Single Object Tracking with Organic Optic Attenuation
Final Project Report

Ibraheem Saleh
California State Polytechnic University, Pomona
3801 W Temple Ave,
Pomona, CA 91768
iysaleh@cpp.edu

Abstract

The problem of object tracking given real world video surveillance footage is one that is being studied widely in the field of Computer Vision. In their paper, "Hierarchical Attentive Recurrent Tracking" (HART) [7], Kosiorek et al introduced a biologically inspired recurrent tracking architecture that they demonstrate as being superior to currently researched techniques for the purposes of single object tracking. The purpose of this research is to delve further into the biological tie-ins presented by Kosiorek et al and more explicitly understand how the work presented in the HART framework adequately mimics attention mechanisms present in human optics. We also attempt to execute HART on the KITTI test datasets [3] and a self-created experimental dataset to verify the accuracy metrics presented in their paper and better understand the object tracking instances where the framework fails to produce acceptable tracking results. In the end, a novel idea in the field of biologically-inspired object tracking is suggested which might be pertinent in addressing some of the failure cases that HART encountered in the test data evaluation.

1. Introduction

The age-old idiom, "A picture is worth a thousand words", refers to human-kind's ability to see an image and derive an understanding of its content by judging the meaning of each object in the image and the relationships between them. For a computer, the science of deriving information offered by an image source is much more underdeveloped than human judgment but significantly improving with the advent of computer vision, deep learning techniques and, recently, attention mechanisms. Object tracking in computer vision refers to the science of following objects as they change through time given a sequence of related frames. This task of tracking is very difficult as the images provided to a machine often exhibit significant continuity breaks due to motion blur, brightness differences, occlusion interference from other physical objects, source image noise and any number of other factors that must be taken into account. On top of that, the process of correctly segmenting the object of interest from its environment is also highly prone to error.

The true importance of efficient, accurate and reliable object tracking algorithms is realized in a plethora of modern computing applications. For example, object tracking algorithms are paramount to the success of optics based self-driving automobile solutions or artificial intelligence (AI) driven robot creations. If self-driving cars or robots are required to be able to make decisions in real-time, they need to be able to decipher any dynamic changes to the environment that is external from their control in real time. With the Hierarchical Attentive Recurrent Tracking (HART) framework introduced by Kosiorek et all [7], AI solutions are presented with an extremely performant way to understand how objects in the world morph through time.

In this paper we take a deep look at how the HART framework is built and try to understand its tracking results. The rest of this paper is structured as follows: Section 1.1 briefly reviews related research efforts that HART is heavily influenced by. Section 2 gives a detailed explanation about the challenges commonly faced by all object tracking solutions. Section 3 briefly tries to elaborate upon the design decisions which influenced HART and then quickly dives deeper into the actual components that comprise the model and how they are connected together. Section 4 goes over the datasets used to evaluate the HART framework and then moves onto both computational results and visually verified results of HART on multiple datasets. Section 5 concludes this research and briefly mentions a novel approach for the direction of future work which might improve upon some of the failure cases that HART exhibits. Finally, references for this research are mentioned at the end.

1.1. Previous Work

In 2015, Karol Gregor et al introduced a Deep Recurrent Attentive Writer (DRAW) neural network architecture for image generation [5]. At the heart of the DRAW framework was a pair of recurrent neural networks (RNNs): The first, an encoder that was utilized to downsample the source images that were fed to the neural network during training, and the other, a decoder that was utilized to generate images based on input "codes". The model was designed to have both a read mechanism that would extract a subsection of

the input image and learn the features which made it unique and a write mechanism that would reconstruct a generated output image in a contiguous manner similar to how a human might paint an image. Beyond that, [5] claimed that their attention mechanism emulated foveation techniques similar to those found in the human-eye through the step-by-step generation of an output image.

