A Study in Human Attention to Guide Computational Action Recognition

by

Sam Sinai

B.S., Massachusetts Institute of Technology(2012)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2014

Author .....
Department of Electrical Engineering and Computer Science
January 17, 2014

Certified by .....
Prof. Patrick H. Winston
Ford Professor of Artificial Intelligence and Computer Science
Thesis Supervisor

Accepted by .....
Prof. Albert R. Meyer
Chairman, Masters of Engineering Thesis Committee

A Study in Human Attention to Guide Computational Action Recognition

by

Sam Sinai

Submitted to the Department of Electrical Engineering and Computer Science
on January 17, 2014, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Computer Science and Engineering

Abstract

Computer vision researchers have a lot to learn from the human visual system. We, as humans, are usually unaware of how enormously difficult it is to watch a scene and summarize its most important events in words. We only begin to appreciate this truth when we attempt to build a system that performs comparably. In this thesis, I study two features of human visual apparatus: Attention and Peripheral Vision. I then use these to propose heuristics for computational approaches to action recognition. I think that building a system modeled after human vision, with the nonuniform distribution of resolution and processing power, can greatly increase the performance of the computer systems that target action recognition. In this study: (i) I develop and construct tools that allow me to study human vision and its role in action recognition,(ii) I perform four distinct experiments to gain insight into the role of attention and peripheral vision in this task, (iii) I propose computational heuristics, as well as mechanisms, that I believe will increase the efficiency, and recognition power of artificial vision systems. The tools I have developed can be applied to a variety of studies, including those performed on online crowd-sourcing markets (e.g. Amazon's Mechanical Turk). With my human experiments, I demonstrate that there is consistency of visual behavior among multiple subjects when they are asked to report the occurrence of a verb. Further, I demonstrate that while peripheral vision may play a small direct role in action recognition, it is a key component of attentional allocation, whereby it becomes fundamental to action recognition. Moreover, I propose heuristics based on these experiments, that can be informative to the artificial systems. In particular, I argue that the proper medium for action recognition are videos, not still images, and the basic driver of attention should be movement. Finally, I outline a computational mechanism that incorporates these heuristics into an implementable scheme.

Thesis Supervisor: Prof. Patrick H. Winston
Title: Ford Professor of Artificial Intelligence and Computer Science

Acknowledgments

In 2005, when I was graduating Allameh Helli high school in Tehran, Iran, I would never believe that I would be researching in Artificial Intelligence at MIT nine years later. It has been an amazing period of my life, and this work reflects a small fraction of that journey. It would have not been possible without the support of the people I interacted with over these years, and I would like to use these few lines to express my immense gratitude towards them:

To Prof. Patrick H. Winston, who supervised me on this project, and has been one of the most inspirational people I have met in my academic career. I learned a lot from his patient, wise and supportive supervision. I became a better thinker, a better writer and a better speaker. Wherever I go as a scientist, engineer or who-knows-what in the future, Prof. Winston's impression on me will be one of the most significant. I am also indebted to him for offering me the freedom to explore my scientific interests in any area I liked. Without that, I would have been lost today.

To Nikola Otasevic, and Pedram Razavi, helped me with parts of this project. They both provided insights and actual material that I have used in this thesis. Working with them has been a pleasure for me, and I hope that I will have the great chance to work with them in the future.

To my professors at MIT, in particular Prof. Robert C. Berwick and Prof. Constantinos Daskalakis. For the undeniably excellent education they have blessed me with.

To Melina, for without her incredible patience, support, and help during my most difficult moments, I would never be here, writing this.

To My Family: Khosrow, Farah, Gizella, and Alma. For understanding and supporting me to travel over continents and oceans, in pursuit of a better education. And for their endless and unconditional love.

To my friends at the Genesis group, MIT, and Boston, who made this experience outstanding.

To you, dear reader, that somehow found the time to read this page, and hopefully

this thesis.

And of course to my anonymous subjects, for their time and effort.

The experiments in this study have been reviewed and approved by Committee on the Use of Humans as Experimental Subjects at MIT.

This research was supported, in part, by the Defense Advanced Research Projects Agency, Award Number Mind's Eye: W911NF-10-2-0065.

Contents

List of Figures

List of Tables

Chapter 1

Vision

1.1 The Problem

Consider the following: A man *walks* into a room carrying a briefcase. A woman is *standing* on the other side of the room. The man *approaches* the woman and *puts down* the briefcase. The woman *gives* the man a folder. The man *opens* the briefcase, *puts* the folder in the briefcase, *picks up* the briefcase and *leaves* the room.

What you just read was a sequence of actions, performed by two people. It is not difficult to believe that if you saw this scene you could easily write up a summary similar to what is written above. Nevertheless, such an inference cannot as easily be made if a computer were looking at the same scene. In fact, recognizing actions such as those in the first paragraph is a challenge that is yet to be solved by computer vision. Furthermore, note how the human visual system is able to infer all these actions from a single frame or a very short sequence of frames. For instance, I invite the viewer to look at figure 2-1, *before* reading on.

While you were observing the figure, you probably looked at multiple locations, including the child's face and where he was looking. If I asked you whether the child is holding the stick, what would you answer? Most people say no and that's correct. However, more importantly, where did you look? I am willing to bet my degree that you immediately looked at the child's right hand. Maybe later you scanned the length of the stick, specifically the end that is caught in the child's pants to confirm your

Figure 1-1: What is happening?

answer, and find an alternative explanation for the state of the stick. Had I not primed you and instead let you freely look at the photo before asking you the question, it might have taken you longer (in milliseconds) to find the important location. But that is besides the point: While attempting to answer this question, you directed your *attention* to multiple locations, and analyzed each sequentially.

There are two characteristics of human vision that could help a computer recognize actions more efficiently:

1. **Attention.** Attention is a high level cognitive process that is cross-modal. Namely, you can use your attention in vision, hearing, touching or multiple of these at the same time. *Visual attention* in particular refers to the characteristics of attention when it is deployed in the vision domain. I am interested in investigating the role of this feature in action recognition.

2. **Visual acuity.** The acuity of vision in humans and most animals is not uniform across the visual field (Kranz, 2012). A central region of our vision consists about 1% of the visual field (known as the *fovea*), whereas approximately 50% of the visual cortex is dedicated to the analysis of that region (Kranz, 2012). While one usually attends to the region in the visual field with highest acuity,

these are distinct features of human vision. It is therefore interesting to study the role of foveal and non-foveal (peripheral) vision in action recognition.

Current systems designed to address questions about actions rarely use attention (or foveation). More importantly, they rarely work with dynamically changing input (videos), and instead restrict the analysis to static frames. This is understandable, as we are fairly ignorant about the underlying processes that are responsible for visual search in humans, in particular when action recognition is concerned. Most studies in the psychology domain have been focused on the phenomenological and behavioral aspects of visual search. Computational models of visual attention exist, but are applied to more primitive aspects of visual search. I discuss these models, as well as their shortcomings in the next chapter. For now, it suffices to know that most of the success in computational approaches was achieved when basic image features drove attention (bottom-up attention). If you were to look at Figure 1-1 without being asked a question, most likely, you would use a part of the bottom-up apparatus to begin your search. However, for the second question, you probably used your top-down attention, driven by higher level cognitive processes such as memory, to direct visual search. Computational models of attention have been less successful in this second domain. Before we proceed, let me demonstrate an example of the performance of a bottom-up computational model proposed by Walter and Koch (Walther and Koch, 2006). Running this method on Figure 1-1, one can see that the algorithm does reasonably well in finding the locations that human's attention is drawn to.

I used, improved and developed tools to study human attention. I used these tools, described in the next section, to formulate a subproblem in the domain of attention. Further, I suggest how the insights gained from my experiments can be used to make more efficient visual systems, built to recognize actions.

1.2 The Approach

To work towards a better understanding of the problem at hand, and taking steps towards solving it, I have pursued research in several parallel lines. Thus, my re-

Figure 1-2: Running the bottom-up attention model, the essential parts of the image are readily attended. In this case, the algorithm starts from the face and then attends to the hand, somewhat similar to what the freely-observing human would do

search entails efforts in multiple different types of research including tool development, human experiments, and computational modeling. To clarify the motivation and progress of each, I will summarize them in this section.

1.2.1 Tool development

Throughout this project, I have developed two separate tools that I used in order to studying human foveal vision and attention:

1. **Eye-tracker development.** During this project I worked with the original developers of the PUPIL(Kassner and Patera, 2012) eye tracker headset to improve their device and augment it's functionality to be used for my own purposes. I built a first-prototype headset for PUPIL, using Xbox cameras. I also received another headset built by the developers of PUPIL, which has a smaller eye-camera with specifically designed hardware and a world-view camera

with a fisheye lens. I have used both of these devices to collect data from approximately a dozen human subjects.

2. **Mechanical Turk platform.** For large scale data collection, I also developed two distinct pipelines through which I was able to publish vision experiments on Amazon's Mechanical Turk. These tools enabled me to perform simple experiments on attention and acuity of vision.

   These tools are explained in greater detail in appendix A.

1.2.2 Human experiments

In order to gain a better understanding of how human attention works, in particular when applied to action recognition, I have designed and conducted four experiments using the tools I described above. The questions that motivate these experiments are listed below:

1. **Do all humans look at the same locations when searching for the same verb?** I have conducted two studies to address this question. The first is a pilot study using the eye-tracker. The second is a large scale study that I have conducted on Amazon's Mechanical Turk. Chapter 3 is dedicated to these experiments.

2. **How critical is the role of peripheral vision in action recognition?** I have conducted two online studies for this question. In the first (pilot) study, I partially eliminate peripheral vision, and look the information that is lost as a results. In the second study, I tried to strengthen the peripheral vision by allowing subjects to attend the periphery, without allowing them to foveate at those locations. The results are discussed in chapter 4.

1.2.3 Discussion and insights for computational models

I have used the results from these experiments, in order to provide better heuristics for computational action recognition. In chapter 5, I propose several heuristics that

can improve our current approach to action recognition. Furthermore, I propose a general, but implementable mechanism, that could use these heuristics in order to generate an attentive vision, dedicated to action recognition.

1.2.4 Contributions

Based on the experiments in this thesis, I provide three contributions to the understanding of attention, as it is applied to action recognition.

1. I observe that **when multiple humans attempt to determinate the occurrence of a particular action in a scene, they look at similar locations.** These observation drawn from the experiments in chapter 3.

2. I propose that while foveal vision is the primarily apparatus for action recognition, **peripheral vision**, **is key to guiding attention towards the location of action**. Hence, the indirect role of peripheral vision is also fundamental to successful action recognition, as it guides the limited-capacity and expensive attentional processing to the locations where action recognition is made. The suggestive evidence for this proposal is provided in chapter 4.

3. I suggest that for building computational systems, **videos, as opposed to static frames, are the appropriate input for training action recognition algorithms**. I address the shortcomings of the static frames in chapter 2. In chapter 5, I propose heuristics for efficiently using videos as inputs.

Chapter 2

Background and Previous Work

In order to motivate the experiments, I provide a background of what is already known in the areas related to attention and computational action recognition. Specifically, I provide an overview of the previous work done in three areas:

1. Psychological studies on human visual attention and visual acuity.

2. Computational approaches to modeling attention.

3. Computational approaches to action recognition.

While the emphasis of this chapter is placed on studying and modeling attention, my work is only partially rooted in these studies. One has to appreciate that even though much progress has been made in each of these domains, due to extreme complexity of the matter, we are still far from a thorough and objective description in each of the named categories. Moreover, it is beyond the scope of this thesis to offer a comprehensive overview of three active areas of research. Thus, one should read this chapter to gain a general understanding of each of these fields, while keeping in mind that this is by no means the entire story. When you are done reading this chapter, I hope that I have convinced you that attentional models should be included in computational action recognition systems. I also demonstrate that a great amount of research in modeling attention is conducted with static images as inputs, which is not similar to how humans learn to recognize actions. I argue that human

action recognition is heavily influenced by movement, and therefore static images, that eliminate movement, reduce recognition ability by omitting a key feature. If you already agree with these conjectures, you may skip this chapter.

Section 2.1 is based on Wolfe's excellent review of visual attention (Wolfe, 2000). Section 2.2.'s primary source is a review by Tsotsos and Rothenstein (Tsotsos and Rothenstein, 2011), and section 2.3 is based on reviews by Otasevic and Aggrawal (Aggarwal and Ryoo, 1986; Otasevic, 2013). Unless otherwise cited, or my opinion is expressed, the information in this chapter should be considered their work.

2.1 Psychological studies on human visual attention

Let us begin with an exercise: stare at the red cross in Figure 2-1, and try to find the white square enclosed by a grey circle. After, doing so, without looking back, try to recall if there was a black circle enclosed by a grey square.

The first target element is located at the top right corner, while the second target is absent. There are two important take aways from this exercise. First, in order to answer these questions, you have to dedicate your attention to each stimulus until you find the target. Second, you are able to pay attention to a target without looking at it directly. Attention is a cognitive process, and while it can be paired with foveal vision, it is not completely bound to it. In fact, it is not bound to any cognitive modality (vision, sensory, auditory, etc.). It is extremely difficult to be attentive to more than one stimulus. Moreover, while you stared at the red cross, about half of your visual processing apparatus is stuck at the cross (Kranz, 2012). The resolution of your visual field drops as the stimuli get farther from the foveal region. The picture is drawn such that you can still recognize the shapes, despite the lower resolution, but I hope you agree that this is a much more arduous task in comparison to the case that you are allowed to search the image freely. Consequently, the task at hand is not only concerned with attention, but also with the variable resolution of images

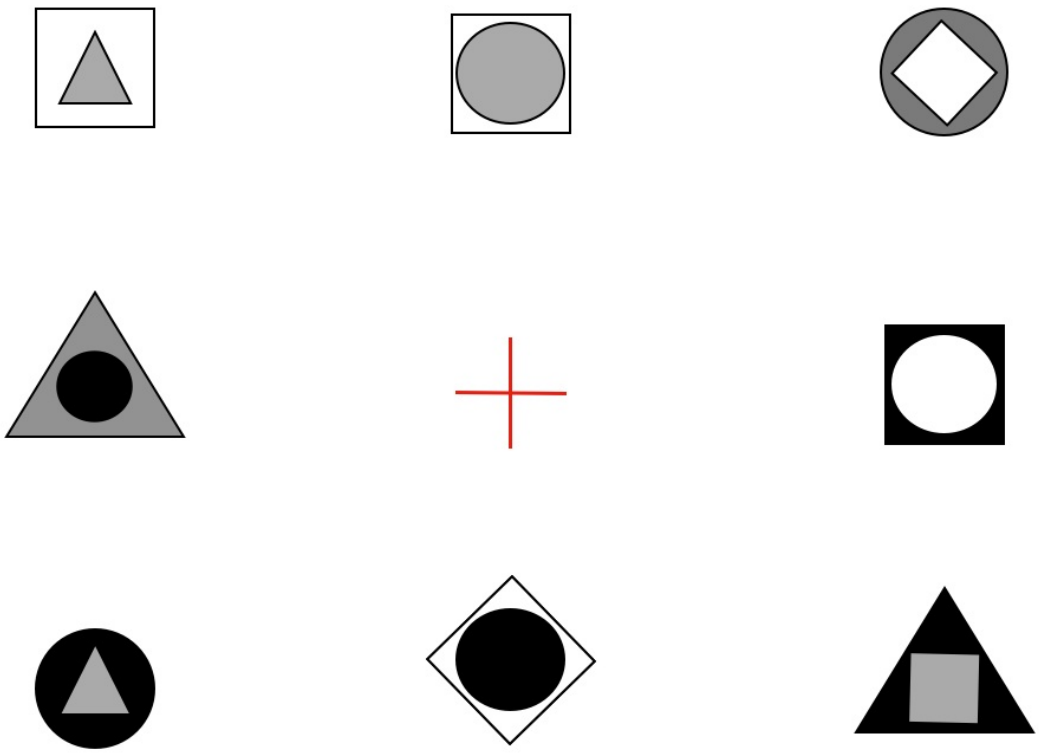

Figure 2-1: Stare at the red cross, and try to answer the questions asked in the text, figure adapted from(Wolfe, 2000).

processed by the human brain. I address these issues separately in my experiments, and provide insights from both categories.

2.1.1 Visual Attention

This section covers what you need to know about visual attention before proceeding to my human experiments. There are four important components to the study of visual attention:

1. **Pre-attentive vision.** Before you really pays any attention to a stimulus in an image, some pre-attentive processing occurs at once over the entire image. Even before you attempted to answer the question that was asked in the beginning of this section, you knew *something* about what the picture consisted of. You

knew to expect geometric shapes, and not rabbits.

2. **Attentive vision.** While attention is dedicated to a particular stimuli, additional information is extracted from the source. To find the target from among the many presented stimuli, you dedicated attention and could infer whether the described target was indeed present.

3. **Post-attentive vision.** Once attention has moved away from a stimulus, is there additional processing dedicated to that stimulus? What if a stimuli is changed?

4. **Attention-less vision.** Similar to pre-attentive vision, this refers to stimuli that you never attend to, either by a cognitive impairment or by simply because the stimulus has disappeared before it was attended.

Most computer vision programs do not include an attention driving mechanism. Hence, pre-attentive or attention-less vision are more analogous to those systems. For this reason, I cover the understanding of pre-attention, and later develop a simple computational approach that models such vision. I then cover the features of attentive and post-attentive vision, that are central to my human experiments.

Pre-attentive Vision

During this stage of vision, the entire visual field is processed at once. This processing is obviously incomplete, as one does not know everything about an image without invoking attention. However, as humans deploy attention rather quickly upon presentation of an image, it is hard to separate pre-attention from attentive vision in humans without attentive impairments.

