ABSTRACT

Subjects performing visual target tracking tasks have been shown to utilize perceptual organization. This organization has been shown to have both Gestalt features and goal-oriented features. Previous studies have attempted to use memory recall techniques to examine potential cognitive groupings in air traffic control tasks, with negative results.

Analysis of eye movements has shown similar patterns of organization to the underlying visual tasks. Experiments were performed to evaluate whether recall or eye-tracking techniques can be used to extract perceptual groupings. Subjects’ memory of scenario information is generally poor, except where significant manipulation of targets occurred. For this reason it is suggested that recall techniques may not be able to elicit subjects’ cognitive groupings. Fixation data, however, indicates clustering consistent with Gestalt factors. Goal-oriented factors did not seem to affect grouping.

Thesis Supervisor: Thomas Sheridan

Title: Professor of Engineering and Applied Psychology Emeritus
TABLE OF CONTENTS

Acknowledgments.................................................................................................................. 9
Perceptual concepts in air traffic control .............................................................................. 11
  1.1 Introduction.................................................................................................................. 11
  1.2 Perception................................................................................................................... 11
    1.2.1 Visual perception................................................................................................. 12
  1.2.2 Visual perception in air traffic control................................................................. 13
  1.3 Attention..................................................................................................................... 14
  1.4 Memory...................................................................................................................... 16
    1.4.1 Short-term memory............................................................................................. 16
    1.4.2 Working memory............................................................................................... 17
    1.4.3 Long-term memory............................................................................................ 19
    1.4.4 Episodic memory............................................................................................... 19
    1.4.5 Semantic memory............................................................................................... 20
  1.5 Situational awareness and mental models..................................................................... 21
Perceptual grouping concepts ................................................................................................. 23
  2.1 Introduction.................................................................................................................. 23
  2.2 Object vs. space-based theories of attention............................................................. 24
  2.3 Bottom-up vs. top-down............................................................................................. 25
    2.3.1 Gestalt psychology............................................................................................... 26
  2.4 Perceptual group formation.......................................................................................... 28
Eye movements....................................................................................................................... 31
  3.1 Introduction.................................................................................................................. 31
  3.2 Eye movement basics.................................................................................................. 31
  3.3 Scanpath...................................................................................................................... 32
Supervisory control concepts................................................................................................ 35
  4.1 Introduction.................................................................................................................. 35
  4.2 Workload and complexity........................................................................................... 37
  4.3 ROC analysis............................................................................................................. 38
Target tracking experiments .................................................................................................. 39
  5.1 Introduction.................................................................................................................. 39
  5.2 Experiment 1............................................................................................................. 39
    5.2.1 Setup and design................................................................................................. 39
    5.2.2 Results................................................................................................................ 42
    5.2.3 Experiment 1 – discussion.................................................................................. 46
  5.3 Experiment 2............................................................................................................. 47
    5.3.1 Setup and design................................................................................................. 47
    5.3.1.1 Eye movement monitoring system............................................................... 48
    5.3.1.2 Yufik software............................................................................................... 50
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Spotlight locus of attention.</td>
<td>24</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Factor of proximity (Wertheimer, 1923)</td>
<td>26</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Factor of similarity (Wertheimer, 1923)</td>
<td>26</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Kanisza figure.</td>
<td>27</td>
</tr>
<tr>
<td>Figure 5</td>
<td>More Kanisza figures.</td>
<td>27</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Scanpath (Privatera, 1999).</td>
<td>32</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Ambiguous figure.</td>
<td>33</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Mental and physical processes required in ATC.</td>
<td>36</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Experiment 1 screen shot.</td>
<td>40</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Definitions</td>
<td>42</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Correct detection rate</td>
<td>43</td>
</tr>
<tr>
<td>Figure 12</td>
<td>False alarm rate</td>
<td>43</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Detection lead times vs. trial number</td>
<td>44</td>
</tr>
<tr>
<td>Figure 14</td>
<td>False alarm rate vs. detection lead time</td>
<td>45</td>
</tr>
<tr>
<td>Figure 15</td>
<td>ROC for target tracking task</td>
<td>46</td>
</tr>
<tr>
<td>Figure 16</td>
<td>Missed acceptance counts.</td>
<td>53</td>
</tr>
<tr>
<td>Figure 17</td>
<td>Acceptance lag times.</td>
<td>54</td>
</tr>
<tr>
<td>Figure 18</td>
<td>ROC curve</td>
<td>55</td>
</tr>
<tr>
<td>Figure 19</td>
<td>Specific recall performance</td>
<td>56</td>
</tr>
<tr>
<td>Figure 20</td>
<td>Raw workload averages</td>
<td>57</td>
</tr>
<tr>
<td>Figure 21</td>
<td>Fixations vs. recall</td>
<td>58</td>
</tr>
<tr>
<td>Figure 22</td>
<td>Fixation categories.</td>
<td>59</td>
</tr>
<tr>
<td>Figure 23</td>
<td>Transitions between pairs</td>
<td>59</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

The author wishes to thank Prof. Sheridan for his guidance and support. It has been a privilege to work with and learn from someone of such great experience and insight. I also wish to thank Dr. Yan Yufik for his contributions to this research. His enthusiasm and contributions have been invaluable.