Significantly influenced by the discoveries in [5], Kahou et al created their Recurrent Attentive Tracking Model (RATM) [6] which utilized both attention mechanisms and recurrent neural networks to track bouncing balls, moving digits and people. Kahou et al theorized that the process which allowed the reconstruction of handwritten digits in a sequential manner in [5] could be useful in object tracking applications where the tracker mechanism is tasked with tracing an object through frames fed time-sequentially into the model. Their model included a convolutional neural network (CNN) to extract features and an RNN to track the object as it moved. Though [6] introduced a unique approach to object tracking, their final results were ultimately fairly poor at between 50 and 65 percent accuracy despite using the relatively simple ball tracking, KTH and mnist datasets which don't present real world tracking challenges such as dynamic lighting changes, object occlusion and other distractors.

In [7], Adam Kosiorek et al introduced their "Hierarchical Attentive Recurrent Tracking" (HART) framework that incorporates concepts strongly influenced from both of the aforementioned papers in an attempt to create a system that would track objects in a manner akin to the way the human eye functions. The approach to tackle the problem of object tracking utilized by the HART framework will be the subject of the rest of this paper.

2. Problem Description

The challenge of creating computer applications that have the ability to track objects given a series of 2D-digital images is essentially as old as image processing itself.

The difficulty of object tracking stems from a huge number factors that affect how and what must be tracked. For example, input images which contain moving objects generally have substantial and varying degrees of motion blur that obfuscate the object being tracked. In addition to the visual blur that accompanies motion, the direction of the motion itself is often entirely unpredictable by heuristic based tracking models which might use velocity or other mathematically-derived probabilities to determine the general region where to expect an object in the next frame given a current frame. This is because organic objects, or organically controlled objects, can shift direction at any moment based on the objects internally derived intention to move. The object being tracked can also exhibit rotational orientation changes or scale shifts caused by turning around or moving closer to or farther from the camera recording the images. There is also the possibility that the camera recording the scene is moving in tandem with the object being tracked which can further complicate the tracking task.

In many cases, object tracking algorithms must also account for varying levels of partial occlusion or even momentary total occlusion from other world-objects such as, for example, trees on a path. Even more challenging for this task is when the object of interest becomes temporarily occluded by an object that contains some structural or feature-based similarities to the object desired to be tracked. Lighting variations which cause the tracked object's illumination to lighten or darken rapidly or its shadow to be cast unpredictably at any given moment can increase the complexity of the scene that the tracker must process.

Beyond the aforementioned challenges, many object tracking applications levy the extra requirement that tracking programs must be capable of processing streams of input data in real-time. This requirement immediately disqualifies many traditional tracking approaches that achieve their success by processing the entire image of every frame with various post-processing computer vision techniques or computationally intense motion prediction algorithms.

3. Methodology

The "Hierarchical Attentive Recurrent Tracking" (HART) model introduced by Adam Kosioerek et al [7] attempts to tackle the challenges presented in the domain of object tracking by having their model imitate human eye visual mechanisms which are generally capable of handling all of the aforementioned difficulties. HART presents, "a biologically-inspired recurrent model for single object tracking in videos...inspired by the general architecture of the human visual cortex and the role of attention mechanisms" [7] In order to understand the ways in which the HART model mimics human optics, one must first understand the organic components which enable human vision. After that, it is then possible to delve deeper into the HART model and connect each component to its respective human counterpart.

3.1. Human Vision

Contrary to many modern computer based vision models, human vision doesn't work by processing every part of every image that the eyes transmit to the brain for processing while it is tracking an object. Instead, biological studies have shown that when the eyes receive an image from the combined visual field, it transmits the picture to the primary visual cortex which then subdivides the image processing stage between the ventral stream and the dorsal stream parts of the human brain. [9] This processing path is diagrammed in figure 1 below.