In order to study pre-attentive vision, researchers generally focus on tasks in which pre-attentive processing can make a difference in the attentive behavior. Hence, the measurements regarding pre-attentive processes are indirect.

One common approach is to ask the subjects to perform a visual search task, where they must find a target item among distractors. The efficiency of the search

task is measured by the subject's reaction time (RT) with respect to the number of distractors present (set size). The key assumption is that effective pre-attentional processing allows quicker-than-random reaction time. This is in fact observed in experiments for particular stimuli. Color, for instance can be used as a pre-attentive feature to guide attention. However, letters are not generally identifiable with pre attentive processes. Finding an 'N' among many 'M's is not any quicker than random search, while finding a red square among many blue ones is very quick, and remains so with increasing set size. This phenomenon is known as the *pop-out effect*. For certain types of features, humans are able to direct attention to the target almost regardless of set size (Figure 2-2). The pop-out effect usually happens when search is performed on a single basic feature, and the set of basic features (color, orientation, size, motion, etc.) is limited. Higher level features, such as closed shapes could also benefit from pre-attentive processing, however, they are thought to be an upshot of pre-attentive processes applied to the basic features. Finally, it is sometimes claimed that faces, or particular states of faces (gaze location, emotions) may be processed pre-attentively. Evidence for this phenomena is controversial, and hence I do not base this study on such assumption.

It is noteworthy that the effect of pre-attentive processing is not necessarily symmetric: namely, if finding a horizontal line among many vertical lines is efficient, it does not provide that the reverse is also true. Finally, as with most searches, determining the absence of a stimulus is harder than determining its presence. On the other hand, if such pre attentive guidance is missing, then search occurs serially over items, thus increasing reaction time as the set size grows.

Pre-attentive vision is not limited to single feature pop-out effect. It can also be used to reduce the domain of more complicated searches, such as *conjunction* search. The efficiency of these searches are worse than the single-feature pop-out effect, but are better than inefficient serial search. An example of such search is if a subject is searching for a red-P among multiple green and red Bs.

For consistency with the literature, it is worth mentioning that some sources (Hulleman, 2013)refer to searches guided by pre-attentive vision as *easy* and those that

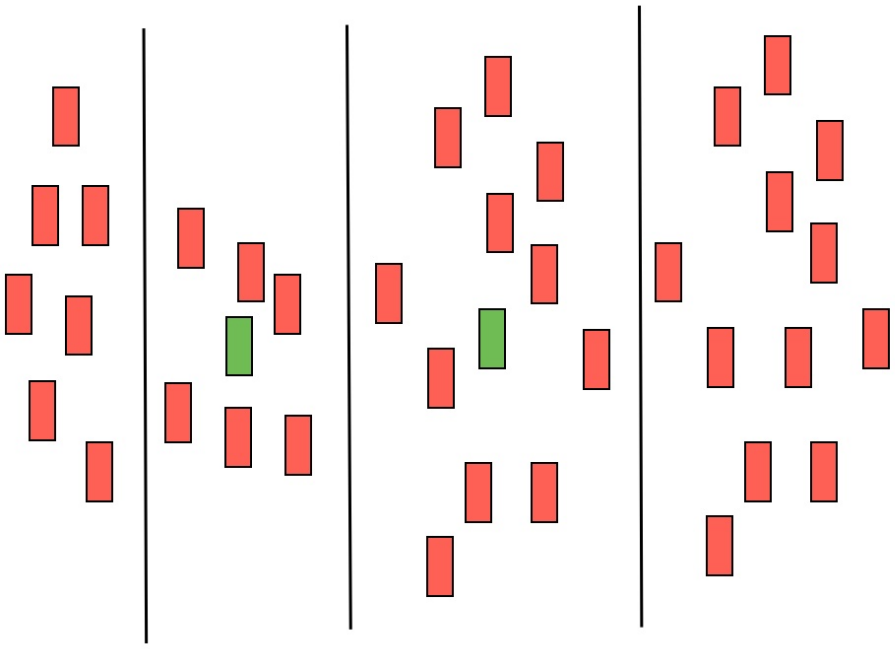

Figure 2-2: Pop-out effect: For each of the four groups, is a green rectangle present? Despite the larger set size for the groups on the right, the reaction time typically stays similar to that of the groups on the left.

require a serial process of investigating samples as *difficult*. Also in some literature (Schneider and Shiffrin, 1977), when reaction time is independent of sample size, the search is labelled as *automatic* while inefficient search is known as *controlled*.

Attentive Vision

In order to understand, and model any behavioral phenomenon, one should understand the extent and limitations of that phenomenon. In order to understand what attention can, and cannot do, I expand on this part of the background more thoroughly. There are two models of how attention is deployed across the visual field.

- **The spotlight model**: One can think of this model as a moving mental spotlight across the visual field. Stimuli that fall under the spotlight are attended, the rest are unattended. Note that this is not the same as the foveal spotlight, which is fixed at the center of gaze. For something to be foveated, one needs

to look at it. However, one can attend to a particular region in the visual field without looking there directly (although it is difficult). Though these models work well for a variety of tasks, like serial object recognition, they are not as successful in explaining subject behavior when multiple stimuli are attended at once.

- **The zoom-lens model**: This is really an extension of the spot-light model. In this model, the spotlight exists, but it can be adjusted in size. In this setup, one can use all of their attention apparatus to a small area, or they can distribute it over the entire visual field. It could be thought of as a zoom-lens, where either everything is in focus but with little detail, or there is one point of focus with a lot of detail.

These hypotheses are not mutually exclusive, but the extent to which attention behaves in one particular way is still to be understood. There is also an ongoing debate on whether attending to many objects at the same time is possible, and if so, to what extent.

Attention is thought to be an *enabling* process. For example, we could think of attention as a process that enables our face recognition apparatus to study one face at a time. Fine perception is impossible without attention, thus foveal vision and attention are almost always coupled, unless they are explicitly de-coupled to study vision without attention (which is done later in this study). The enabling role of attention is applied in a variety of ways, which are explained below.

- **Stimulus selection in space**: Most of a human's visual processes cannot be applied to every stimulus present in the visual field at once. There is a selective process by which stimuli are attended (and foveated), and passed to the other processes with limited capability, e.g. face recognition, serially. It is clear why this is a nice way of addressing face recognition. If one is presented with many faces at once, and tries to recognize each face by attending to eyes, nose or lips (or a combination of these) , one also has to keep track of which eye belongs to which face and so on. Analyzing faces serially, especially in the type of

encounters that occur regularly in daily life, seems to be the better approach. This process is not perfect however, and especially in spatially adjacent stimuli, some interference is observed (Miller, 1991).

- **Stimulus selection in time**: There is also attentional selection in time. *Inhibition of return*, a property that is popular in computational models, is a form of attentional selection in time. Attended stimuli are less likely to be re-attended in a brief period after the attentional processing. This is a useful mechanism, especially in order to search the scene for interesting objects effectively. The second form is known as the *attentional blink*. In attentional blink experiments, subjects search for a particular stimulus within a stream of presented stimuli. While subjects can easily report the absence of the target or appearance of one, they fail to spot a second target if it is immediately presented (200-500 milliseconds) after the first target.

- **Vigilance**: In the presence of multiple events occurring, and with limited processing ability, attention can provide additional vigilance. In order to detect important, potentially life-threatening, events deploying the visual processing uniformly across the entire visual field may not be sufficient. It would be more informative to process a set of 'interesting' spots in more detail. This process is both bottom-up and top-down. There is evidence that any abrupt changes in the visual field will capture attention, unless it is inhibited by top-down processes.

- **Reduction of uncertainty**: This is an implicit point about attention. Knowing something about the target a priori can help the subject find it more effectively. If a subject is asked to find the "odd" object from several samples, they would present shorter reaction time if the type of oddness is also specified (e.g. color, orientation,etc).

- **Enhancement of the signal**: Attention seems to improve the quality of visual signals, both on a neural and behavioral level. Discrimination of fine signals is

only possible when the stimulus is directly attended.

- **Efficiency of processing**: There is evidence that suggests attention increases the speed of processing. In an intelligent experiment, Shimojo et al. flashed a line on the screen and asked the subjects whether the line is drawn from the left or from the right, while in fact the line was drawn at once (Shimojo et al., 1992). Subjects reported each side equally. When subjects were cued to attend to one end, however, they overwhelmingly reported that the line was drawn from that side. This suggests that the attended end of the line has reached awareness more quickly.

- **Modulation of visual processing**: As seen in the previous point, attention can change the perception of reality. This is not only limited to the line motion effects, but also to adaptation to stimuli. While attention is not required for adaptation, it can modulate it. For example, in the presence of several adapting stimuli, attending to one can change how that particular stimulus is perceived and adapted to.

- **Feature binding**: Triesman and Gelade suggest that features (color, orientation,etc) are only approximately bound to the region that they exist within before they are attended (Treisman and Gelade, 1980). However, the precise binding between the feature and the object, is believed to happen after the object has been attended. The research on generation illusory conjunctions (where features are bound to the wrong objects) provide more support, though not a proof, for this theory(Prinzmetal et al., 1995).

Post-attentive Vision

There is some evidence that previously attended stimuli, if not memorized explicitly, cannot increase search efficiency any more than pre-attentive processing. In repeated search studies, subjects were asked to find a target object from stimuli similar to those in Figure 2-1. If the target is not there, presumably, the subject must have attended to every object before responding "no". However, if the subject is then asked again

to search for a different target within the same image, the search time does not seem to be improved, even though the stimuli have all been attended. One must be careful in such studies for artifacts, because even if improved reaction time is observed, it could be due to memory search, rather than visual search.

Another upshot of the lack of persistent consequences for attended stimuli is illustrated by the phenomenon known as *change blindness*. In the lack of sudden and dramatic visual transients that attract attention, change in unattended or previously attended stimuli is often not perceived.

Vision without Attention

There is some debate in the literature about what type of stimuli are truly unattended. Most studies in this domain jam the attention apparatus with some task and then query the subjects for the unattended stimulus. While some studies show that subjects can recall some basic properties of the unattended, others challenge this view by using more elaborate tricks, such as attentional blink. For the purposes of this study, we adhere to the categorization provided by Wolfe (Wolfe, 2000). Wolfe categorizes the stimuli present in the visual field into three different groups: (i)The explicitly attended group, where objects are recognized and thoroughly processed, (ii) Stimuli that have never been attended explicitly, but leave some impression (through pre-attentive processing) of their basic features on the observer, and (iii) the stimuli that are never attended and do not leave any recallable mark of their features on the observer.

2.1.2 Foveal and Peripheral Vision

The highest density of light receptors is located near the center of the retina (within a highly pigmented structure known as the macula). As it can be seen in figure 2-3, which are adapted from (Hunziker, 2006; Wandell, 1995), the highest acuity of human vision corresponds to the location where the density of receptors is the highest. However, this only comprises about 1% of the total area in the visual field. A tangible approximation of the size of this area would be twice one's thumbnail width held at

arm's distance from the eye. Astonishingly, about half of the visual cortex is dedicated to the analysis of the signals in this small region.

This non-uniform resource allocation has obvious implications for vision. Most importantly, humans use their foveal vision, coupled with visual attention, for visually demanding tasks that involve recognition. Reading is a simple example. While reading often occurs with well-separated, high contrast stimuli (black ink on white paper), it is still difficult to read when foveal acuity is missing. Action recognition is even more challenging, because the medium is often wildly volatile, and actions are generally not recognized in a static manner. Therefore, it is reasonable to assume that foveal vision must be deployed to locations of some importance. Generally, this goes hand in hand with attention.

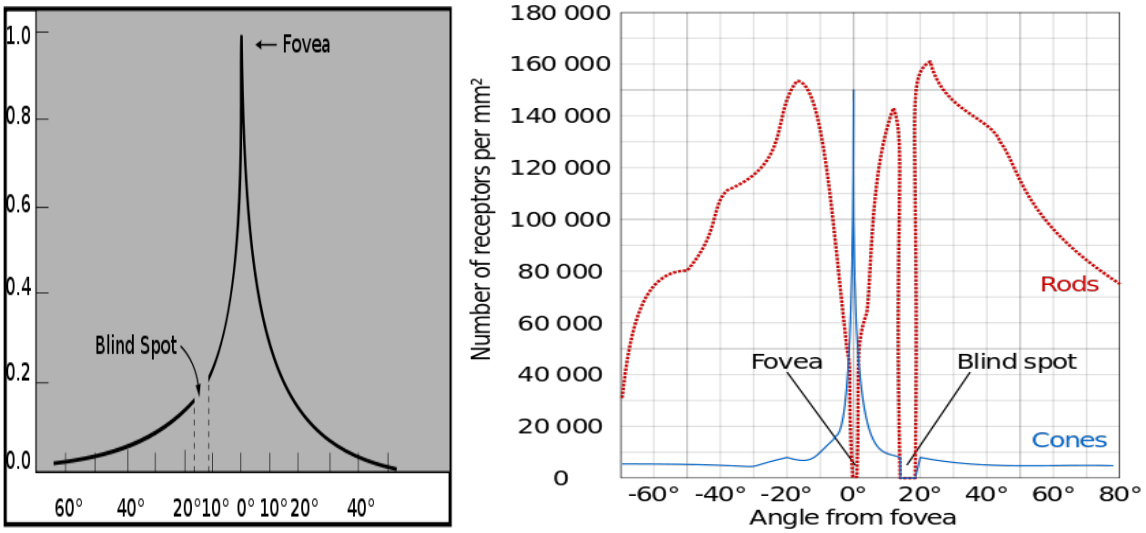

Figure 2-3: Left: The acuity of human vision across the visual field, normalized by foveal acuity (taken from (Hunziker, 2006)). Right: The number of receptors present across the retina. Note that cones, which are the high acuity receptors, are heavily concentrated in the fovea(taken from (Wandell, 1995)).

The resolution of the human visual field drops as an object falls farther from the foveal region. This is not hard to implement, as the algorithm would simply treat the input data with less precision as it falls outside foveal vision. However, it is non-trivial to find out where to deploy the foveal apparatus computationally. Moreover, it is also not clear how much data humans collect from peripheral vision, especially when it concerns action recognition.

2.2 Computational approaches to modeling attention

As discussed in the previous section, attention is a natural algorithm that reduces the analysis space required to infer events and find objects of interest. Though models can be descriptive of a natural algorithm in many ways, including mathematical and non-mathematical descriptions, I am particularly interested in models that are implementable, at least to some extent. More concretely, I am interested in reducing the search space for models that can be formally specified, algorithmically approached and implemented. This is in line with Marr's (Marr, 1982) opinions on the approach to creating systems capable of performing task similar to those that humans can perform.

2.2.1 Descriptive representations

Using attention as a possible mitigator of the information load has been proposed as early as 1972 (Uhr, 1972). Since, many models have been proposed in computer vision. They operate on four themes, that describe how attention would be defined in a computer vision, suggested by Tsotsos and Rothenstein (Tsotsos and Rothenstein, 2011):

- **Interest point operations**: After some shallow parallel processing, the analysis space is narrowed down to points of interest. These could be points of movement, action or objects. The points of interest are then analyzed serially.
- **Perceptual organization**: While processing an image pixel by pixel may be arduous, it is not impossible to achieve this using modern computers. This does not draw the entire picture however, as the composition, distribution of object and many combinatorial relations that may exist between every two objects in a scene also adds to the complexity. How such perception is organized is another feature of attentive vision. We saw the analogue of this problem in the psychological overview in humans.

- **Active vision**: Humans do not look at their surroundings statically. Instead, they actively move their eyes and head in order to discover events and follow objects. Some computer vision systems emulate this property, either by moving the camera, or by software tricks.

- **Predictive models**: Tsotsos (Tsotsos, 1989) formally demonstrated that the intuition that task-specific knowledge can reduce the processing domain. Such information can then be used in face and motion recognition.

2.2.2 Algorithmic models of attention

The four general themes presented in the last section define attention in a manner that can be implemented computationally. Over the past three decades, many models of computational attention have been presented. While these models draw from the themes presented above, from an algorithmic perspective, they implement one or several of the following mechanisms that are suggested as a possible parallels (either algorithmically or functionally) to human attention.

- **Selective routing**: In this mechanism, higher layer neurons (or circuits) modulate the lower lever elements with regards to which signals will pass and which signals will be suppressed on a lower level. This creates several 'attentional beams' with different signal intensities, which are ultimately selected from. The STBind model by Tsotos et al (Tsotsos et al., 2008) and SCAN model by Postma et al. (Postma et al., 1997) are largely influenced by this mechanism.

- **Saliency maps**: These mechanisms are among the most widely studied, and still a subject of intense research due to their relative success at producing bottom-up attentive behavior. This mechanism is comprised of five stages:

    1. The entire image is first processed in parallel, generating a raw representation of the image based on basic features.

    2. A saliency map is created based on the each of the features existing in each location in the image.

3. Each location is then mapped into a non-topological representation of the saliency.

4. Some selection mechanism operates on this representation, generating a ranking of conspicuity, and then implementing a winner-takes-all approach. Attention is then dedicated to that locale.

5. This selected location is then inhibited after a brief period, and attention is directed to the second most conspicuous location.

As noted before, this domain of mechanisms is still actively researched. It is not clear if the brain produces such saliency maps, and if so, which one is produced. The attentional pattern presented in Figure 1-1 was generated by the Walther et al. implementation of this mechanism (Walther and Koch, 2006). In Figure 2-4, I demonstrate the intermediate saliency maps for the analysis.

- **Temporal tagging**: This mechanisms suggest that attention is allocated to neurons that present 'synchronized' oscillations in response to a stimulus, as opposed to the neuronal groups that are fired with less synchrony. These models received more recent biological reinforcement, as Bauer (Bauer et al., 2009) and Neibur (Niebur, 2009) suggest. The HumBie model by Hummel and Biederman (J. E. Hummel, 1992) is an early implementation of such models. Later, this was incorporated into other models that draw influence from multiple categories, for instance in the AIM model by Bruce and Tsotsos (Bruce and Tsotsos, 2009).

