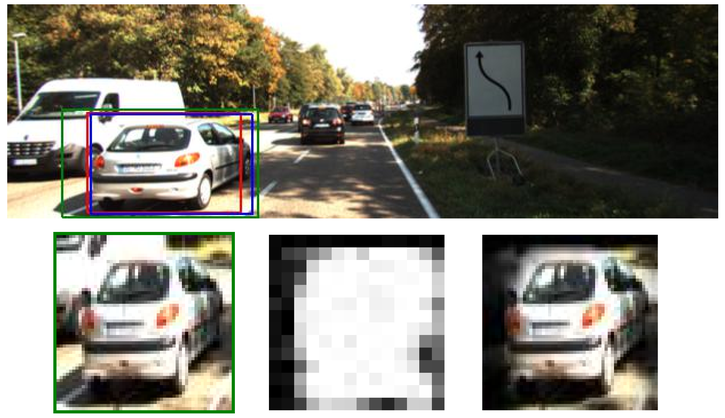
Hierarchical Attentive Recurrent Tracking

Adam R. Kosiorek, Alexy Bewley, Ingmar Posner

Applied AI Lab, Department of Engineering Science, University of Oxford, UK

Hierarchical Attention



Ground-truth and predicted bounding boxes and an attention glimpse. The bottom row shows the three layers of attention:

1^{st} layer extracts an attention glimpse (left)

2^{nd} layer uses appearance attention to build a location map (middle)

3^{rd} layer uses the location map to suppress distractors (right)

Loss

Directly optimise Intersection-over-Union (IoU) and guide attention mechanisms.

$$\mathcal{L}_{\text{HART}}(\cdot) = \lambda_{\text{t}}\mathcal{L}_{\text{t}}(\cdot) + \lambda_{\text{s}}\mathcal{L}_{\text{s}}(\cdot) + \lambda_{\text{a}}\mathcal{L}_{\text{a}}(\cdot) + \beta R(\cdot)$$

Tracking: Negative log of Intersection-over-Union.

$$\mathcal{L}_{\text{t}}(\mathcal{D}, \theta) = \mathbb{E}_{p(\widehat{\mathbf{b}}_{1:T}|\mathbf{x}_{1:T}, \mathbf{b}_1)}\Big[-\log \text{IoU}(\widehat{\mathbf{b}}_t, \mathbf{b}_t)\Big]$$

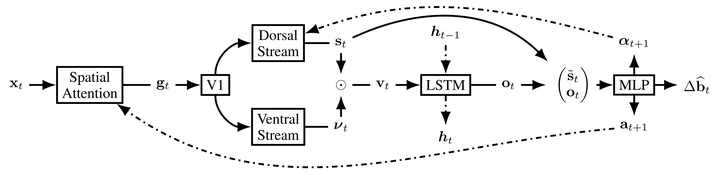
Spatial Attention: It follows the object, but shouldn't be too big.

$$\mathcal{L}_{\text{s}}(\mathcal{D}, \theta) = \mathbb{E}_{p(\mathbf{a}_{1:T}|\mathbf{x}_{1:T}, \mathbf{b}_1)}\left[-\log\left(\frac{\mathbf{a}_t \cap \mathbf{b}_t}{\text{area}(\mathbf{b}_t)}\right) - \log(1 - \text{IoU}(\mathbf{a}_t, \mathbf{x}_t))\right]$$

Appearance Attention: Cross-entropy with dynamically created target mask $\tau(\mathbf{a}_t, \mathbf{b}_t)$.

$$\mathcal{L}_{\text{a}}(\mathcal{D}, \theta) = \mathbb{E}_{p(\mathbf{a}_{1:T}, \mathbf{s}_{1:T}|\mathbf{x}_{1:T}, \mathbf{b}_1)}[H(\tau(\mathbf{a}_t, \mathbf{b}_t), \mathbf{s}_t)].$$