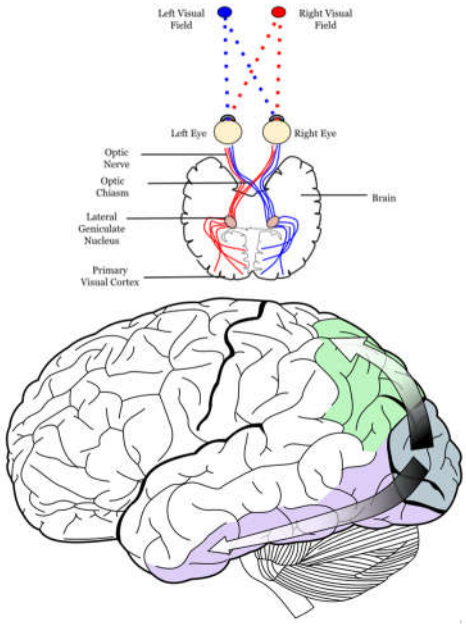

Figure 1: Diagram of human vision processing pathways. In the bottom portion, green represents the dorsal stream and purple represents the ventral stream. The blue section represents the primary visual cortex. [11]

The dorsal stream component of human image processing is believed to be responsible for discerning where the object of interest is in any given scene. It is responsible for blurring out and quickly discarding from vision the parts of the scene that are not relevant to what the eyes are attempting to focus on. Human vision relies on this spatial attention mechanism due to limited processing power within the human mind. Kosiorek mentions, "Whenever more than one visual stimulus is present in the receptive field of a neuron, all the stimuli compete for computational resources due to the limited processing capacity [of the human mind]" [7]

The ventral stream component of human image processing is believed to be responsible for targeting and focusing on the features of any tracked object that make it unique from its environment. [9] By learning what the object is, the ventral stream enables the human mind to track a moving object through time despite the fact that it may blend into its surroundings or be partially occluded by other objects.

Beyond the raw image processing aspects of vision perception in general, the human brain functions with memory which plays a crucial role in remembering the features that are unique to the object that is being tracked and providing informational clues to the spatial attention mechanisms that enable it to track the object of interest despite significant external changes over time.

3.2. The HART model

In order to attempt to mimic the structure behind human vision mechanisms, the "Hierarchical Attentive Recurrent Tracking" model creates a series of different machine learning components and connects them. A diagram for the HART model can be seen in figure 3 below and each of its components will be described in this section. At the root of the HART model, Akosiorek et al created a spatial attention mechanism which extracts glimpses from the input images. From there, the glimpses get passed onto the V1 component which is intended to imitate the function of the human primary visual cortex. From V1, the glimpse is then split and passed onto the Dorsal and Ventral Stream components. The output of both of these components is then combined and passed onto an LSTM which serves the function of human memory. The LSTM output is then used to compute predictions for the attention components in the next source image frame.

3.2.1 Spatial Attention and Glimpses

As shown in figure 3, the first component of the HART model that the images pass through is the Spatial Attention mechanism. This component constructs two matrices, one relating to the x-dimension and the other the y-dimension. Each matrix contains one Gaussian per row which eventually construct the final glimpse. Figure 2 shows an example of a glimpse extracted from an input image from a video sequence.

Figure 2: Example of glimpse extracted from input image

The position of the initial glimpse is manually inputted per this model as a 4-paramater system which bounds the object desired to be track: (y_start,x_start,y_gain,x_gain). After the initial time-step, the position of glimpses from future image frames is determined from the output of the LSTM which calculates the strides and centers of the positional variance of the object from the hidden layers of its RNN model.

In order to combat the chance that the extracted glimpses in future time windows both contains the object being tracked and does not grow too large, the loss function shown in Equation 1 is utilized.

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

Equation 1: Spatial Attention loss function designed to stop the glimpse from growing too large or too small over time. [7]

The variables utilized in the equation are as follows: L_s is the spatial attention loss function. D is the input image dataset and θ represents the neural network parameters. E_p