- **Emergent attention**: These models are somewhat similar to selective routing, except that there is no explicit selection process. Instead, attentional beams compete with each other and according to some top-down bias may 'win' the loci of attention. The SAIM model by Heinke and Humphreys (Heinke and Humphreys, 2003) is largely influenced by this paradigm.

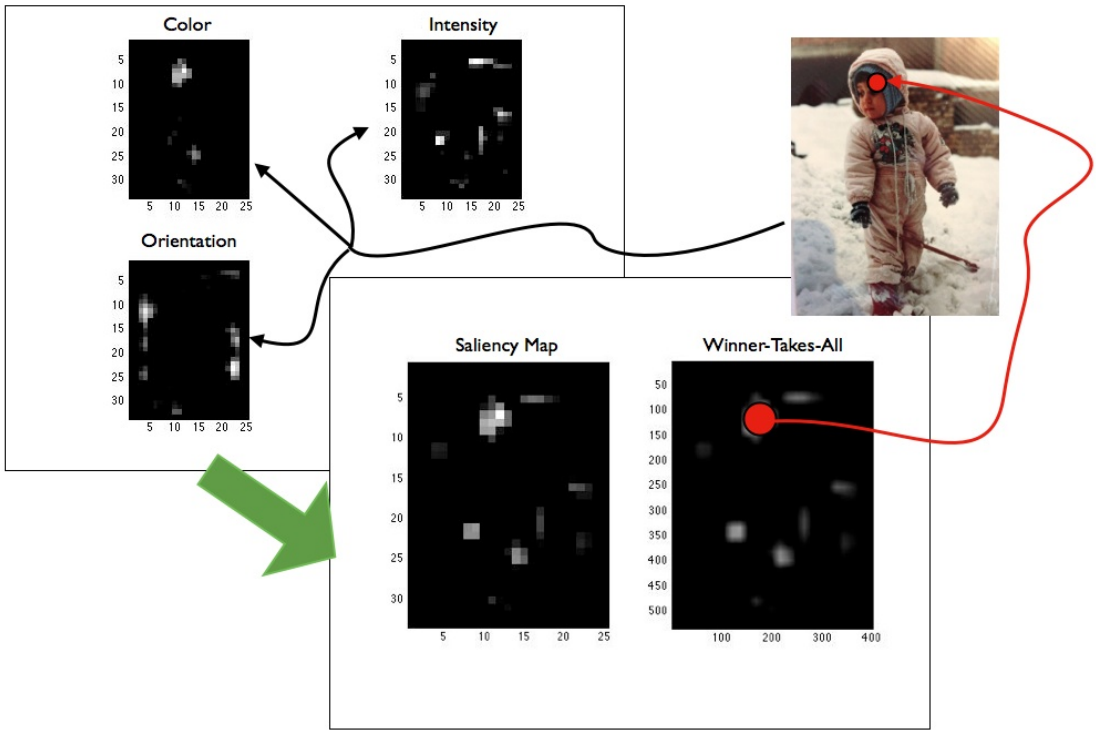

Figure 2-4: Running the saliency map algorithm, as implemented by Walther et al. (Walther and Koch, 2006). The next step (not shown) involves inhibiting the winning location such that attention is directed towards a new locus.

2.2.3 What is missing?

It is also noteworthy that most of the human studies performed to inform the computational models have been done in the domain of static images (Tatler et al., 2011, 2006), where subjects would look at images and their gaze is tracked. Thus the dominant modeling framework relies on bottom-up control of attention by finding salient features, in which the task driven properties of vision are not present(Tatler et al., 2011). Such salient features include pre-attentive capture of pop-out features, and could be modeled using salience maps based on luminance, color, or orientation of the object. These models have some prediction power, but they do not perform impressively better than a model that always predicts the gaze point at the center of the screen. Using object-level information has significantly improved the performance of

these models (Tatler et al., 2011). Salience models usually have a set of accompanying assumptions and shortcomings:

- They assume that pre-attentive features drive selection and there is a default, bottom-up mode of selection
- They do not account for the variable resolution of the visual field
- They mostly depend on spatial selection, and temporal information such as fixation duration and order of fixations are not incorporated
- They follow visual saccades rather than the location of processing as a whole, sometimes the item of interest is not exactly at the fixation point but close by.

There has been alternative accounts to address these shortcomings. For example introducing peripheral acuity limits (Wischnewski et al., 1973), modeling fixation durations (Nuthmann et al., 2010) and using top-down factors(Ehinger et al., 2009; Baluch and Itti, 2011). Rao (Rao, 1998) introduces a model that makes direct use of Ullman's visual routines, which are both top-down and bottom-up(Ullman, 1984). Most of this work is still related to static image viewing however, hence the niche for a dynamic video or live action tracking and the corresponding computational modeling of visual attention is nearly unoccupied. A study by Tom Foulsham et al (Foulsham et al., 2011) is the closest to such efforts. In that study, subjects would go out for a 15 minute walk to buy coffee while wearing an eye-tracker similar to PUPIL. Their gaze was recorded, and later the subjects would look at the video of the walk, and their gaze was tracked again. The team then reported the number of times and the duration that each subject would look at a particular object. One observation of this study was that people tend to turn their head in the live scene, which is the more natural behavior, rather than moving their eye around. The study does not go into any action study or modeling of this data.

Hence, there is definitely a large area of study on the effect of foveal vision and attention on action recognition. The understanding of attention is incomplete, and computational models in this domain are promising but far from perfect. Nonetheless,

the advantages that attentive vision offers in the domain of action recognition is clear. This is where attention can be used, even with limited understanding, to improve our computational approaches towards action recognition. That is addressed in the next chapters, however, before we proceed, it is useful to have a brief overview of the current work on computational approaches to action recognition.

2.3 Computational approaches to action recognition

Action recognition has been a topic of interest in computer science for years. Many systems have been developed to understand a variety of actions, mostly human actions. They range from basic, common, actions like walking, to very complex actions such as braiding hair. The systems developed to recognize human actions rely on some categorization of lower level primitives. For instance, Aggrawal and Ryo use gestures as action primitives(Aggarwal and Ryoo, 1986), while Otasevic uses poses (Otasevic, 2013). Actions are then defined as a composite of these poses or gestures. Complex actions are then defined as those than entail more than one action or subject. While these are the general themes of how actions are broken down for recognition, several approaches are used in single-level systems in order to classify actions:

- **Space-time approaches**: In these approaches the actions are characterized by concatenation of frames within a time dimension. Hence an action is described as stream of 2D images arranged on a time dimension. These approaches make use of space-time volumes of silhouettes, trajectories of points of interest, a 3D volume of interesting features, or a combination of these. Figure 2-5, taken from (Sun et al., 2011), is an example of the tracked interested points in a space-time volume.

- **Sequential approaches**: These approaches are based on time-sequential methods that deal with the sequence of events, poses, or observation vectors in any given window. Otasevic (Otasevic, 2013), and Yamato et al. (Yamato et al.,

1992), approach the problem in this manner, and both use Hidden Markov Models (HMM) to extract actions from a sequence of representations.

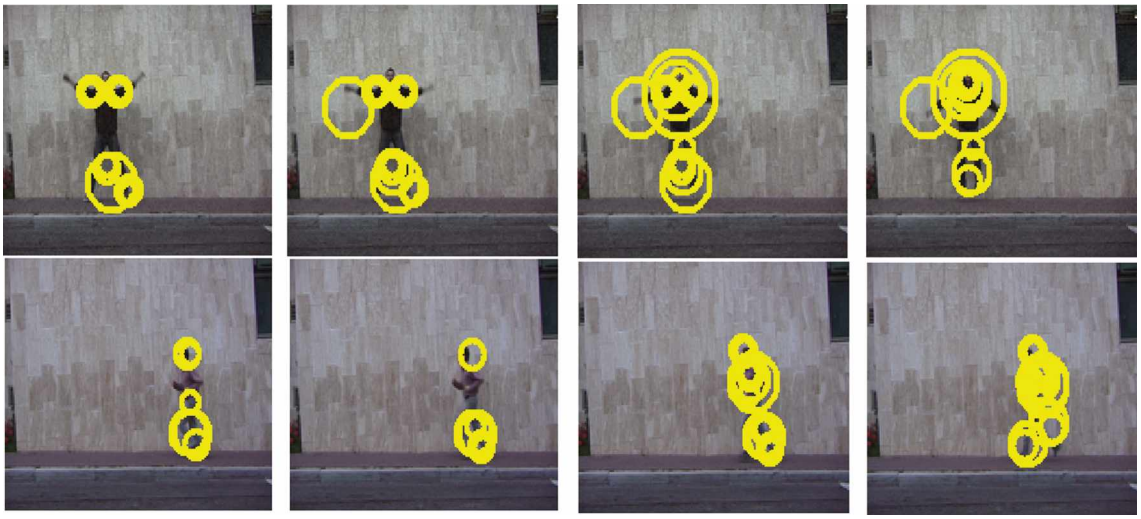

Figure 2-5: Space-time interest points for a person doing jumping jacks (above), and jogging (below), taken from Sun et al (Sun et al., 2011)

While these approaches can be used by themselves to recognize simple actions, more complex actions are usually approached by augmenting these systems with hierarchical structures. Actions such as fighting include more than one person performing multiple basic actions such as walking, kicking, punching, etc. Higher level recognition systems are used to extract such complex actions from the basic sequence of lower-level actions.

Computational models of action recognition may start from annotated images where people and objects are labeled, or they may build their own people detection mechanisms. In both of these cases, some global or local image representations are used. The choice of this representation, is a significant factor in how the mechanism as a whole performs. In my study, I hope to elucidate aspects of human vision that can improve our image representation and feature selection for action recognition.

2.4 Conclusion

After reading this chapter, I hope you are convinced that attention can offer a great computational advantage in vision. While the amount of progress in both the psychological studies and computational modeling has been immense, there are a few short-comings that I try addressing, although in part, in my thesis.

First, I am primarily interested in how to improve action recognition. Most of the psychological studies are conducted to understand the underlying principles of attention. This is the ultimate goal of the study of attention in psychology, but certainly beyond the scope of this thesis. For this reason, most of the studies on attention have to do with more basic roles of attention, such as recognition of target letters among distractors, or more natural experiments that involve visual search for target faces or objects. Hence, the body of knowledge on how attention is directed in particular to recognize actions is scarce. Therefore, with a few of the experiments in the following chapter, I hope to add to this particular domain of understanding.

Second, many studies and models address the case of static images. The subject, often head-fixed, stares at an image for a period that is long enough for the subject to direct attention to multiple locations. While this simplifies the modeling and doing experiments, it is not how humans have evolved to see. In a static image, the ranking of the 'salient features' is static as well. In a real-life situation, this 'saliency map' needs to be updated very quickly. To address this, I tried using videos or even real-life acts when possible. One important pitfall in watching movies is that movies may be cut, and especially when gaze tracking, this produces an unnatural break in the gaze trajectory. Therefore, all experiments with videos were done with movies that are uncut throughout to avoid this pitfall.

Third, freely observing humans often move their heads, rather than eyes, to follow objects in their visual field, which means the visual field itself is changing dramatically. Further, the environment itself is constantly changing. Thus, when possible, I tried not to restrict subjects in the way they observe images. Of course, this will generate more noisy data, but the purpose of my study was to provide qualitative, rather than

quantitative insight. When noise seemed to affect qualitative results, the data was aborted altogether.

Chapter 3

Is a Visual Agreement a Verbal Agreement?

This chapter includes two experiments that are addressing a basic question: *To detect a verb, do humans look at the same place?*. In this chapter I argue that the answer to this question is likely to be *yes*. I provide suggestive evidence that humans target the same location in an image when they try to answer the same question. For instance, when a human is trying to determine if someone is *catching* a ball in an image or video, she would look at the actor's *hand* and the *ball*, and they would do so consistently across images. Moreover, all humans tend to be in agreement as to where they should look in order to determine the occurrence of an action.

3.1 An eye for action - a pilot study using PUPIL

Initially, I was interested to study how humans observe events and use those observations to recognize actions. In particular, I investigated how humans control and guide their foveal vision, accompanied by attention, when observing actions. Hence, I built a head mounted eye tracking device, PUPIL, based on the work by Patera and Kassner (Kassner and Patera, 2012). This device enabled me to track subject's gaze point while they view a scene. The goal was to understand whether a particular part of a scene is more important than another in recognizing an action, and inspire

better methods of computational action detection through vision.

When a human watches a scene, most of the information that he perceives from the fovea is based on a very small area of his entire vision. Accordingly, it could be argued that in order to recognize an action — *pick up* for instance— it would suffice to analyze a small section of the entire visual field, in order to fully infer the occurrence of the action of interest. Certainly, this line of deduction has two premises:

1. Humans can successfully infer actions in a scene by only processing the foveal region: the premise claims that the visual data acquired from the foveal center suffices for the subject to successfully interpret the action. For the purpose of this research however, a concentric region that contains the foveal vision will satisfy the premise will do, assuming its not so large that it lacks any specificity and the region has a biological correspondent in human vision.

2. All Humans infer similar actions when they look at the same location in an image. This premise suggests that because a majority of adults agree on the reported action, they must find similar parameters in the visual data. *Assuming* the first premise is true, the similarities and differences likewise, must be inferable by the information acquired through the foveal vision(or a reasonably larger superset). Actions such as *pick up*, or *throw* occur over a rather short time frame, hence how the foveal vision is dedicated to find the fundamental characteristics of an action is very important. Finally the word 'similar' is defined to point to conceptual similarities between two scenes. For instance, it could be the case that all observers must look at an actor's hand (or have his hand within a reasonable distance of the foveal field) in order to recognize the action *drop*. Of course, if everyone is observing the same exact scene, it might be the case that they look at same spatial coordinates during the key moments.

In order to construct a vision system based on human vision, it is necessary to establish whether these premises are true, or partially true. While the first premise is intuitive and has been studied before (Chun and Marois, 2002), the second one is less so. The first step in my research was to investigate the second premise. In other

words, I have to demonstrate the validity of this premise based on human experiments. For this purpose, I used PUPIL to record where my subjects look when they observe scenes in which actions of interest are carried out. Intuitively, one could argue that humans must allocate their attention similarly if they infer a similar action, and hence premise two has some support. My experiments confirmed the extent to which this claim is true. It is noteworthy that once these premises are established, they will create a solid foundation based on which a computational model for vision can be constructed. For instance, if every subject looks at the actor's hand while the actor is releasing an object in order to recognize *drop*, then we can create our model such that it has the same property.

Before I proceed, I remind the reader that the problem of guiding visual attention is a different one from the problem of inferring an action based on what is seen. Most of this project will address the inference problem. However, if the collected data suggests a certain model for visual guidance, then modeling the guiding problem would be a reasonable path to take. This cannot be decided a priori, and will materialize further after the data is collected and analyzed.

3.1.1 Experiment set up

I designed a protocol for experiments to be done on my initial trials. Subjects were shown a selected clip of a DARPA-provided video. The video spanned two minutes in which 3 soldiers interacted with two civilians performing actions such a give, throw, carry, etc. The subjects were asked to watch the video on a computer screen or a flatscreen TV twice. Subjects spent about 20 minutes for calibration and familiarity with the set up, and then the experiments began. In the first test, the subject freely observed the video and no prior information was given. This was called the *non-primed* run. Afterwards, the subject observed the video once more, but this time he was asked to call out when one of the verbs — give, take, pick up, put down, throw — occurred. Accordingly, this was called the *primed* run. The goal of this design was to observe if there are obvious differences between the occasions when the subject is freely looking at the scene and when they are actively searching for a particular verb.

Moreover, I was looking for signs of similar behavior, both in eye-movement and what the subjects reported, when looking at the same scene, after different instructions.

3.1.2 Data collection and processing

Five volunteer subjects were used for the experiments. The gaze-tracking data that was further processed as shown in Figure 3-1 to map subject gaze points into a unified framework. This was done in order to avoid restricting the subject's head, and find a unifying frame of reference between trails and subjects.

Although this data is not sufficient to draw conclusions, it provides a suggestive evidence that there is some similarity between subjects' behavior. Further data collection was hindered due to lack of volunteers and the unreliable performance of the eye-tracker (as it was in the early development stage) for maintaining calibration over two minutes. Furthermore, because there was no recording time marker/surface detection system, which is usually provided with eye-tracking software, I implemented a post-collection mechanism shown in Figure 3-1. However, the process required extensive noise canceling for each case, and therefore was hard to automate. In the next section, I have performed this noise correction for two subjects across two trials and compare them as a evidence. I present the data hoping that further studies would corroborate these observations. This would be a great place for further investigation, if one is able to use a more reliable version of the eye-tracker, especially if marker/surface tracking is in-built.

3.1.3 Results

For all subjects, the non-primed case of free-observation contained more variation than the primed case (see Table 3-1). Should this be confirmed with more data, this observation is interesting, because it is not obvious that free looking subjects should be moving their gaze more often. The likely explanation for this is that in the primed case, the subjects knew where to look for the particular task of interest (as they have seen the video already).

Figure 3-1: A general demonstration of how the remapping algorithm works for each frame A: a single frame with superimposed gaze position (red circle) B: The edge detection mechanism finds the monitor screen, and calculates the distance of each edge to the global view(blue dotted arrows), the green arrows are the coordinates of the gaze according to the world view and are already recorded during the experiment. C: based on the information calculated in B, one can derive the coordinates for the new frame D: a picture of the gaze point remapped to the video, after this stage we can superimpose gaze-points from multiple experiments and track similarities

For the rest of this section I will focus on the results from two people, each undergoing two trials. These results were chosen based on the quality of calibration when recording. I then have performed extensive processing in order to match the frames across subjects and trials reliably, using the approach that was demonstrated in Figure 3-1.