Thanks also goes to Sandy Lozito, whose sponsorship through NASA/Ames grant number NAG 2-729, is the reason this research exists at all.

Personal thanks go to all the professors I have learned from while here at M.I.T., including R. John Hansman, Jr., Jim Kuchar and Amadeo Odoni.

Finally I wish to thank my family, my friends and fellow students, including Michael Mashtare, Dr. Mary Christy, Todd Farley, Debbie Hyams, Lee Winder and all the rest of the ICAT group, as well as my fellow HMSLers.
Chapter 1

PERCEPTUAL CONCEPTS IN AIR TRAFFIC CONTROL

1.1 Introduction

Studying the concepts of cognitive psychology is essential to understanding not only why the air traffic control (ATC) system succeeds despite a very high degree of task difficulty and virtually no automation, but also how to safeguard that performance in the face of increasing automation and free flight concepts. There is still considerable debate about the nature of perception, attention, memory and comprehension. Models have been developed and tested which although not definitive, provide enough predictive capability to examine where cognitive limits are being pushed in a task such as air traffic control. This section will introduce these concepts, and show how they apply to ATC.

The experiments that form the basis of this thesis dealt only with the visual aspects of a target-tracking task like air traffic control. For this reason, the following discussion will focus on that aspect of perception. ATC is, however, both a visual and an aural task. Voice communications form at least as big a part of the task of the controller as the visual aspects. Not only do more errors originate from voice communication difficulties, but also the controller's situational awareness would be compromised without the information received in this manner.

The aural aspects of ATC are better understood than the visual aspects, however. Furthermore, many of the proposed changes to ATC will impact the displays the controllers use. There have been no conclusive studies on how controllers are able to perform their job, nor on what circumstances lead to errors. This deficit needs to be remedied in order for designers to be able to predict the impact of proposed changes.

1.2 Perception

The physical aspects of visual and aural sensation are known. Once the stimuli leave the sensory organ involved, however, there is little understanding. This is the area of concern, "the way in which we interpret the information gathered
(and processed) by the senses. In a word, we sense the presence of a stimulus, but we perceive what it is” (Levine and Sheffner, p.1).

This is not to say there are not sensory detection problems in ATC. Digits on the radar scope are unclear, the display is generally cluttered, and communications are garbled. These problems are mostly understood, however, and often easier to tackle than perception problems, given proper resources.

1.2.1 Visual perception

The message that reaches the visual cortex is a processed and summarized record of the original stimulus. The most important information (edges, contours, changes) is transmitted, while the steady-state information is omitted. This information is taken in during discrete fixations, interrupted by sweeps of the eye called saccades. The movement is fast (about 30-100 msec), although it takes about 200 msec for a fixation. Normal visual processes are mostly suppressed during saccades, as information received during an eye sweep would be blurry. Given these times, people are capable of about three or four visual cycles per second, although far less time is required to actually perceive visual scenes.

This feature can be demonstrated by examining our perception of lightning. Bolts of lightning generally consist of three or four component bolts of about one msec each, with a separation of about 50 msec, for a total of about 200 msec. How is it that we can examine a scene illuminated by lightning, and why does the illumination appear to fade instead of dissipating instantly? This is due to “visual persistence”, the apparent persistence of the stimulation beyond its actual duration. Visual stimuli must be held in a memory register briefly. This register is referred to as “visual sensory memory.”

Sperling (1963) tested the capacity and duration of this storage. Subjects were presented arrays of letters and digits for very brief durations (5 to 500 msec), then asked to report what they remembered from the display. Sperling found that four or five items were reported correctly.

Sperling then showed that more items were available but faded from visual sensory memory before they could be retrieved. Given three rows of four items each presented again very briefly, subjects could retrieve three out of four items on a row when prompted for a particular row after the display went off. Other tests showed that this “iconic” memory is erased by subsequent visual stimulation, such as a new icon or a very bright screen. This ensure was dubbed “interference,” a term to which we will refer several times in this paper.
Once the image is observed and placed in visual sensory memory, it must be comprehended. The characteristics of comprehension are somewhat unclear. Reading is typically used as an example. "Template" matching (where one matches a template of what a particular letter looks like) and feature matching (where one matches the major features of a letter) arguments do not capture all aspects of recognition. Consider that when one reads sentence, one often fails to detect misspelled words (or even missing words, as is the case in the preceding clause - a missing "a" before "sentence"). Context and expectation significantly influence perception. This interpretation process is automatic, and is nearly indistinguishable from actual sensation.