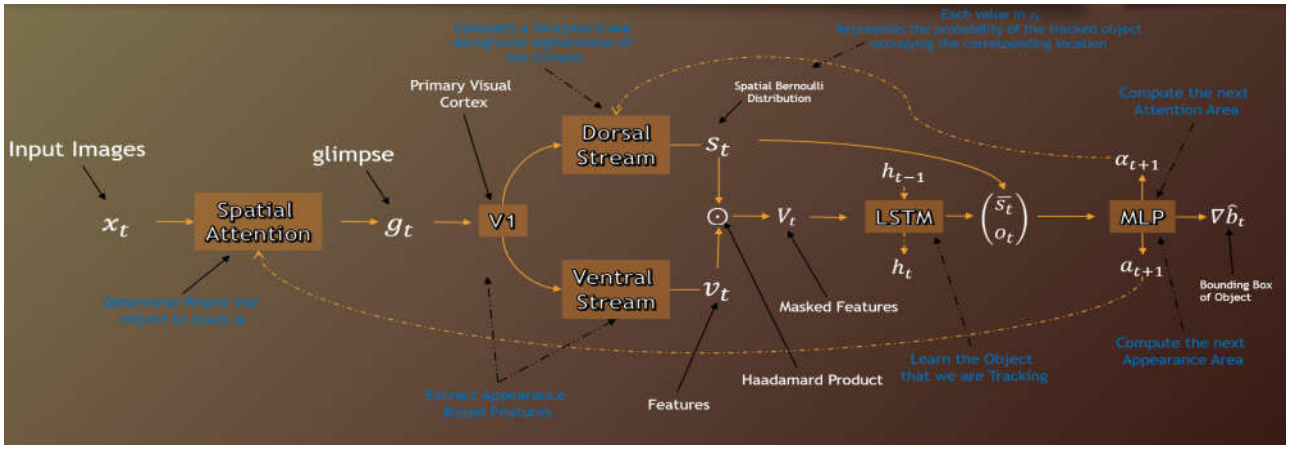

Figure 3: Hierarchical Attentive Recurrent Tracking framework with brief explanations for all components. Dotted lines are temporal connections while solid lines show the information flow within 1-timestep

refers to the expectations of the probability. a_1 refers to the attention calculation from the LSTM, T refers to the complete range of time from t=0 to t=T. b refers to the bounding box that gets outputted from the MLP. x is the input image. IoU means intersection over union—It is used to calculate the error of a bounding box given another bounding box by taking their intersected areas divided by the complete area of both boxes unioned.

The first loss term in equation 1 attempts to ensure that the predicted attention is limited to cover only the bounding box while the second loss term tries to stop it from growing too rapidly. The logarithm of both terms is taken with every frame step to avoid rapid changes in the overall loss.

3.2.2 V1 and the Dorsal and Ventral Streams

Once the glimpse is extracted from the spatial attention mechanism, it is then passed on to the primary visual cortex component V1. V1 is implemented as a convolutional neural network (CNN) which transforms the glimpse through a series of convolutional and max-pooling layers. The output of V1 is then shared between the ventral and dorsal stream processing components (see figure 3).

The ventral stream is also implemented as a CNN and is designed to output feature maps from the input visual features from V1. In our case, it utilizes pre-trained weights from Alexnet [4]. These weights are taken from the network at the conv3 56x56 layer instead of the later layers since weights from later layers would require further image downsampling to be compatible as a 14x14 which would be detrimental to the object tracking task since that is likely not enough resolution to accurately identify anything.

The dorsal stream is implemented as a dynamic filter network (DFN) [1]. Different from a CNN, a DFN outputs filters which describe an image and continues to learn weight shifts in real-time instead of stopping all model-training after the initial training phase. The dorsal stream is responsible for handling spatial relationships between the object being tracked and its environment. It achieves this by computing a background and foreground segmentation of the glimpse and ultimately creating a location map which acts as a kind of image mask that can be used to identify where in the glimpse the object of interest lies.

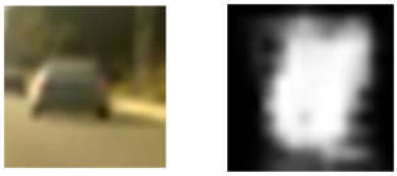

Figure 4: Example dorsal stream location map output from glimpse image passed through V1.

Using the location map from the dorsal stream and the appearance attention from the ventral stream, the haadamard product of both outputs is calculated to create the final output vector containing appearance and spatial information about the tracked object which will be passed into the LSTM. The combined dorsal and ventral stream output mimics the distractor suppressing behavior of the human-brain. [7]

Figure 5: Example output of final attention mechanism. Leftmost box is attention output. Middle box is the location map from the dorsal stream. Rightmost box is the combined output which is fed into the LSTM. [7]

The overall appearance attention mechanism is further constrained to ensure that it suppresses distractors by masking the output where the bounding box from the spatial attention mechanism does not match the features from the appearance attention mechanism. The overall loss function for the appearance attention mechanism can be realized in equation 2 below.