In Figure 3-2, I demonstrate the difference between the primed and non-primed trial for one subject. The other subjects demonstrate the same trend. After a period

$\frac{Var(x_{np})}{Var(x_p)}$	1.26
$\frac{Var(y_{np})}{Var(y_p)}$	3.22

Table 3.1: Ratio of variance measure for all six subjects across 15000 data points. Subjects tend to look around more freely in the y-axis when not primed. Most of the events occur in different locations along the x-axis, hence the difference in variance for the two cases is less clear. $x_p : x_{primed}$, $x_{np} : x_{non-primed}$, $y_p : y_{primed}$, $y_n p : y_{non-primed}$.

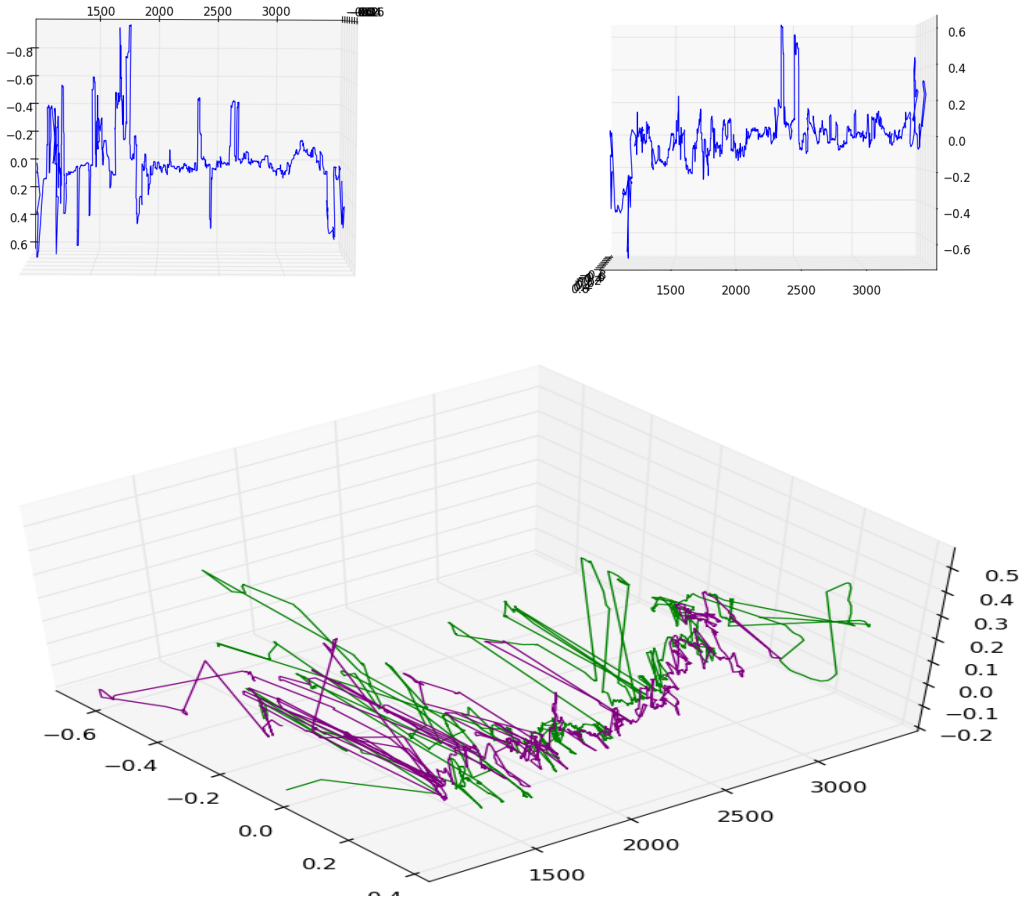

Figure 3-2: Difference between freely looking subject and the same primed subject. Top left: distance between the primed and non-primed gaze point on y-axis. Top right: distance between the primed and non-primed gaze point on x-axis. Primed (purple) and non-primed (green) trajectories superimposed. The decimally increasing axes correspond to normalized distance from the center. The integer axis corresponds to frames.

of highly wide-ranged gaze movements, as observable in Figure 3-3, there are intervals of strong overlaps between frames 2000-3200. During these frames, up to 4 actors interact with each other, which is the main event within the scene. Subjects seem

the be drawn to those events regardless of whether they are primed to look for them or not. During the primed trial the subject seems to stay more concentrated on the group of people with less frequent deviations. This is likely due to the fact that the subjects did not want to miss detecting a verb. Hence, it appears that most of the important events, human actions in this case, are well tracked by the observer even without advanced priming.

I stated in premise 2 of this chapter that humans may be looking at similar locations of a scene to infer similar actions. This experiment was set up to confirm this premise. The current data is not sufficient, but across all six subjects (two best calibrated cases shown here for clarity), the trajectories seem to agree at critical intervals, especially in the primed case. Therefore my current observations seem to support premise 2. These observations are shown in Figures 3-3 and 3-4. In the

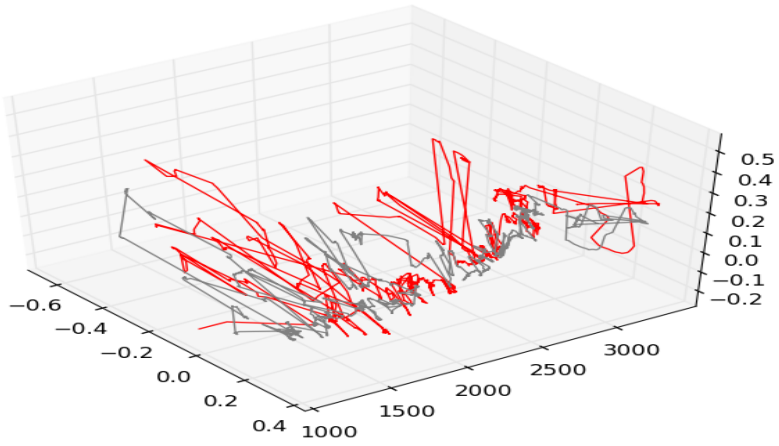

Figure 3-3: Freely-observing eye trajectories across 2500 frames, for two subjects colored red and gray. The decimally increasing axes correspond to normalized distance from the center. The integer axis corresponds to frames.

primed case, (Figure 3-4), the two subject's gaze location converges when two people enter from the corner of the screen. While the two people walk in parallel, the two subjects seem to track different targets. In this case, both subjects look away less frequently than in the free-observing case. On all but one occasion all six subjects were able to detect all verbs of interest. One example is given, in Figure 3-4, where

both subjects look at the area where give occurs and successfully report it. One one occasion a subject missed a *pick up* when he was primed for it (red subject in Figure 3-4). My gaze-tracking data showed that the subject was looking at a different part of the scene in that case, this event hints towards the confirmation of premise 1.

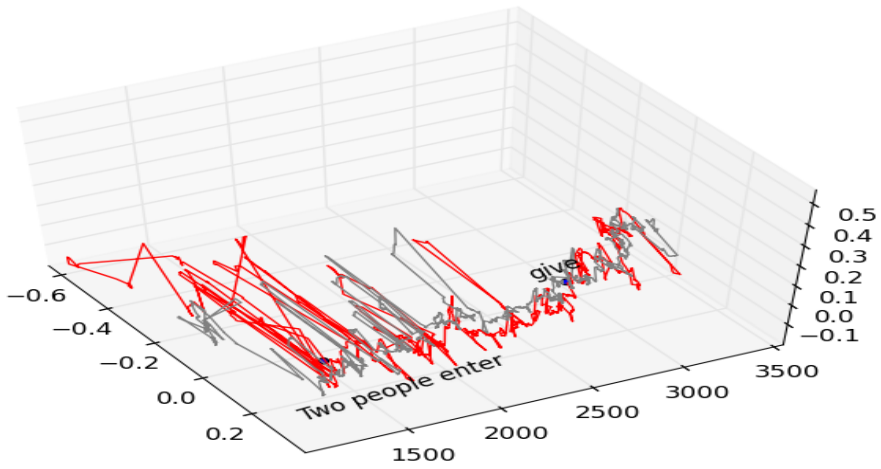

Figure 3-4: Primed subject's eye trajectories across 2500 frames, for two subjects colored red and gray corresponding to figure 3-3. Note that both subjects seem to follow two people that enter the screen from a corner and walking in parallel, however, they are mainly focusing on different people, hence the parallel lines. An example location and occurrence of a target verb , *give*, is also demonstrated. Both subjects detected this event. The deviation seen at the end of frame 3500 for both subjects corresponds to a *throw*. The decimally increasing axes correspond to normalized distance from the center. The integer axis corresponds to frames.

This data manifests an important characteristic of videos. In videos, our saccades do not land on different locations of the same frame. Instead, our gaze is directed towards different frames of the video, because frames are dynamically changing to each other. The gaze path is not a trajectory through 2D space. It is a trajectory through many frames, that change constantly. We do not infer verbs from one frame alone. For many actions, such as throw, the total time available to recognize the action is quite short: enough to allow for 2-3 saccades at most. It is therefore reasonable to assume that very few gaze points, along with a vague idea of the surroundings, may suffice to infer an action. Moreover, in this aspect videos are also closer to the real

life scenes. I investigate this further in the next experiment.

3.2 A one-shot gaze point Q and A

The short pilot study described above provided me with interesting information regarding the similarities of the gaze trajectories. Unfortunately, for several reasons including the low reliability of the eye-tracker, it was extremely difficult to scale that experiment up to collect more data. Therefore, I designed a scalable experiment that was cheaper and more effective, and hence allowed scaling up to many subjects. The price I had to pay, however, was using static images instead of videos. The main question in this study is similar to the one I introduced in the beginning of this chapter: In order to recognize whether an action is occurring, do we all look at the same place? Does the location we look at affect our judgement? I seek to provide insights into these questions further with this study conducted on Amazon's Mechanical Turk.

3.2.1 Experiment set up

Subjects were instructed to read a question presented at the top of the computer screen. Questions asked whether a person is performing a particular verb. An example would be: *Is someone dropping the ball?*. When subjects were ready, they would click at the center of the testing window, and they would be shown a picture for 2 seconds. The picture would not appear unless the subject clicked within 4 pixels of the middle of the testing window. This was required to eliminate the biases introduced from different starting position of the mouse. Subjects were instructed to click on the first location they looked at, in order to answer the question that had been posed. Once the subject clicked on the location of gaze, the picture would disappear and a dialog box would ask the subject whether the verb was occurring or not. Of course the assumption here is that subjects do click in the same place that they look at more often than not.

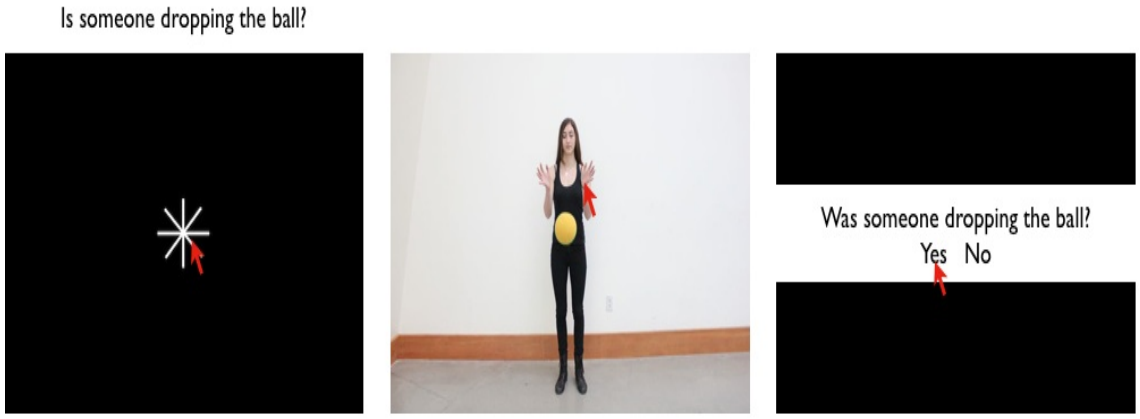

Figure 3-5: The steps of experiment from left to right: (i) reading the question and clicking at the center of the star, (ii) clicking at the first point of gaze, in order to determine the answer to the question, (iii) answering the question.

3.2.2 Data collection

Subjects were recruited on Amazon Mechanical Turk, as well as volunteers. A total of 70 unique subjects successfully performed the task. Each subject completed 35 questions by the end of the task.

3.2.3 Results

Out of 35 questions, 32 distinct pictures were displayed, so that each subject answered 3 random questions twice. This repetition was included to measure consistency. On average, the distance between the first and second click on the same picture by the same subject was small, as shown in Table 3-2, with more than two-thirds of the repeated hits landing within 25 pixels of the original hit.

mean distance in pixels	33
median distance in pixels	16

Table 3.2: The size of the display box was 563 by 375 pixels. The mean and median distance between two hits on the same picture, by the same person (observed among ~ 200 points) suggest good consistency of hits when repeated.

Once I established that people tend to be consistent with themselves when report-

ing the location of gaze, I proceeded to study the consistency across the population. As showing all of the pictures is beyond the scope of this chapter, I will display a few in this chapter and provide a few more in appendix B.

Notwithstanding a few borderline cases, the majority of questions asked were simple to answer. This is also evident by the strong majority agreement of the color of the clicks ('yes' responses are marked as blue, and 'no' is marked red). While observing figures 3-6 through 3-10, note that a majority of subjects click at a certain object, rather than exact location, unless these two properties coincide. For example in figure 3-6, when subjects were asked whether the person is touching the box, people either click at the hand, or the box. While a majority of them do click at the important location (the point of touch), almost all of the click on either one of these two objects. Because these locations are close, it is reasonable to assume that one could look at the middle of the box, and still answer the question correctly, even though it reduces accuracy. The case in Figure 3-7 is different. Subjects now seem to be looking at two distinct locations when the ball is far from the hand. It is almost certain that the subjects will saccade to both the hand and the ball and then choose one of the two to click on. This becomes much clearer in the second picture of Figure 3-7, where the ball and the hand are at distant locations, and the click are distributed almost equally between them.

In Figure 3-8, I demonstrate another point about the gaze, anticipation. The subjects try to choose between two actors, and find out whether one is catching the ball. While it is hard to distinguish a catch from a throw in a still image, the subjects tend to recognize these events well. Moreover, many of the clicks tend the be where the subjects *expect* to see an action. By looking at the second image in Figure 3-8, we can observe that most subjects consider the *absence* of a pose that indicates an action by the female actress as evidence for the lack of action, even though the male actor is performing a task that could be otherwise classified as catch (and it is by a few subjects).

In Figure 3-9, the human's actions are almost the same in both cases. Once again, both the hand, the box draw attention. However, in order to distinguish the two cases,

many subjects noticed the state of the thread that was connecting the two.

Finally, it is interesting to look at the results obtained in Figure 3-10. While smelling is not necessarily an easily observable action, subjects tend to strongly agree about which picture involves smelling and which one does not. While most subjects click on the nose to tell the two cases apart, the distinguishing factor appears to be whether the actress kept her eyes open. This strongly indicates the presence of a 'mental' driver of attention in this case, as opposed to a 'visual' driver.

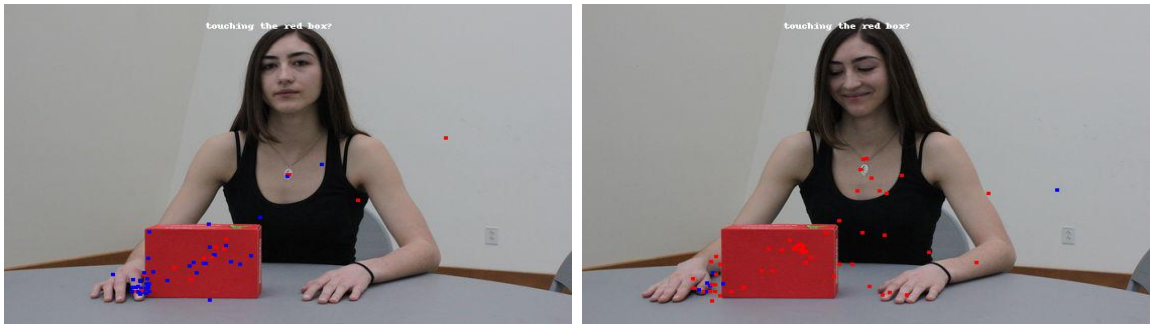

Figure 3-6: Question asked: *Is someone touching the box?* Each point indicates a click by the subject. Red points indicate that the subject has answered 'No' to the question, while blue dots indicate that the subject has answered 'Yes'.

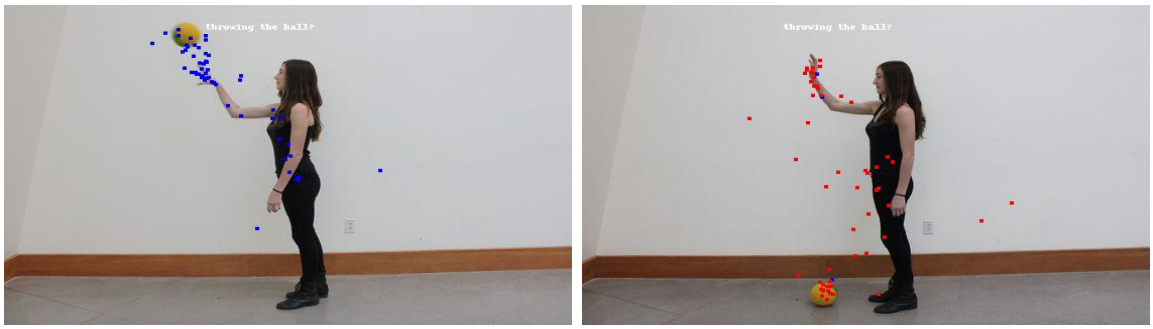

Figure 3-7: Question asked: *Is someone throwing the ball?* Each point indicates a click by the subject. Red points indicate that the subject has answered 'No' to the question, while blue dots indicate that the subject has answered 'Yes'.

