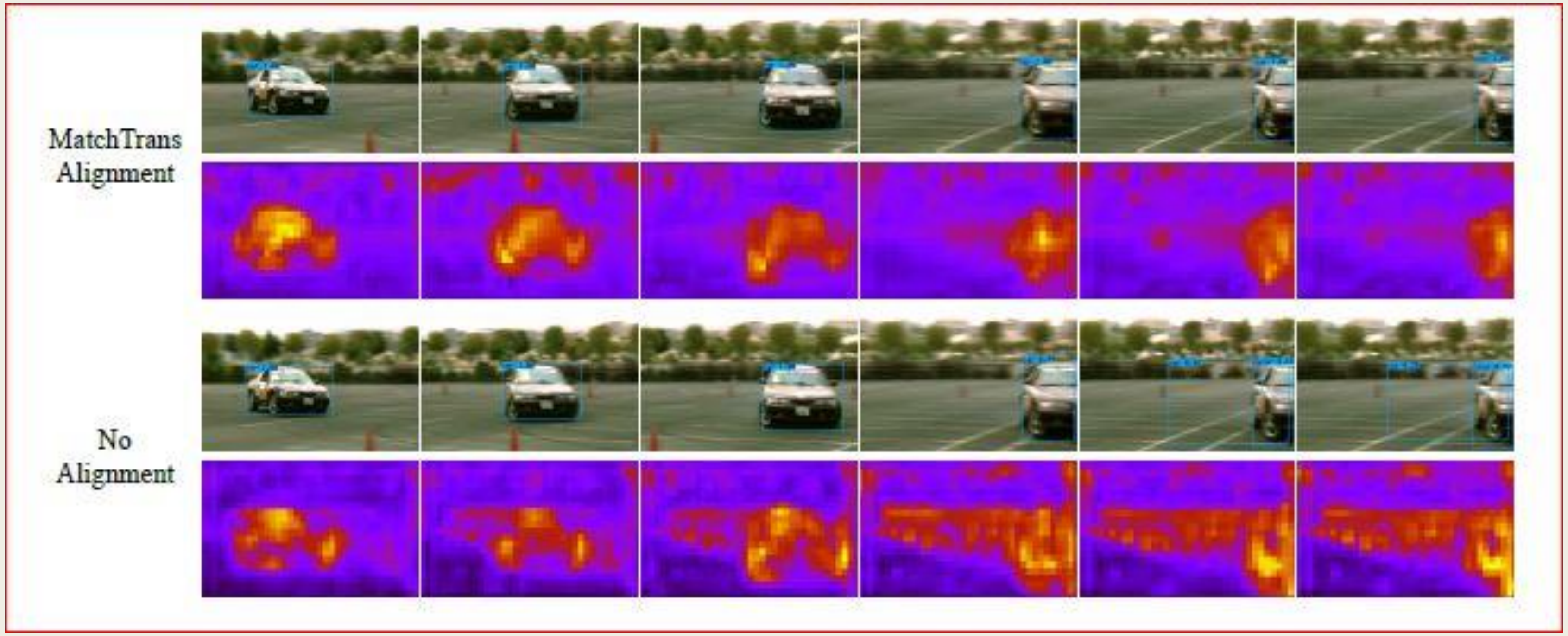
$$
\begin{aligned}
z_t &= \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)\\
r_t &= \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)\\
\tilde{h}_t &= \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)\\
h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t
\end{aligned}
$$

Zt : update gate
Rt : reset gate
h~t : Current memory
Ht : Final Memory

Methodology(Contd..)

- Here the weights of ConvGRN are replaced by weights of RFCN Static Image detector, and continue to fine tune it on ImageNet VID videos.

Problem of hallucination

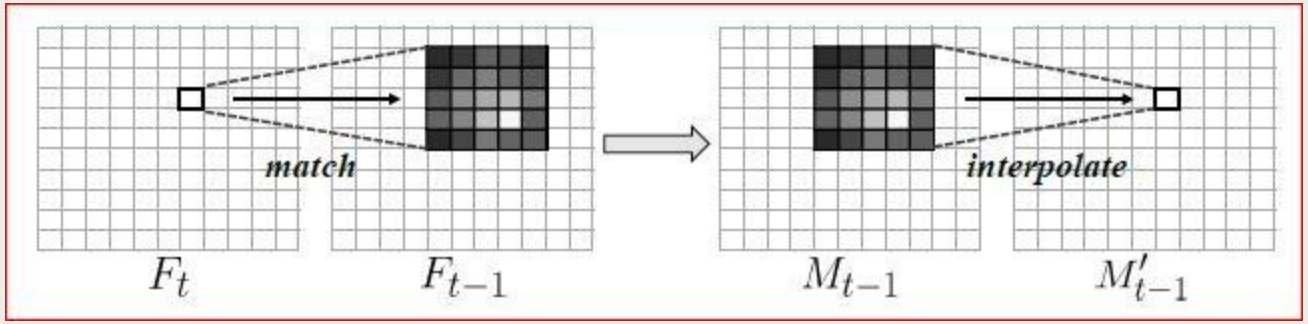


Methodology(Contd..)

- To avoid hallucination we use MatchTrans for proper alignment of memory

Transforming Coefficient

$$\Gamma_{x,y}(i,j) = \frac{F_t(x,y) \cdot F_{t-1}(x+i, y+j)}{\sum_{i,j \in \{-k,\dots,k\}} F_t(x,y) \cdot F_{t-1}(x+i,y+j)},$$



Memory Update into alignment

$$M'_{t-1}(x,y) = \sum_{i,j \in \{-k,\dots,k\}} \Gamma_{x,y}(i,j) \cdot M_{t-1}(x+i,y+j).$$

Dataset

- ImageNet VID Videos

Plan

- 1^{st} quartile : Reading research papers and gathering required information and dataset.
- 2^{nd} quartile : Implementing the paper.
- 3^{rd} and 4^{th} quartile :Trying new techniques to solve the problem and optimizing

References

1. http://fanyix.cs.ucdavis.edu/project/stmn/project.html (Research paper)
2. https://medium.com/@george.drakos62/what-is-a-recurrent-nns-and-gated-recurrent-unit-grus-ea71d2a05a69 (Article)
3. Kang, K., Li, H., Xiao, T., Ouyang, W., Yan, J., Liu, X., Wang, X.: Object detection in videos with tubelet proposal networks. In: CVPR (2017)
4. Zhu, X., Dai, J., Yuan, L., Wei, Y.: Towards high performance video object detection. CVPR (2018)
5. Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y.: Flow-guided feature aggregation for video object detection. ICCV (2017)