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

Equation 2: Appearance Loss Function as formulated in [7]

Variable references in equation 2 are mostly the same as those found in equation 1, however, new to equation 2 are L_a which stands for the appearance attention loss. H refers to a cross-entropy calculation where $H(p,q) = -\sum_z p(z)\, log\, q(z)$. s_t refers to the spatial Bernoulli distribution output from the dorsal stream. The τ operator simply performs a 0/1 type image masking operation given an attention and a bounding box at time t.

3.2.3 Memory Based LSTM and the MLP

The final components of the HART model are the Long-Short-Term-Memory (LSTM) RNN module and the multi-layer-perceptron (MLP) that ties all of the model pieces together through time. An LSTM is used here since, as shown in [5,6,7], it is able to learn the features and spatial information provided from the combined dorsal and ventral stream output in a temporally-relevant manner while being fed time-related frames. This property of LSTM type neural networks allows for the DFN in the dorsal stream and the spatial attention mechanism to dynamically learn and adapt over time according to how the tracked object transforms through the input image sequence. The LSTM assumes a Markovian state property about the motion of the object being tracked in that it only regards information about the objects location and features from the previous frame as being pertinent to learning its current location. This is in contrast to traditional models which generally utilize all the location information gathered from time and use heuristics to predict where it might be in the next frame.

The output of the LSTM and a vector relating to the masked component of the dorsal stream output is passed to the final MLP layer which then handles the current frame bounding box computation and passes on spatial and appearance information to the spatial attention mechanic and the dorsal stream mechanism for use in the subsequent frame processing.

4. Results

For the purposes of evaluating the Hierarchical Attentive Recurrent Tracking framework presented in [7], we will conduct both computational and visual verifications over two datasets: 1-The KITTI dataset and 2-the "experimental dataset".

4.1. Datasets

There are two datasets which will be evaluated: the KITTI dataset [3] and an experimental dataset created specifically for the purpose of evaluating HART under varying real-world environment and subject circumstances.

4.1.1 KITTI Dataset

The Karlstuhe Institute of Technology and Toyota Technological Institute (KITTI) [3] dataset is a collection of 21 training and 29 test video sequences that are used to evaluate object tracking algorithms in real world scenarios. The primary focus of this dataset is to evaluate an algorithms ability to track objects from Car and Pedestrian classes given instances where the subjects of interests might disappear from view temporarily due to occlusion from the environment or might change their orientation over time or be subject to different amounts of lighting. The video sequences in this dataset are akin to the kind of images that a self-driving car might generate given their position relative to the external buildings, pedestrians and cars and the way that the source camera moves only along straight roads and making simple 90 degree angled turns.

4.1.2 Experimental Dataset

For the purposes of this research, the term "experimental dataset" refers to the collection of 18 video sequences which include people walking down streets, cars driving on the road, cats moving in a house and apples rolling on a table. Unlike the KITTI dataset which contains only about 1 frame per second of input data, this collection includes 23 frames per second and feeds every frame into the HART model for testing purposes. Since these video sequences are recorded by a human, they experience source camera shaking and recording height variations. No ground truth data is provided with this dataset and all evaluation metrics must be visually deduced.

4.2. Dataset Object Tracking Results

When ground truth data for the bounding box which encompasses an object to be tracked is provided, the accuracy of a model can be evaluated using a simple intersection over union evaluation of the model's predicted bounding box with the human provided bounding box. If this is not provided, as is the case with the KITTI test datasets and the experimental dataset, only visual accuracy conclusions can be drawn.