As we can see in all of these pictures, there is a great amount of consistency among people and the location of gaze seems to be significant in the subject's answer. The promise of these results is apparent, and these experiments could be repeated in laboratory, with a reliable eye-tracker, which eliminates the need for clicking. This

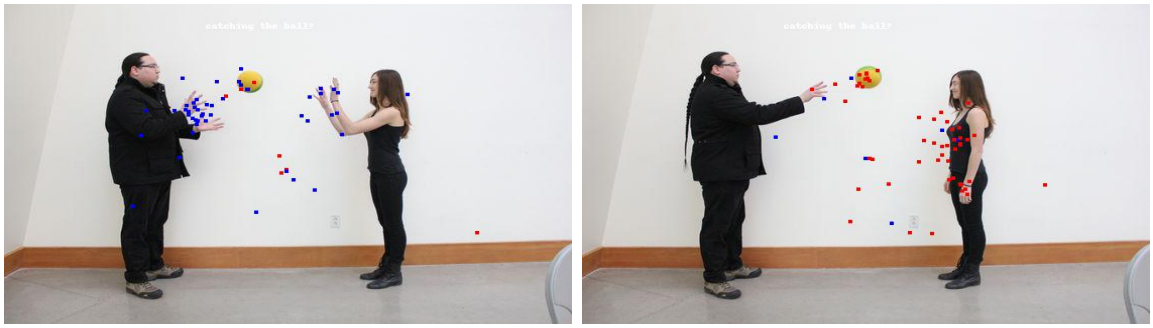

Figure 3-8: Question asked: *Is someone catching the ball?* Each point indicates a click by the subject. Red points indicate that the subject has answered 'No' to the question, while blue dots indicate that the subject has answered 'Yes'.

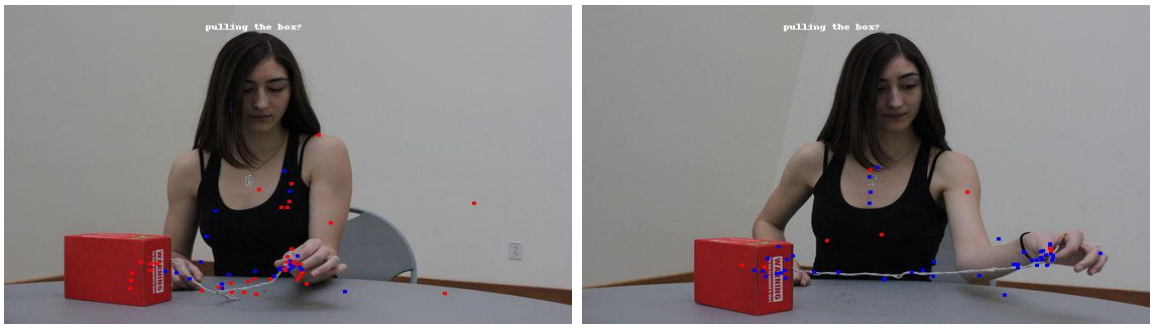

Figure 3-9: Question asked: *Is someone pulling the red box?* Each point indicates a click by the subject. Red points indicate that the subject has answered 'No' to the question, while blue dots indicate that the subject has answered 'Yes'.

Figure 3-10: Question asked: *Is someone smelling flowers?* Each point indicates a click by the subject. Red points indicate that the subject has answered 'No' to the question, while blue dots indicate that the subject has answered 'Yes'.

would allow for the pictures to be flashed for even a shorter period, and to remove the artifacts of multiple saccades, or errors in self-reports.

3.3 Conclusion

I began this chapter with a question. I promised that by the end of this chapter, you would be better equipped to answer that question, than you were before reading this analysis. *Do humans look at the same locations when they are trying to detect the same verb?* We saw in the first pilot experiment with the eye-tracker that looking for particular verbs (in the primed case) narrowed down the search space significantly. We also observed that the occurrence of some verbs may automatically draw attention, because subjects looked at them in both the primed and non-primed cases. Perhaps most importantly, the two subjects seemed to converge more than before when they were told to look for the same verbs. One may say that the subjects memorized the locations where the verbs occur. I think this objection is valid, but insignificant. First, as we saw in the pervious chapter, post-attentional affects are almost non-existent. Second, on several occasions, subjects that actually correctly looked at the location of a verb in the free-observing case, did not do so in the second time.

In the second experiment, we also observed that subjects are mostly in agreement on the location that they look at to respond to a question. We also saw the significant draw of objects and hands for action recognition, as predicted by previous studies(Ullman et al., 2012; Wolfe, 2000). While faces seem to have a lot of attraction when freely observing subjects look at an example, hands and extremities are more important for tracking actions. This is emphasized in the second experiment. In chapter 5, we discuss the implications of these observations on computational models.

Chapter 4

The Secrets of the Periphery

Peripheral vision, the visual region outside what is processed by the fovea, has a much lower resolution than that of the foveal vision. Furthermore, visual attention is rarely allocated to vision outside fovea, as most interesting targets are instantly foveated by the observer. Accordingly, you may ask if peripheral vision plays a significant role in action recognition. In this chapter, I argue that it does, although *indirectly*. First, I provide evidence that suggests that a weakened peripheral vision does not affect action understanding significantly. Then, I also show that a strengthened peripheral vision, one with full attention allocated to it, is still 2 to 3 times weaker for action recognition in comparison to foveal vision. Hence, I argue that peripheral vision is not good for directly recognizing actions, however, it plays a role in guiding attention (and foveal vision) to the location on the image where action recognition occurs.

4.1 Bull's eye - a pilot study

In this pilot study, conducted by Pedram Razavi and me, we tried to eliminate peripheral vision, and guide the subject's attention and fovea together. We were interested in observing the effects of this elimination on the subject's understanding of the actions that occurred in some short clips. This study was conducted online, where subjects observed two single-sequence videos, and explained what they understood.

4.1.1 Experiment set up

For each subject, one movie was played normally. The second video, not seen by the subject before, would be blurred in the area that wasn't foveated. This effect was introduced to emulate the drop in acuity across the visual field, and also drove the subject's attention, and point of gaze towards the desired location. The 'gaze-point' was generated by the tracking the eye of a person observing the video for the first time. Figure 4-1, demonstrates screen-shots of the videos in blurred and non-blurred editions. Each video was approximately 2 minutes long. After the video, subjects were asked to describe the events that occurred in the video in 5 sentences, and later also chose the observed actions from a multiple-choice list. The actions in the list ranged a variety of events that happened, e.g. "man tipping the doorman" to "picking up a suitcase", and a few distractors similar to these, which did not occur in the videos.

Figure 4-1: Two screen shots from the videos are shown, each with or without blurring. Each subject would see a blurred video and a different non-blurred video.

4.1.2 Data collection

Twelve volunteer subjects were recruited to conduct the study online. All video's were played from Youtube.com and subjects were asked to watch the video in full screen mode. Each video was shown to half of the subjects in the blurred version, and to the other half in the normal form.

4.1.3 Results

As with the previous pilot study, we provide suggestive evidence. The first observation comes from the short video clip taken from the movie *Good fellas* made in 1990. Five people out of six who watched this clip mentioned that the man is well-known or famous, while none of the people who watch the version with the blurred periphery mention this. It seems that the blurred vision caused the subjects to fail to see the point-of-gaze of the other people in the video and their gestures towards the couple. Interestingly, the person whose recorded trajectory was used for the video had also noticed the popularity of the couple. It may be that the eye-tracker failed to register quick saccades, or alternatively, peripheral vision sufficed to induce such conclusion. Apart from this difference, the subjects tend to describe the story similarly. Two excerpts are reproduced below.

Subject with blurred video:

"A man and woman, presumably a couple, walk through a fancy reception area. They give some money to a doorman and then go through a series of narrow corridors, eventually going through a kitchen. There is a lot of action in the kitchen and seem to be very busy. There is a private room of sorts where they end up, where they are seated and presumably are welcomed to the dinner."

Subject with normal video:

"A guy and a woman are going to a formal club. They enter through the back entrance. Apparently the guy is famous or friends with the owner. The guy pays the a[sic] man to enter. The place seems full but they get a table. They sit in the place."

Another expected result was that subjects who were shown the blurred videos

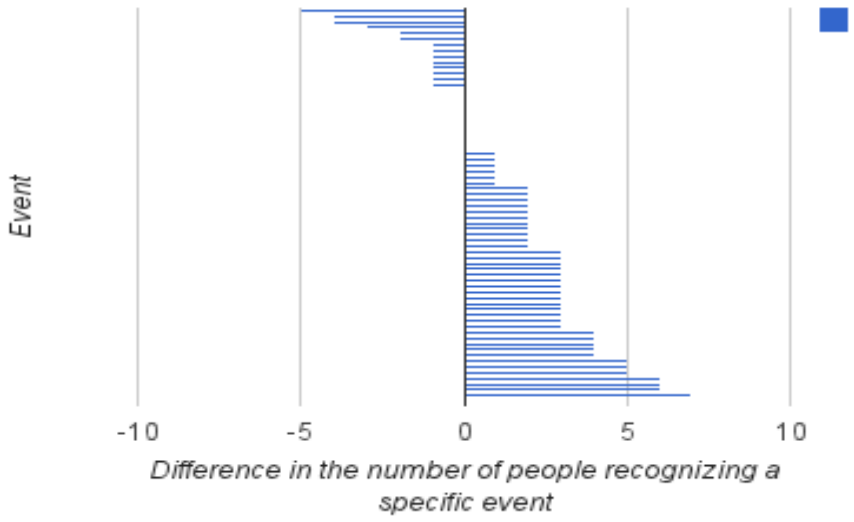

Figure 4-2: Subjects could choose the occurrence of an event among 68 different possibilities. The number of correct responses for the normal group was subtracted from the number of correct responses in the blurred group for each event. A bar to the far right indicates that many more people with the normal video noticed the presence of an event. A bar to the far left indicates many more people with the blurred video noticed the presence of an event.

tend to miss more of the occurred actions (queried in the multiple choice questions) than those that get to observe the videos normally. This is consistent with our observations. Like the previous case, most of the lost information has to do with the context, rather than the main story line. In Figure 4-2 I provide a schematic of the difference between the correctly reported events for the two groups.

4.2 The unknown in the periphery

In contrast to the study I just discussed, I conducted a large scale experiment in which subjects did not foveate, although they did attend, verbs in the periphery. As I discussed in the second chapter, gaze and attention are separate entities. Directly looking at an object affects the visual resolution. Attending an object involves different processes that I discussed. This experiment was designed to study the effects

of attention in action recognition, outside foveal resolution. In our daily lives, many actions occur outside our foveal vision, but we are still more or less aware of them. When we watch a movie, although we focus on a particular location, we are aware of other events that happen in a scene. This experiment studies how well humans can infer verbs from the periphery when they do pay attention to those locations, but their gaze is located somewhere else. I then add distractors to occupy attentional processing to some extent. I study how well subjects can detect the occurrence of verbs in the periphery, with or without distraction.

4.2.1 Experiment set up

Subjects were given a short explanation of the process. They were asked to watch 10 short videos ($\sim$ 35 seconds) each consisted of three mini-videos and answer questions about them after each video. While watching the video, they were asked to stare at the red cross in the center, but they knew that they will be asked about the verbs occurring in the periphery. Each video contained three verbs presented in three random locations from either 20 or 64 possible locations. Five videos included no distractors at the gaze point, and five included a distractor - either a text or a video. After each video the subject would view 3 multiple choice questions (18 identical choices) and pick the verbs that they observed in each video. If distractors were present, subjects would also be asked about them. Subjects were asked to use full screen mode when watching the video. This was checked programmatically and the questions would not appear if the subject failed to watch the entire video in full screen. In such cases, another random task would appear and the subjects could reattempt. For online experiments, subjects were not payed by correct results, but only by effort and successfully completing all 10 tasks. A small number of in laboratory experiments were run as controls to compare the results with those of the remote subjects. In the in-laboratory experiments, subjects wore the eye-tracker to ensure their gaze did not deviate from the fixation point. The entire task lasted about 25 minutes. Subjects were asked to keep their eye-level along the point-of-gaze and adjust their distance to the monitor such that their eyes were about the same distance from the monitor as

the diagonal size of the monitor itself. This way the visual angle across the monitor can be calculated as:

$$V = 2 * arctan(\frac{S}{2D})$$

Where S is the size of the diagonal and D is the distance from the monitor. When $S = D$, V becomes constant across all subjects. The $S = D$ was fixed because it was the easiest yardstick available to every online subject. This way the visual angle can be unified across the study cohort.

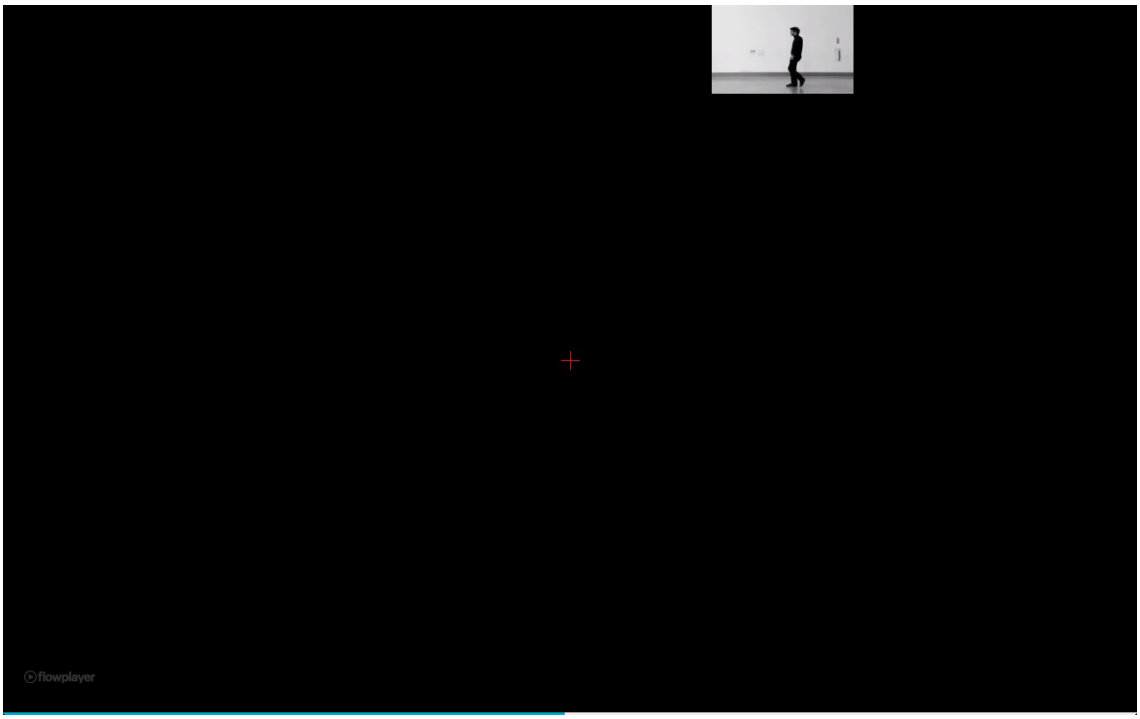

Figure 4-3: Subjects are asked to stare at the red plus in the middle. Three short videos appear in different locations on the screen before the subject is asked to report them. One of the mini-videos is visible in this screen shot.

4.2.2 Data collection

A total of 76 subjects were recruited on Amazon Mechanical Turk. Each subject provided about 30 evaluations. Four subjects were tested more extensively in laboratory while wearing an eye-tracker, with each subject providing 60 evaluations. The in-laboratory subjects were also tested with different mini-video sizes.

4.2.3 Results

In laboratory results (without distractors).

In this section, I present the results obtained from subjects that performed the experiments in laboratory. In this set of experiments, I collected most of the data when the screen was divided into 20 separate locations where the video could appear. The task video size was 200 by 150 pixels. The combined results between all subjects is shown in Figure 4-4. It is notable that the pattern for each individual does not necessarily conform to the overall pattern. This can be caused by two separate factors: First, the detection pattern does not need to be uniform across the visual field (Afraz and Cavanagh, 2008). Second, the difficulty of the verbs shown to each subject at different locations may have been slightly varied, especially when subjects had different opinions about what a verb exactly means. Subjects were not forced to do the task again if they looked away, hence the eye-tracker was used to evaluate how often the subjects did in fact deviate from the center. Subjects would self-report if they did not manage to fix their gaze properly. With this set up, the subjects looked away, and reported their deviation, in $\sim 8.2\%$ of the time, while an additional $\sim 2.1\%$ deviation was not reported by the subjects but observed by the eye-tracking.

The results provide a quantitative sense of how the lower resolution of peripheral vision affects action recognition in humans, even if they are actively paying attention. Of course, in a real-world situation, quick saccades to the target results in foveation. For the farthest locations (~ 35 degrees), the true positive recognition rate was almost half of those near the fovea. While the sample size was small, I ran this as a pilot study to inform the large scale experiment on Amazon Mechanical Turk, which is described in the following section.