Two-stream Attentive Model



$\mathbf{x}_t$ input image

$\mathbf{g}_t$ attention glimpse

$\boldsymbol{\nu}_t$ appearance-based features

$\mathbf{s}_t$ object segmentation

$\mathbf{v}_t$ masked features

$\boldsymbol{h}_t$ hidden state

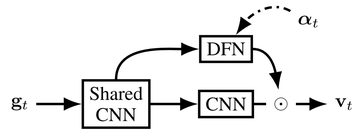
$\mathbf{o}_t$ LSTM output

$\boldsymbol{\alpha}_{t+1}$ appearance

$\Delta\widehat{\mathbf{b}}_t$ bounding-box update

$\mathbf{a}_{t+1}$ spatial attention

Appearance attention architecture:



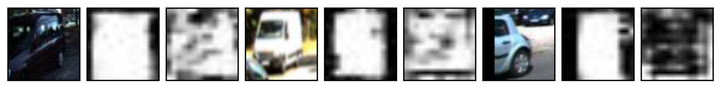
V1 shared CNN

Dorsal Stream Dynamic Filter Network (DFN)

Ventral Stream CNN

Is Attention Loss Important?

Appearance attention loss (top) prevents an ID swap when a pedestrian is occluded by another one (bottom).



Left to right: glimpses and segmentations learned with and without appearance loss. Attention loss leads to distractor suppression.

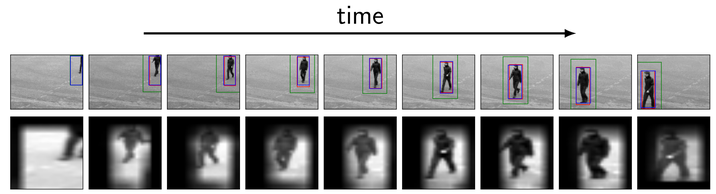
References:

[1] Samira Ebrahimi Kahou, Vincent Michalski, and Roland Memisevic. RATM: Recurrent Attentive Tracking Model. CVPR Work., 2017.

[2] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In ICPR. IEEE, 2004.

[3] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. Int. J. Rob. Res., 32(11):12311237, sep 2013.

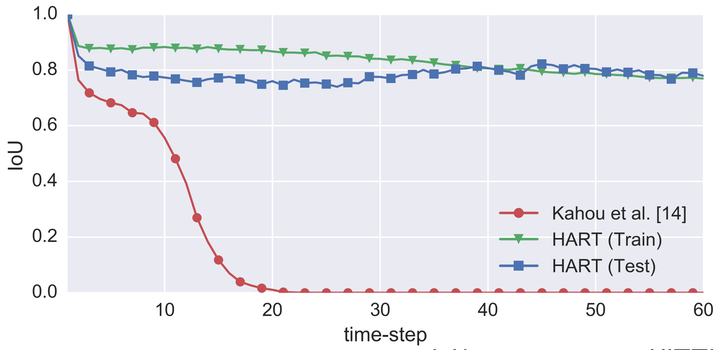
Pedestrian Tracking: KTH Dataset [2]



- Attention, prediction and ground-truth overlap at initialization
- Every 16^{th} frame of the sequence at 25 fps.
- 2^{nd} row shows attention glimpses multiplied with appearance attention

Intersection over Union	
Kahou *et. al.* [1]	Ours
55.03%	77.11%

Scaling to Real-world Data: KITTI [3]



Average IoU on KITTI over 60 time-steps			
Kahou *et. al.* [1]	Spatial Att	App Att	HART
0.14	0.60	0.78	**0.81**

IoU curves on KITTI over 60 timesteps. HART (train) presents evaluation on the train set.

Conclusion

Bio-inspired: Neural Recurrent tracking with Attention Mechanisms.

Interpretable: Important features selected by spatial attention and object segmentation mechanisms.

Scalable: Auxiliary loss terms allow scaling to complex real-world datasets.

Efficient: > 120 fps on a laptop!

Future Work: Multi-object tracking.