4.2.1 KITTI Train Test Split Computational Results

In order to evaluate HART with the KITTI train dataset which included labels with bounding boxes for each frame for each object in the scene, Kosiorek et al divided the dataset 80/20 for train and test. After training their network with parameters tuned for the task, HART demonstrated on average 81% intersection over union area for their predicted box versus the provided ground truth bounding boxes from KITTI [7]. These results are superior to the tracking performance achieved by similar, previous models.

4.2.2 KITTI Test Data Visual Verification Results

The test image sequences provided by the KITTI dataset [3] do not include the associated tracking labels that are necessary for the computation of the intersection-over-union bounding box errors that were calculated in the 4.2.1 results. Since HART out of the box isn't capable of doing object class identifications and is inherently a single-tracking model where KITTI presents a multi-class multi-object tracking challenge, they aren't quite compatible. Therefore, the only verification of KITTI test data sequences that can be done is manually through visual verification.

For the majority of tested KITTI test data sequences, HART performed fairly well and tracked the pedestrians and cars with little error. There were, however, one general fail case that was observed.

The major failure that was observed with HART was its inability to learn medium distanced-cars that completed a turn within a low number of frames (see Figure 6 below).

Figure 6: KITTI Test Dataset Example of HART model tracking loss of focus on distant car that makes a turn over a period of a few seconds. The red bounding box is attention and the blue box is the object tracking prediction.

In the KITTI test image sequence relevant to figure 6 above, the HART model had no difficulty in tracking the vehicle at various distances so long as it was positioned parallel to the source camera direction. However, as soon as the vehicle initiated a left turn, both the spatial and appearance attention mechanisms began to place weight on the background instead of dynamically learning the shifted orientation of the car intended to be tracked. In this case, the entire system proved that it could not learn the changes presented by the tracking subject fast enough to track it.

4.2.2.1 Experimental Data Results

Similar to the KITTI tracking test dataset, visual verification of results on the experimental datasets demonstrated that, in real-world scenarios, the HART model tracks objects fairly well even if they undergo distance and perspective variations or exhibit partial occlusion. There were, however, two consistent tracking failures that the framework handled poorly for this dataset.

The first and most concerning test failure pattern that was observed over multiple video sequences is that whenever the subject being tracked would pass through a region of significant darkness or significant lightness, the area of the prediction box would become greatly expanded beyond the region that the subject actually occupies. In the case of the subject entering a dark environment, IE figure 7, the spatial attention mechanism would still encompass the object being tracked but the appearance features being learned would cause the LSTM to expand the prediction box significantly.

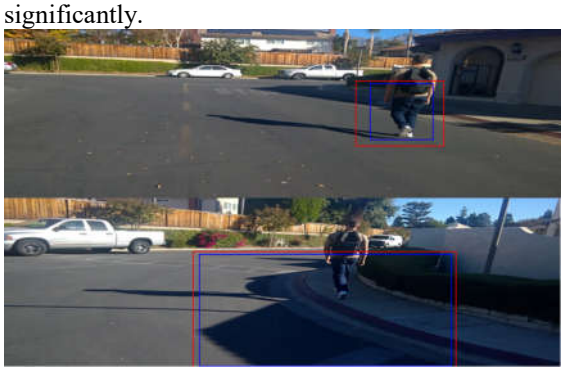

Figure 7: Experimental Dataset demonstrating how dark lighting scene shift causes significant prediction accuracy degradation. The red bounding box is attention and the blue box is the object tracking prediction.

In the case of the strong light source entering the attention box, the dorsal and ventral stream mechanisms appeared to become fixated on the strong source of light and completely lost focus of the object supposed to be tracked (figure 8).

Figure 8: Experimental Dataset - Presence of strong light source in environment causes complete loss of focus of subject being

tracked. The red bounding box is attention and the blue box is the object tracking prediction.

The second test failure pattern that was observed on the experimental dataset is that whenever the object being tracked would temporarily go off screen, even if it was just for a moment while the source recording camera adjusted for a turn, the spatial attention mechanism responsible for locating the glimpse establishing where the object might be in the scene was never able to recover and find the object being searched for. Instead, it immediately began to provide irrelevant information to the primary visual cortex which would then begin to learn whatever was being tracked despite not being the intended focus (Figure 9).