Online results (without distractors)

This experiment is similar to the pilot study discussed in the previous section, with two important caveats. The mini-video sizes used now are smaller, in order reduce the errors of momentary deviations from the gaze point as the online subjects were not controlled. Second, apart from the programmatic measures to control subject's

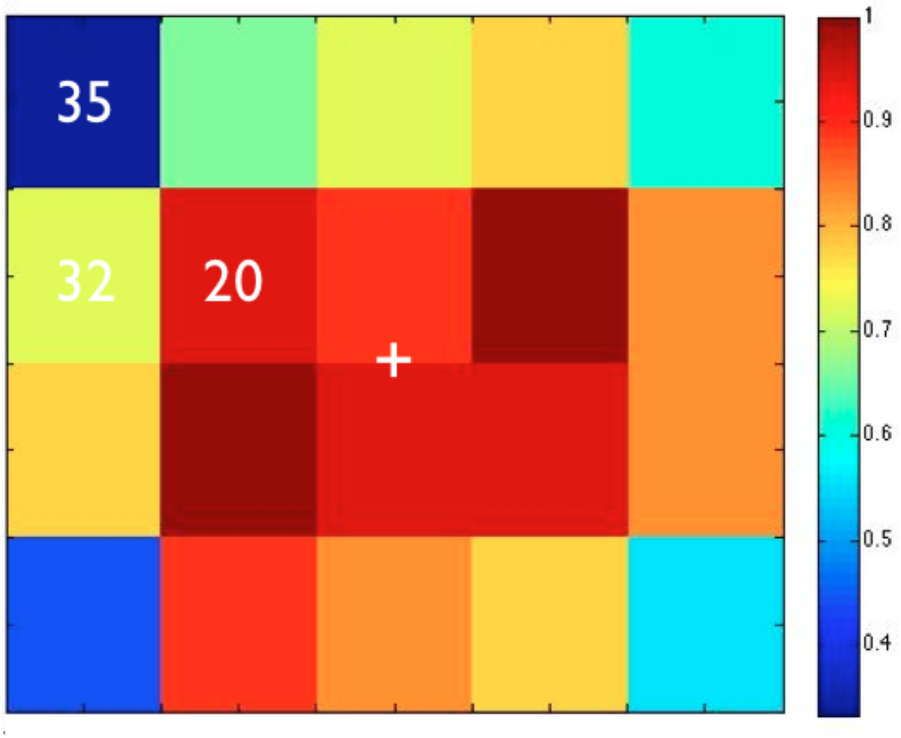

Figure 4-4: Four subjects were asked to stare at the red plus in the middle. Their eyes were tracked for deviations. Three short videos appear in different locations on the screen before the subject is asked to report them. The results is color coded to indicate the accuracy of recognition. This heat map is produced from 240 recognition tasks, with 12 questions per box. The video's approximate angular distance from the point of gaze is shown in white.

behavior (e.g. requiring full screen mode), no real evaluation of subjects adherence to the instructions was available. This was partly due to the challenges of implementing such a system, and partly due to the Amazon Mechanical Turk policy that subjects may not be identified by the experimenter. The experiment was designed such that subjects would take the same amount of time, whether they cheated or not, to finish the task. Additionally, they were payed based on their effort and not performance. If they looked away from the desired point, they could indicate so without a penalty. In about 11% of the cases, subjects reported such incidents. Therefore, a reasonable error rate due to such artifacts should be between around $14-17\%$, if online subjects behave slightly less faithfully than those in laboratory.

It is also noteworthy that there seem to be no particular performance effects for each eye's blind spot. As the subject keeps both eyes open, every location that falls under one eye's blind spot is covered by the other. Although this means that the data is initially processed by less resources in the cortex, it does not seem to affect the performance on the task in an objective manner.

Appearance	True Positive Rate	Total Points
Verb 1	0.52	617
Verb 2	0.51	528
Verb 3	0.51	557

Table 4.1: Subjects perform similarly regardless of the order in which verbs are shown, suggesting they do not forget their recognition. The total points considered is not equal because boxes with too little data were eliminated.

Given that the subjects watch 3 different mini-videos before answering questions, one may also ask whether there are memory effects that change the rate of recognition for the subjects. As shown in Table 4-1, such effects are not suggested by the subject's performance.

While the results follow the same general trend as that of the in-laboratory experiments, the heat map is not as clearly distributed as in the previous case (Figure 4-5). This could be due to the non-uniform distribution of data points across subjects. The order in which tasks were given was randomized, so that the subjects could reattempt if they were not happy with their progress. Therefore, some locations received less coverage, or contained easier verbs. The overall trend of lower recognition rate with higher distance from fovea does not change, as it can be observed in Figure 4-6. The reader may ask what is meant by easier verbs? Does it relate to complexity? Similarity to other actions? The time-frame in which the action happens? I do not propose such definitions. Instead, I retrospectively evaluated the performance of the subjects when they were presented with different verbs (See Figure 4-7). There is some correlation between the difficulty of the verbs as defined in chapter 2, and what is observed in this experiment. For example, *Walk* is an easy verb, while *Catch/Throw* are harder. One factor for this extra difficulty is that catch and throw require an object as well as the actor, while for walking one actor suffices. Another factor is

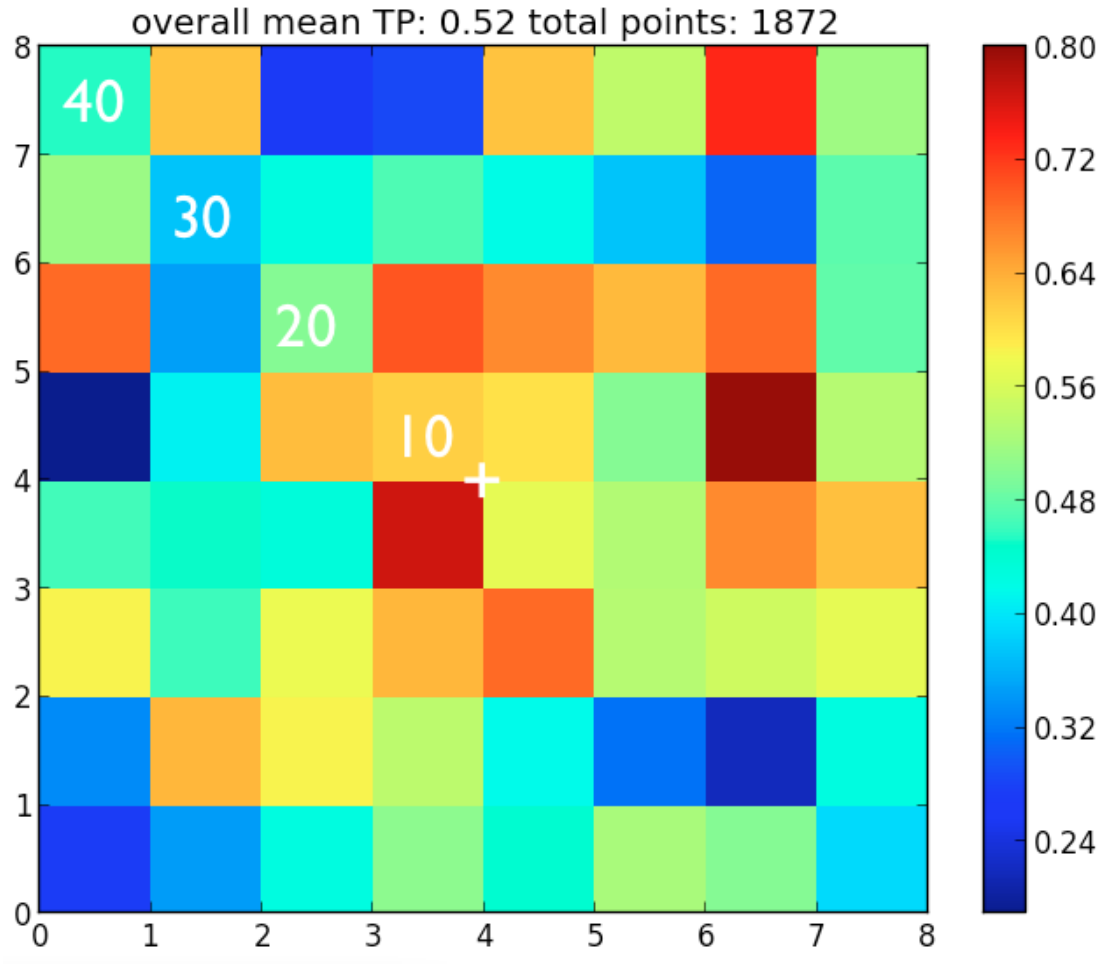

Figure 4-5: Subjects are asked to stare at the plus sign in the middle. The results is color coded to indicate the accuracy of recognition. Note that the video size is smaller than in the in laboratory case. The true positive recognition rate is about 52%. The video's approximate angular distance from the point of gaze is shown in white. All boxes had a minimum of 5 responses, with an average of 29 responses per box.The display distribution was slightly more biased towards the distant locations, i.e. more data points were collected from those areas, as subjects had a stronger tendency to look away for distant videos.

that catch and throw occur in a very short time period. This is further supported by the observation that videos that had a quicker *Catch/Throw* progression, were less often correctly categorized than the slower variants. Interestingly, some verbs that are similar in complexity were often harder to categorize than others. For instance, *Put down* appears to be more difficult for subjects than *Pick up*.

When humans observe a scene, even in the periphery, one of the cognitive tricks

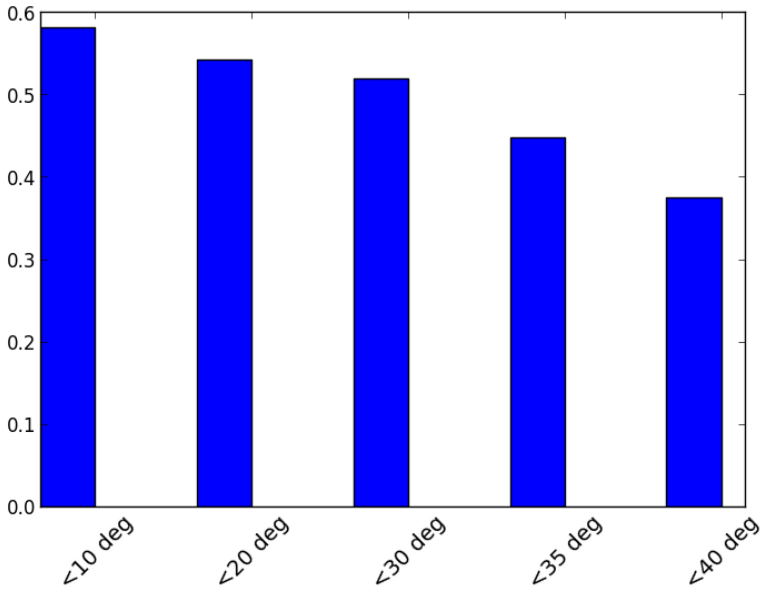

Figure 4-6: True positive recognition rate, with respect to the angular deviation from fovea in degrees. The sets are not overlapping, i.e. < 20 does not include the data from < 10.

they use is to reduce the search space by assuming the possible verbs available. Swimming is not a possible verb in the desert. In this study, I already performed this reduction for the subjects, by giving them multiple choice options. This helped with the analysis, and also unified the subject's reference of selection. However, I also intentionally left out one verb that would appear, but was not among the multiple choice options: *Crawling*. In the in-laboratory experiments, this verb was recognized as often as *Running*. However, when it was eliminated from the options, so that subjects had to manually enter the verb if they saw it, this recognition rate dropped significantly. Of course, a possible source of error in this case is that subjects are simply unwilling to take the time and type in their response. For this purpose, I also included one task in which subjects had to manually write all responses after they had completed four multiple choice tasks. While the recognition rate for all verbs dropped, crawl was still not recognized as often as expected by a low complexity verb.

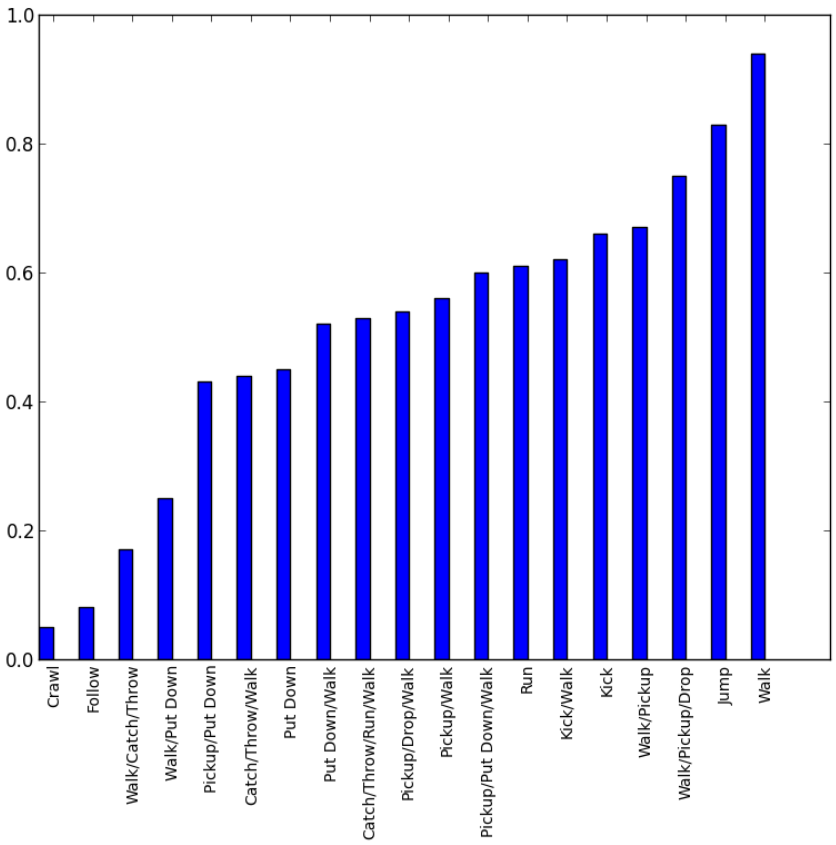

Figure 4-7: Accuracy of recognition for different verbs when no distractor was present. The case for crawl is special, because while crawl was shown, it was not explicitly among the multiple choice options. Subjects could fill in an 'other' box if they saw a verb that wasn't part of the multiple choice options.

Online results with distractors

In real-life situations, humans often direct both their gaze and attention to the same locus. Hence, not only the resolution in the periphery is lower, but also the stimuli in those regions are rarely attended. Without any attention this case is not too interesting, because there is very little recognition possible. The more interesting case is when the subject has to track two events at the same time. While for static, or slowly changing images, this might be achieved through saccades, in a busy scene, one often has to make a choice to either divide attention serially, or in parallel. In an extension to my experiment in the previous part, I looked at the effects of distractors. These distractors were not supposed to occupy attention fully, as in those cases the

chance of deviation from the fix point was larger (i.e. it could not be done online). However, as seen in table 4-2, even a mild distractor can affect the recognition results.

Distraction	True Positive Rate	Total Points
None	0.50	1821
Text	0.49	387
Video	0.44	1521

Table 4.2: Subjects perform worse when part of their attention is occupied by a distractor near the point of gaze. If only a text (a verb) is used as a distractor, this drop is negligible. But if a video, similar to the target videos is shown, the interference becomes more obvious.

Note that in order to make a direct comparison, I had to eliminate some of the points from the no-distractor cohort, because they were very close to the fix point. In the distractor included group, no videos were shown in those spots, because they would overlap, or be very close to the distractor itself. Therefore, the no-distractor group's recognition rate is now slightly smaller than what was reported in the previous section.

4.3 Conclusion

So how important is peripheral vision for action recognition after all? It seems that peripheral vision can be eliminated to a great extent without the loss of too much information. However, this is assuming that attention and gaze are automatically guided, which is not the case in real life situations. In my experiment, I artificially forced the subject to follow an attentional path that was recorded from another person. That attentional path seemed to contain a significant amount of the data that was necessary for interpreting the general flow of events in the scene. Nonetheless, the foveal trajectory, and consequently the blurred periphery, would be impossible to generate had the original subject been deprived of peripheral vision. Hence, it may be reasonable to say that peripheral vision, in this case, is not about *What is there?*, but rather it addresses the question of *What may be there?*.

When attention is coupled with the peripheral vision, a great amount of infor-

mation can be retrieved. This case is of course also artificial, because in real life, if full attention is dedicated to a location it is usually accompanied by gaze. I observed that subjects recognized almost half of the verbs on average, if they were helped by a pre-constrained search space. Additionally I observed that as attention is reduced, even mildly, the recognition rate drops tangibly. Extensions to this study may include more attention-heavy tasks as distractors, or eliminate the constrained search space.

Chapter 5

The Attentive Computer

The observations I made in the previous chapter are not general enough to be conclusive, but they are consistent. Hence, they can be used as reasonable heuristics to construct and improve algorithms for action recognition, specifically by attentional processing. Like human brains, computers also have to make do with limited capacity processors. Moreover, some computations are expensive to uniformly apply across the entire input. In this chapter, I propose general heuristics that can be used in any action recognition algorithm. Furthermore, I sketch out a mechanism for an improved, task-specific, action recognition system using these heuristics.

5.1 Heuristics

We observed in chapter 2 that the current systems of computational attention generally drive their systems using the visual properties of an image, often static frames. Many systems perform reasonably well in predicting the most attractive (attention drawing) components of an image, somewhat similar to human pre-attentional processing.

However, there is an element of symbolic thinking involved in action recognition that is not easily resolved with visual stimuli alone. As we saw in chapter 3 (Q and A experiment), when posed with a question to begin with, humans look at objects that are required for that verb. Even in the case of smelling, which is invisible by itself,

many people drove their attention to the nose, at least partially, although there was no visual information there corresponding to the verb *smelling*.

Heuristic 1. *For action recognition, start from videos, not still images.*
Humans do not learn actions from still images. In fact, I think it may be impossible to teach a human an action with a single still image. At the very minimum an action is consisted of multiple still images. However, many actions, such as kicking, compared to touching with the foot, are indistinguishable if the temporal aspect is eliminated. When presented with a still image, humans have enough time to saccade to multiple locations and make inferences as do computers. In live scenes or videos, a lot of the information in the periphery is never foveated by a human, and even if mostly eliminated, a large fraction of the story can be recovered (chapter 4, experiment 1). Furthermore, you can teach a human what walking is more easily by a stick figure animation, rather than a still image of a human. We should use the same approach for computers. The extra load of processing videos, rather than images, can be resolved by attention. How should we drive attention on a video? As I discussed, the basic feature maps widely used for pre-attentive behavior are not sufficient for action recognition. In order to drive attention on videos, I suggest that movement is used as the basic heuristic.

Heuristic 2. *Movement is a key driver.*
When humans observe a live scene, everything is changing. Objects that are closer change positions faster in the visual field, than those afar. The closer objects, people and actions are naturally more important to the human observer. In action recognition, a key driver of attention is movement or change. Movement is easy to detect in videos, while it is almost impossible to accomplish in still images. Even when people are presented with still images, they hallucinate movement (Figure 3-8). Almost all actions require movement, hence, using heuristics one and two together and *attending* movement can improve mechanisms for action recognition.

Heuristic 3. *Hands, bounded objects, and faces are useful, in that order.*
When it comes to real life videos, humans have inherent biases. From a very early age,

humans look at hands (Ullman et al., 2012), point of gaze, faces, and other bounded objects. Attention and visual saccades almost always move from object to object (Wolfe, 2000). Introducing a catalogue of such biases can help reduce the complexity of action recognition.

Heuristic 4. *Approach the image with a (even vague) symbolic hypothesis, reducing the search domain.*

This is a rather trivial but noteworthy case. We saw in the fourth experiment of chapter 4 that humans also perform much better when the problem of recognition was reduced to educated 'selection' process among possible verbs, rather than pure recognition. In fact, the general applicability of this 'selection' mechanism by humans further suggests that we should look for definitions of verbs from a symbolic, rather than visual perspective. The space for possible variations of a verb, visually, is infinite, while symbolically, this space is precise, tractable and small. Attention can connect these two together by applying visual processing to the key elements that are required by the symbolic definition of the verb. The verb *throwing* requires an object and an actor. Humans look for both of these elements when they try to recognize throwing, as we saw in the Q and A experiment. Knowing what you are looking for can reduce the search space, and is a good start for action recognition programs.

Heuristic 5. *Devote processing capacity unevenly and use the expensive processes wisely.*

When dealing with an input, one can process all of it evenly, or dedicate more processing to a particular region/aspect of the input. Human attention is an example of the second approach. We saw in chapter 2 how such an approach can increase vigilance, speed of processing, and also enable limited-capacity expensive processes like face recognition to be applied to faces in the image one at a time. In the computational approach, one could use techniques such as intermediate feature computations (Ullman et al., 2002) or pattern matching on the the attended region and pull out information that would be expensive to compute (at least in real time) on the entire image.

With these heuristics in mind, in the next section I propose a rather general

mechanism the can be used for computational action recognition.

5.2 Attentive vision for action recognition

As an example of how one can integrate my proposed heuristics on a video to generate an attentive process, I provide a sketch of a mechanism that can be implemented. The details of the mechanism are not fully worked out, but those are design decisions that are made while implementing a particular instance.

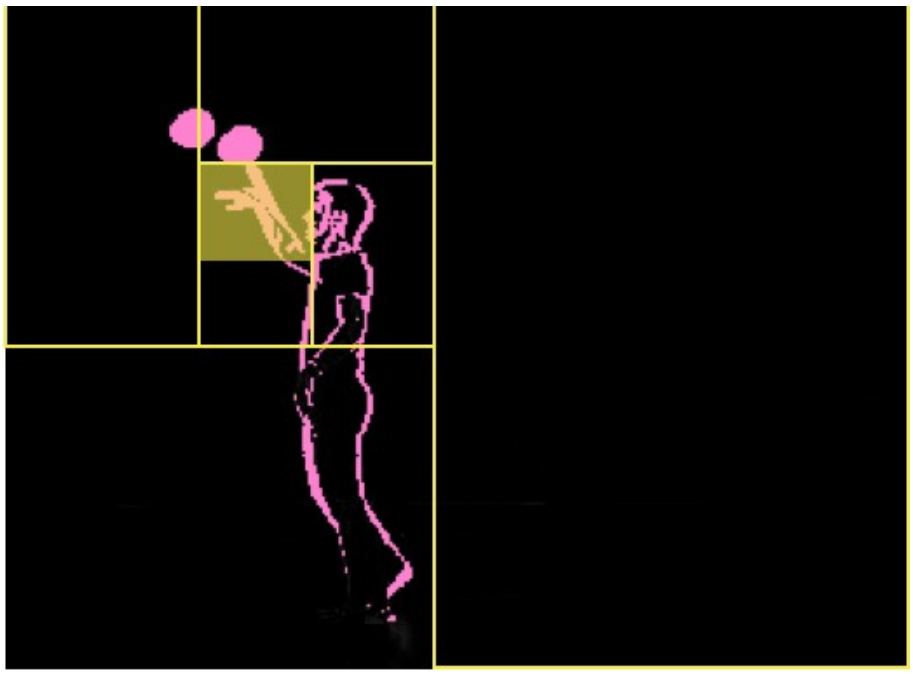

Figure 5-1: A mock-up of the divide and conquer approach that can be used to optimize image processing and finding the best attentional locus. The pink pixels are the non-zero locations of the frame subtraction operation.

---

1. *Initiation:* For the first frame, use a saliency map model, like the one used in chapter 1, to compute the starting draw of attention.

2. Calculate frame difference, between current and the previous n frames. This will turn all moved pixels to non-zero (or above a threshold k), while static pixels (or those below the threshold) are treated as zero. Generally, for 24 fps videos, $n = 1$ should suffice.

3. Search the space progressively (for non-zero pixels) according to one of the algorithms below.

   (a) Brute-force: use a sliding window approach. Score each window according to the number of positive samples (and other heuristics).

   (b) Divide and Conquer: Divide the space S into two and sample $i \propto |S|/2$ points in each. Choose the space with higher number of positive samples, and repeat until the space is divided to a minimum size $|S^*|$. This can be decided based on the problem at hand. Memoize the spatial division for the next round of sampling, report positive samples for each region. This approach can also be optimized by dividing the space in a manner similar to the approach used by Barnes-Hut simulations(Barnes and Hut, 1986). In this context, positive points can be treated as bodies, and the mass of each region is simply the score. The highest gravitational center in the system gets the highest score. A mockup of how the program would divide the space is presented in Figure 5-1.

   (c) 2-D Gaussian function $f(x, y)$: Sample the points based on a Gaussian probability distribution with its mean at the center of current attention. Divide the space into similar sized boxes(again, the minimum size is decided by the problem), and regions are scored accordingly.

4. Score the regions based on some function $f(x_1, x_2, \ldots, x_m)$. The parameters can be based on the intensity of the movement (positive samples), the size of the area (spread of the samples), the distance from the current gaze point, and other criteria such as heuristics 3 and 4 of the previous section.

5. Inhibit the current point of attention.

6. Choose from the ranked region of points probabilistically, according to a distribution that includes the scores.

7. Move on to the next frame, return to step 2.

This mechanism is provided to serve as an example of how the heuristics proposed in the previous section can be incorporated into a computational and implementable system that provides attention for action recognition. The mock-up shown in Figure 5-1 is generated by a partial implementation of this algorithm, without the well-trained scoring schema in place. I believe that an intelligent scoring schema, as well as features such as face and hand recognition, can make this mechanism robust. I hope to pursue this line in the future research.

Chapter 6

Contributions

I conducted this research to provide insight for the development of computational mechanisms for action recognition. I think that a good understanding, and more importantly a working computational model of attention can greatly improve current action recognition systems. Accordingly, I took numerous steps to clarify aspects of attention in order to facilitate its incorporation into computational models. These steps are listed below:

1. I **justified** the performance advantage, at least in action recognition, of attentive vision over a non-attentive one by discussing the previous results in psychology and computational modeling of attention.

2. I **identified** numerous gaps in the understanding of attention, and its applications to action recognition, by studying the psychological and computational understanding of attention, as well as computational approaches to action recognition. I argued that static images lack movement, which is fundamental in action recognition, and therefore are not suitable as primarily inputs for a learning algorithm targeting action recognition.

3. I **proposed and conducted** two experiments on how attentional search occurs in humans. In the process, I elucidated aspects of attention and gaze that can be useful in improving algorithms for action recognition. For instance, I demonstrate that there is consistency of visual behavior among multiple subjects

when they are asked to report the occurrence of a verb. Likewise, the same subject will look at similar locations when searching for the same verb

4. I **proposed and conducted** two experiments on the role of peripheral vision in action recognition and the extent to which its elimination could affect human understanding. I observed that eliminating the periphery, while guiding attention, does not seem to affect human understanding of a scene significantly. I also provided evidence on how well humans recognize an action when it is only presented in the peripheral field. In the process, I also collected results, in various parts of the periphery, about the relative difficulty for humans to recognize various actions. Finally, I demonstrated that while peripheral vision may not play an important *direct* role in action recognition, its indirect role through the guidance of attention may be crucial for that functionality.

5. I **proposed** simple computational heuristics that can be used to improve current models of action recognition based on my findings from human experiments. I also proposed a framework for how these heuristics can be implemented in computer systems.

6. I **refined and helped in improving** the performance of the PUPIL eye-tracker for my use in my own experiments.

7. I **implemented** two distinct tools for massive online data collection, which can be used in similar studies on online subjects.

I hope that these steps can be built upon, and I firmly believe that incorporating attention into computer vision can increase the problem solving capability of these systems significantly. Additionally, I think that attention connects vision to symbolic thinking, and that would be a major step towards building more intelligent systems.

Appendix A

Tool Development

Throughout this thesis I developed and improved a number of tools in order to study human attention and foveal vision. In this appendix, I explain some details about the implementation of these tools.

A.1 Eye-Tracker Augmentations

The PUPIL eye-tracker was designed and developed by Moritz Kassner and William Patera as a part of a Master's thesis at MIT(Kassner and Patera, 2012). When I started this project, this eye-tracker was still in development and many functions were not yet in the state needed for my study. Along with being the first beta-tester of this device, I made several modifications to the device in order to improve its performance for the purpose of my study. My work with this eye-tracker consisted of several steps before I finally used it to collect data and these steps are presented in chronological order.

1. **Constructing a prototype (hardware) for PUPIL, based on the first design** (Figure A-1). During early development, I finished building a first-prototype headset for PUPIL, using Xbox cameras. Later I received another headset built by the developer's of PUPIL, which has a smaller eye-camera with specifically designed hardware and a world-view camera with a fisheye

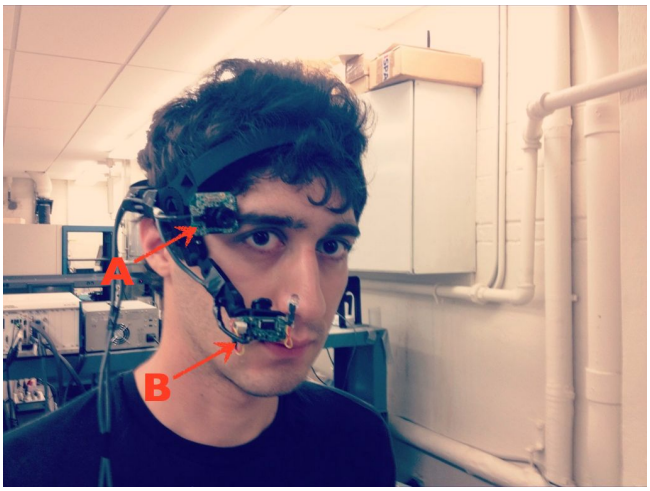

Figure A-1: A picture of the author wearing the first generation headset, A: The world-camera captures the subject's visual field, B: The eye-camera, connected to two infrared LEDs points towards the pupil to track the gaze. The rest of the process is done on a computer which connects to the headset using the USB cables

lens. The second prototype is easier to wear and slightly better in performance. However because of the fisheye lens, after recording, the data must be corrected to correspond to a flat world view.

2. **Working with the developers of PUPIL as a beta tester of the code.** At this stage I mostly helped with debugging and improving the program. During this time the software was upgraded at least 6 times because of the errors that I discovered together with the developers. The first problem was to iron out the platform and library dependencies such as installation of opencv, ffmpeg, numpy scipy, glumpy, PyopenGL and antweakbar in a compatible, functional manner. The second issue was to edit the minor parts of the code that caused compile-time errors, such as problems with recognizing the cameras, being unable to access certain libraries in different running modes (e.g. issues with the appropriate architecture each library was configured to operate with and cross-OS discrepancies). The third difficulty, which took the longest to resolve, was to check the correctness of the tracking results and fix the runtime problems, and return values that were hidden until certain functions were called during runtime. This process was iterative and time-consuming, with the last occasion

of a bug discovery and fix being two months after my first data collection.

3. **Collecting a small sample to refine and optimize PUPIL.** Three volunteer subjects were used for these experiments. Fine tuning the calibration process across subjects was especially challenging. PUPIL uses mechanisms such as canny edge detection and Kalman filtering to detect the pupil of the subject. This method, however, is subject to errors because the position of the eye changes and especially when the illumination on the pupil is different from the original state in which the pupil was detected. In my experience, these errors worsen when the subjects have darker eye color, even though the eye-camera uses IR range for the detection. The fundamental problem, is that once you have detected the pupil and calibrated the pupil location to the world-view, there is no guarantee that the pupil is tracked during the entire time of the recording. This makes the data collection process lengthy and tiresome to some subjects, especially if the environment is changing or the subject is required to move their head. To deal with this, I generally consider the data in which the camera is calibrated and the pupil is detected correctly for more than 70% of the time acceptable, and then apply probabilistic smoothing techniques to recover parts of the missing gaze-path.

4. **Developing the software further and designing new experiments based on a better understanding of the device's limitations and strengths.** In the first set of experiments, the subjects were looking a monitor that did not fill their entire field of view. Therefore, it was impossible to make a comparison between where multiple subjects looked at before remapping all of the gaze coordinates to the same plane. Originally the developers implemented a scale invariant feature transform (SIFT)-based homography algorithm, to remap the world view to the video. This functionality did not work properly so I cooperated with them to fix it. However, even then the results were not satisfactory. Therefore, I developed two other methods using edge detection to determine where the monitor is and remap the monitor screen to the video. The method

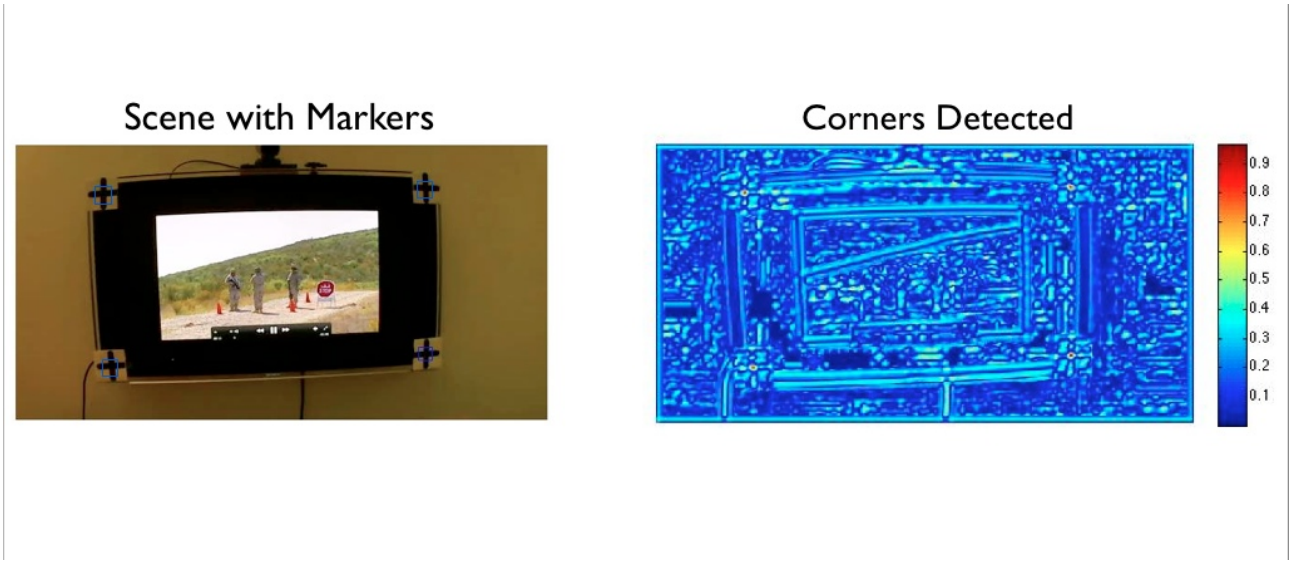

Figure A-2: An example of the performance of the marker (plus-sign) detection extension to the eye-tracker software. The very small red dots on the heat map (right) are the coordinates of interest. All four are robustly detected by the software.

finally used was presented in Figure 3-1. I also developed another more accurate mechanism, which relied on markers using cross correlation analysis between the marker and the target image(See figure A-2) . As most of my data was collected without markers, I did not use this second method for my data analysis.

A.2 Foveal and peripheral vision testing software

For the online and in-laboratory experiments described in chapter 4: 'The unknown in the periphery', I developed a platform to automatically generate a large (more than 1000 tasks) testing set. This pipeline can be used on any type of study regarding the peripheral vision. The fast, in-labortatory version is ideal to be combined with the eye-tracker. The web-based version is designed to collect information from remote subjects, and at the same time minimize cheating possibilities.

The software is an amalgam of different technologies. For generating the tasks, I have used Processing programming language. Although Processing programs can be run directly from within a browser, such use is not recommended as this may slow down the frame rate of the videos significantly. Therefore, I have also developed a

javascript program, based on flowplayer library. This allows me to track user actions specifically for the purpose of evaluating the data. For example, the video will not play until fully loaded and the subject cannot complete the task without watching the video in fullscreen for its entirety. The screen resolution is also measured. The responses were collected in an embedded Google form, in order to eliminate the necessity of a server. However, a server side code (like the one in the next section) can be easily incorporated to send back the results automatically.