Figure 9: Example of Car on Road being tracked and spatial attention mechanism unable to regain focus after going off screen.

5. Conclusion

The "Hierarchichal Attentive Recurrent Tracking" framework presented by Akosiorek et al [7] has demonstrated that it is superior to previous methods in tracking real-world objects through environment based challenges such as partial occlusion, distance shifts and other changes. In this paper, we have taken a look at how previous or related object tracking and neural network models have achieved considerable advancements in the field of computer vision. We then followed the HART design process and traced in which ways the model designed by Kosiorek et al imitates human vision. In our results, we noted that despite achieving fairly high accuracy over the KITTI training an test datasets, there still remains some significant object tracking challenges that HART failed to handle.

5.1. Future Work

One of the object tracking challenges that the HART framework was unable to overcome was the case where the object being tracked would momentarily move off screen. For the purposes of handling this failure case, I propose the following novel idea for future research. Regarding the imitation of human vision, something overlooked in the HART model in the machine to human component connections is that humans don't actually continuously track an image. In reality, every few seconds, at varying intervals, the human eyes blink and transmit a completely dark and blank image to our primary visual cortex. After the eyes complete their blink and reopen, a momentary period of readjustment is undergone by which human optic systems re-find and re-focus on the subject being tracked. To continue the theme of imitating human optics, I propose that HART introduce a blink mechanism based on a recurrent neural network model that is fed the glimpses from the primary visual cortex at a varying rate to simulate the human blinking function. Instead of having this RNN rely on a Markovian state like the LSTM module and only looking at the previous frame for relevance information, it would instead maintain a moderately sized memory of all the previous blink frames which would, periodically, re-search the entire image for the subject of interest and lock on the object which best matches all the images in memory. In theory, this would counteract the problem that occurs when the spatial attention mechanism is unable to recover due to momentary losses of vision of the tracked object.

5.2. Project Resource Links

- Project Code: https://github.com/iysaleh/cs519_project_code
- Experimental Data: Link in Github above
- KITTI Data: http://www.cvlibs.net/datasets/kitti
- HART Code: https://github.com/akosiorek/hart

References

[1] Bert De Brabandere, Xu Jia, Tinne Tuytelaars, and Luc Van Gool. Dynamic Filter Networks. NIPS, 2016.

[2] Brian Cheung, Eric Weiss, and Bruno Olshausen. Emergence of foveal image sampling from learning to attend in visual scenes. ICLR, 2017.

[3] Geiger, Andreas, et al. The KITTI Vision Benchmark Suite, www.cvlibs.net/datasets/kitti/eval_tracking.php.

[4] Guerzhoy, Michael. AlexNet implementation + weights in TensorFlow. November, 2017.

[5] K Gregor, I Danihelka, A Graves, and D Wierstra. DRAW: A Recurrent Neural Network For Image Generation. ICML, 2015
[6] Kahou, Samira Ebrahimi, et al. “RATM: Recurrent Attentive Tracking Model.” [1510.08660] RATM: Recurrent Attentive Tracking Model, 28 Apr. 2016, arxiv.org/abs/1510.08660.
[7] Kosiorek, Adam R., et al. “Hierarchical Attentive Recurrent Tracking.” [1706.09262] Hierarchical Attentive Recurrent Tracking, 5 Sept. 2017, arxiv.org/abs/1706.09262.
[8] Laptev, Ivan, and Barbara Caputo. “Recognition of Human Actions - Database” Recognition of Human Actions, 18 Jan. 2005, www.nada.kth.se/cvap/actions/.
[9] Peter. Dayan and L. F. Abbott. Theoretical neuroscience : computational and mathematical modeling of neural systems. Massachusetts Institute of Technology Press, 2001.
[10] Sabine Kastner and Leslie G. Ungerleider. Mechanisms of visual attention in the human cortex. Annu. Rev. Neurosci., 23(1):315–341, 2000.
[11] Selket. (n.d.). Ventral-dorsal streams. Retrieved November 21, 2017, from https://commons.wikimedia.org/wiki/File:Ventral-dorsal_streams.svg