Once data is collected, the analysis software (written in Python) is also broad and can be modified to analyze the data.

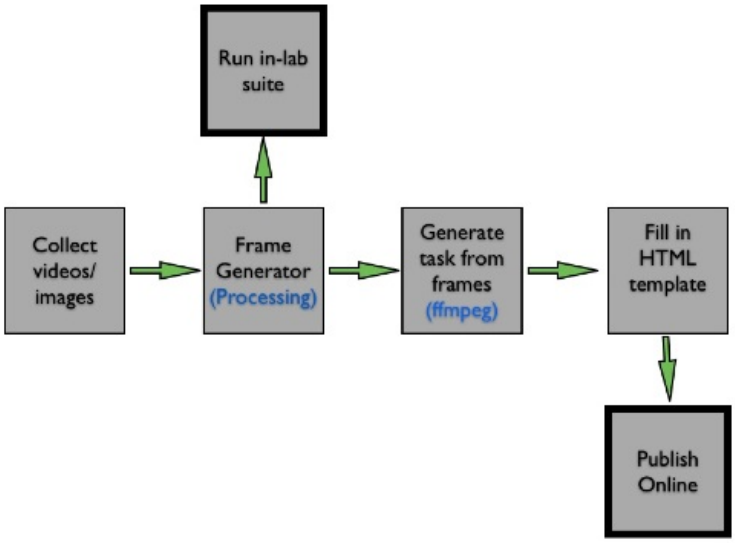

Figure A-3: The pipeline for the foveal/peripheral vision software. After collecting the videos, they can be put into the program's input folder. For in-laboratory experiments, one can simply run the in-laboratory mode at this step. Otherwise, tasks can be generated from frames in the online-mode of the software. Once video's are generated (ffmpeg is recommended), they can be put in the static folder of the web-export, and are ready for publishing.

A.3 Mechanical Turk server

Finally, for the experiments described in chapter 3: 'Q and A', I have implemented a light-weight server to collect clicking data from images from online users looking at a picture. This can be used to serve any type of pictures and collect click data

correspondingly. The program can be obtained from github, and is readily deployable on heroku. The software for this program is also available per request.

Appendix B

Figures

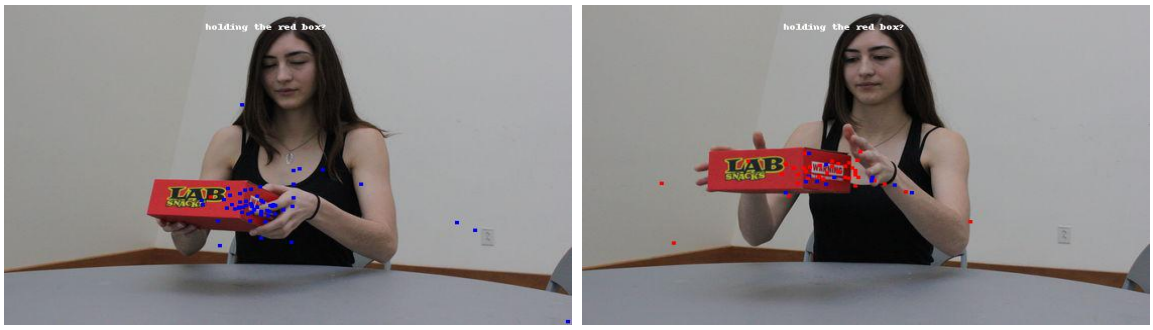

Figure B-1: Question asked: *Is someone holding the box?* Each point indicates a click by the subject. Red points indicate that the subject has answered 'No' to the question, while blue dots indicate that the subject has answered 'Yes'.

Figure B-2: Question asked: *Is someone pouring liquid?* Each point indicates a click by the subject. Red points indicate that the subject has answered 'No' to the question, while blue dots indicate that the subject has answered 'Yes'.

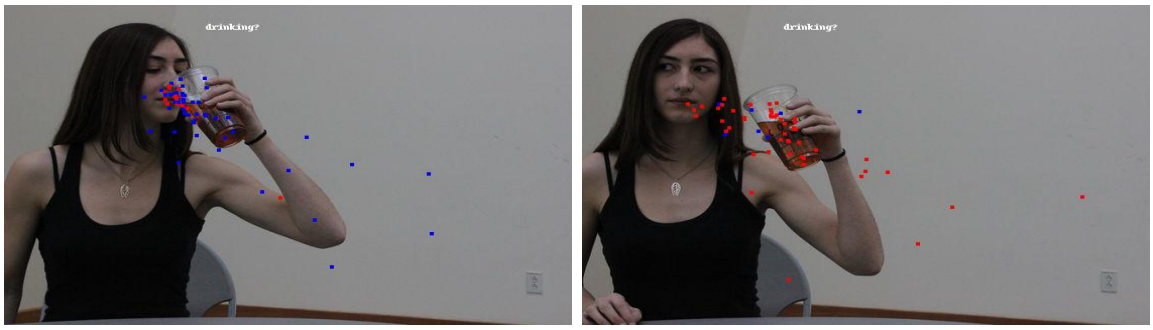

Figure B-3: Question asked: *Is someone drinking?* Each point indicates a click by the subject. Red points indicate that the subject has answered 'No' to the question, while blue dots indicate that the subject has answered 'Yes'.

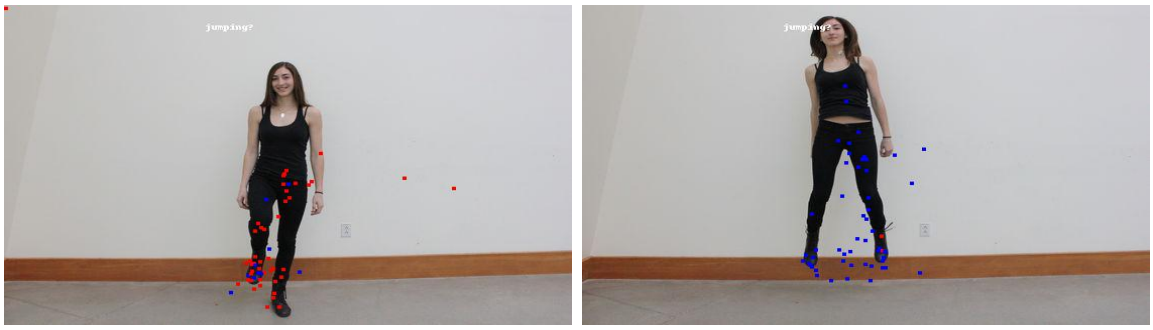

Figure B-4: Question asked: *Is someone jumping?* Each point indicates a click by the subject. Red points indicate that the subject has answered 'No' to the question, while blue dots indicate that the subject has answered 'Yes'.

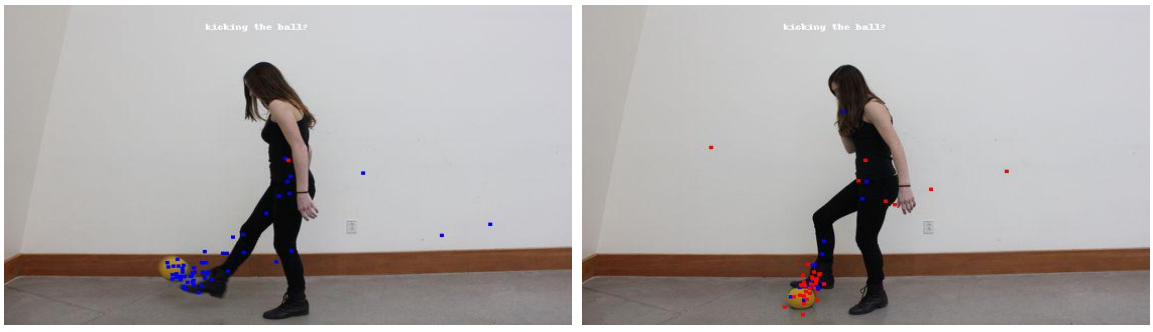

Figure B-5: Question asked: *Is someone kicking?* Each point indicates a click by the subject. Red points indicate that the subject has answered 'No' to the question, while blue dots indicate that the subject has answered 'Yes'.

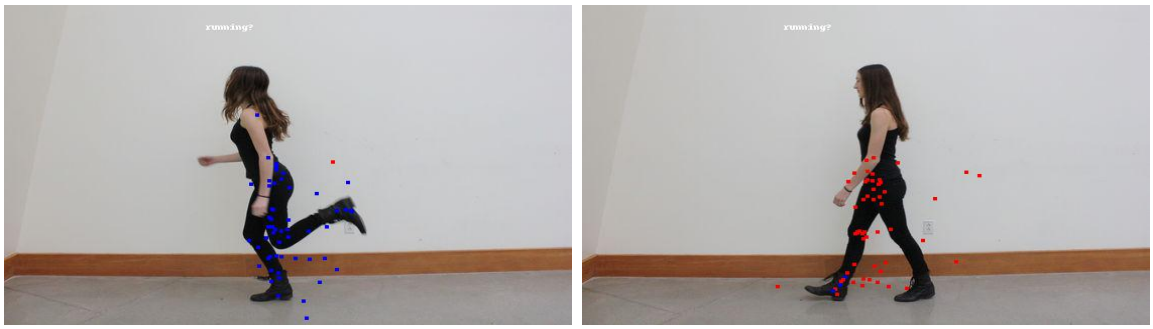

Figure B-6: Question asked: *Is someone running?* Each point indicates a click by the subject. Red points indicate that the subject has answered 'No' to the question, while blue dots indicate that the subject has answered 'Yes'.

Appendix C

Consent forms

This study has been approved by Committee on the Use of Humans as Experimental Subjects (COUHES) at MIT. Both P.H.W and S.S. have undergone human subjects training. The relevant consent forms for subjects are provided in the following pages.

CONSENT TO PARTICIPATE IN
NON-BIOMEDICAL RESEARCH

A Mechanical Turk Study on Human Attention Allocation and Verb Recognition

For subjects participating remotely through mechanical turk

This HIT is part of a scientific research project. Your decision to complete this HIT is fully voluntary. There is no way for us to identify you. The only information we will have, in addition to your responses, is the time at which you completed the survey and optionally your screen size and resolution. The results of the research may be presented at scientific meetings or published in scientific journals. Clicking on the 'SUBMIT' button on the bottom of this page indicates that you are at least 18 years of age,and agree to complete this HIT voluntarily.

CONSENT TO PARTICIPATE IN
NON-BIOMEDICAL RESEARCH
A Mechanical Turk Study on Human Attention Allocation and Verb Recognition
For subjects participating in lab wearing an eye-tracker

You are asked to participate in a research study conducted by Sam Sinai and Patrick H. Winston from the Department of EECS at the Massachusetts Institute of Technology (M.I.T.). This study will contribute to the masters thesis of Sam Sinai. You were selected as a possible participant in this study because you are above 18 years of age. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not to participate.

- **PARTICIPATION AND WITHDRAWAL**

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise which warrant doing so. If you participate in the study truthfully until the end, you will be paid in full.

- **PURPOSE OF THE STUDY**

The study is designed to evaluate the level of verb recognition in peripheral vision.

- **PROCEDURES**

If you volunteer to participate in this study, we would ask you to do the following things:
Step 1:
You will start by wearing the eye-tracker. For about 2 minutes we will spend time familiarizing you with the eye-tracker and calibrate it. You will be asked to look at a few points on the screen.
Step 2:
You will watch a short tutorial video about the procedure.
Step 3:
You will begin watching 30 second videos. During each video you are expected to stare at the red cross in the middle of the screen. If you look away, we repeat the video. You will try to guess what actions have happened on the video in the periphery. Subsequently, you will fill a short survey on the actions you observed.
This step is repeated 10 times.
Step 4:
You will be paid.
The study will take about 20-30 minutes, depending on the time required for calibration of the eye-tracker.

Bibliography

Afraz, S. R. and Cavanagh, P. (2008). Retinotopy of the face aftereffec. *Vision research*, 48(1):42–54.

Aggarwal, J. and Ryoo, M. (1986). Human activity analysis: A review. *G-Animal's Journal*.

Baluch, F. and Itti, L. (2011). Mechanisms of top-down attention. *Trends in Neurosciences*, 34(4):210–224.

Barnes, J. and Hut, P. (1986). A hierarchical o(n log n) force-calculation algorithm. *Nature*, 324(4):446–449.

Bauer, F., Cheadle, S. W., Parton, A., Muller, H., and Usher, M. (2009). Gamma flicker triggers attentional selection without awareness. *Proceedings of the National Academy of Sciences*, 106(5):1666–1671.

Bruce, N. and Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9:1–24.

Chun, M. M. and Marois, R. (2002). The dark side of visual attention. *Current Opinion in Neurobiology*, 12(2). 184-189.

Ehinger, K., Hidalgo-Sotelo, B., Torralba, A., and Oliva, A. (2009). Modeling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, 17:945.

Foulsham, T., Walker, E., and Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision research*, 51(17):1920–1931.

Heinke, D. and Humphreys, G. (2003). Attention, spatial representation, and visual neglect: Simulating emergent attention and spatial memory in the selective attention for identification model (saim). *Psychological Review*, 110(1):29–87.

Hulleman, J. (2013). The effect of task difficulty on visual search strategy. *Journal of Vision*, 13:686–686.

Hunziker, H.-W. (2006). *Im Auge des Lesers: foveale und periphere Wahrnehmung - vom Buchstabieren zur Lesefreude*. Transmedia Stubli Verlag Zurich. ISBN 978-3-7266-0068-6.

J. E. Hummel, I. B. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, 99:480–517.

Kassner, M. and Patera, W. (2012). Pupil: Constructing the space of visual attention. Master's thesis, MIT.

Kranz, J. H. (2012). *Chapter 3: The Stimulus and Anatomy of the Visual System*. Experiencing Sensation and Perception. Pearson Education. ISBN 978-0-13-097793-9.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., New York.

Miller, J. (1991). The flanker compatibility effect as a function of visual angle, attentional focus, visual transients, and perceptual load. *Perception and Psychophysics*, 49(3):270–288.

Niebur, E. (2009). Temporal tagging of attended objects. *Proceedings of the National Academy of Sciences*, 106:2479–2480.

Nuthmann, A., Smith, T., Engbert, R., and Henderson, J. M. (2010). Crisp: A computational model of fixation durations in scene viewing. *Psychological Review*, 117:382–405.

Otasevic, N. (2013). Recognizing simple human actions by exploiting regularities in pose sequences. Master's thesis, MIT.

Postma, E. O., Herik, H. J. V. D., and Hudson., P. T. (1997). Scan: A scalable model of attentional selection. *Neural Networks*, 10(3):993–1015.

Prinzmetal, W., Henderson, D., and Ivry, R. (1995). Loosening the constraints on illusory conjunctions: the role of exposure duration and attention. *Journal of Experimental Psychology: Human Perception and Performance*, 21(6):1362–1375.

Rao, S. (1998). *Visual Routines and Attention*. PhD thesis, Massachusetts Institute of Technology.

Schneider, W. and Shiffrin, R. M. (1977). Controlled and automatic human information processing: 1. detection, search, and attention. *Psychological Review*, 84:1–66.

Shimojo, S., Miyauchi, S., and Hikosaka, O. (1992). Voluntary and involuntary attention directed by the line-motion effect. *Perception*, 22.

Sun, H., Wang, C., and Wang, B. (2011). Hybrid generative-discriminative human action recognition by combining spatiotemporal words with supervised topic models. *Optical Engineering*, 50:027203–027203.

Tatler, B., Hayhoe, M., Land, M., and Ballard, D. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, 11(5).

Tatler, B. W., Baddeley, R. J., and Vincent, B. T. (2006). The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision research*, 46(12):1875–1862.

Treisman, A. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136.

Tsotsos, J. K. (1989). The complexity of perceptual search tasks. *Proc. Int. J. Conf. Artif. Intell.*, page 15711577.

Tsotsos, J. K., Rodriguez-Sanchez, A., Rothenstein, A., and Simine, E. (2008). Different binding strategies for the different stages of visual recognition. *Brain Research*, 1225:119–132.

Tsotsos, J. K. and Rothenstein, A. (2011). Computational models of visual attention. *Scholarpedia*, 6(6201).

Uhr, L. (1972). Layered 'recognition cone' networks that preprocess, classify and describe. *IEEE Transactions on Computers*, pages 758–768.

Ullman, S. (1984). Visual routines. *Cognition*, 18. 97-159.

Ullman, S., Harari, D., and Dorfman, N. (2012). From simple innate biases to complex visual concepts. *Proceedings of the National Academy of Sciences*, 109(44):18215–18220.

Ullman, S., Vidal-Naquet, M., and Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature neuroscience*, 5(7):682–687.

Walther, D. and Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks 19*. 1395-1407.

Wandell, B. A. (1995). *Foundations of Vision*, chapter 3. Sinauer Associates Inc, Reading, Massachusetts. ISBN-10: 0878938532.

Wischnewski, M., Steil, J. J., Kehrer, L., and Schneider, W. X. (1973b1973). *Integrating inhomogeneous processing and proto-object formation in a computational model of visual attention. In Human centered robot systems*, chapter 1.2, pages 93–102. Springer Berlin Heidelberg.

Wolfe, J. (2000). *Visual attention*, pages 335–386. De Valois KK. Academic Press, San Diego, CA, 2nd edition.

Yamato, J., Ohya, J., and Ishii, K. (1992). Recognizing human action in time-sequential images using hidden markov model. in computer vision and pattern recognition. *Proceedings CVPR'9,1992 IEEE Computer Society Conference*, pages 379–385.