1.2.2 Visual perception in air traffic control

As mentioned above, people are able to perceive large amounts of information in a very brief period of time. The data on the scope or the flight progress strip do not have to be present, or fixated upon, for very long for the information to be perceived. However, the presence of continuous interference in the form of new stimuli can prevent all this information from being interpreted and committed to long-term memory.

This is particularly important to our application. A radar display contains symbolic information (the aircraft target), spatial information (sector features, direction of flight, motion of the targets) and text information (identification, speed, altitude, etc.). Given brief glances at the target information, our understanding of memory suggests that controllers could not transfer information about aircraft to long-term memory. If consecutive fixations are very brief (<200 msec), it is unlikely that the controller can even identify what has been seen, since the iconic memory is being erased by interference from the succeeding fixation.

Controllers scan the scope, adding features of relevant information to their "picture" (also known as their mental model) of the traffic. The controller will use deep knowledge of the sector, aircraft characteristics and typical intentions to very quickly understand the situation. The scope appears hopelessly complicated to untrained eyes, but controllers are able to discern features of importance with a short viewing. The perception of this information is almost completely automatic for the controller.

Subjects performing target tracking scan the scope, briefly fixating upon targets and/or target information. The information must be passed into short-term memory in order to be interpreted. I shall come back to what happens in short-term memory in a succeeding section.
A large number of ATC errors are associated with misperceiving data. Some of this is due to attention, which will be discussed below, but many errors are due to applying context inappropriately. Controllers may interpret information incorrectly because of expectations, either failing to perceive the correct data, or inserting data that is not present. An example of this might be a slowed reaction time to detection of an overshooting final, or not initially noticing that an aircraft has continued ascent beyond its assigned altitude.

1.3 Attention

An item missing from the analysis above is attention. It is apparent that a great deal of information reaches the brain only to be discarded if not considered pertinent. This discrimination process is attention. An important result of the research in placing the attention process is that most of the limitations of processing associated with attention occur after the information is accrued in memory.

Early theories of attention (Welford, 1960; Broadbent, 1958) located the process between perception and memory. Later evidence placed attention after the accrual of information in memory, and these seem to hold up better under experimental scrutiny. Numerous experiments demonstrated a widespread and parallel access to memory systems by stimuli, with attention being the control of this information in memory (Keele and Neill, p. 41).

There are several results worth noting. In the first, Morton (1969) had people sort cards on the basis of the numerals 1 to 6, or numerosity of X marks ranging from one X to six Xs, or “redundant” numerals and numerosity (one 1, two 2s, etc). Subjects sorted faster in the redundant task than in either component alone. This finding was confirmed and expanded upon by Ellis and Chase (1971). Redundant information is perceived in parallel, and begins to accrue in memory. This accrual occurs faster given redundant information, resulting in reduced error, reaction time, or both.

The second experiment of note concerns irrelevant information. Numerous studies (Morgan and Alluisi, 1967; Well, 1971; Keele, 1972; and Egeth et al., 1972) demonstrated that some irrelevant information is processed to the memonal stage, but is discarded. Several studies (Garner, 1974; Kahneman, 1973; Lockhead, 1972) have shown that of the irrelevant information that is remembered, only irrelevant information that is integral to the stimulus affects accuracy, reaction time and recall. For instance, if the X-counting experiment above used complex faces instead of X’s, reaction time would be slower than if normal X marks were used but a face were imprinted on the background of the task. The shape of the marks, although not important to the task, is integral to
the stimulus, while a separate background image is not. Integral information, even if irrelevant, is passed along to memory, while nonintegral information is not. This suggests that if the data tag contains information about an aircraft that is not needed to perform the task, its mere presence will affect performance, whereas the presence of other information on the scope display (such as nav aids not involved with the aircraft) will not.

Another result of interest is that of activation (Posner and Snyder, 1975; Neely, 1977). Neely had subjects decide if two words, one shown after the other, were the same or different. The words were preceded by a priming word. Neely told his subjects that if bird appeared, to expect a word like “robin” or “sparrow” to appear next. If the word building appeared, the subject should expect a body part, like “arm” or “leg.” The results showed that only when the prime was bird was there a benefit to priming.

This last result demonstrated that when a stimulus is received, information in memory associated with that stimulus is activated automatically. Stimuli received after this activation and associated with the activated memory benefit from better reaction time and accuracy. The converse is also true. Reaction time will suffer should a stimulus not associated with the activated memory be received (i.e. an unexpected signal).

The concept of automaticity is directly related to attention as well. Automatic processes, such as reading familiar words, hearing one’s name, etc. consume no conscious resources. Those resources can be used to perform other tasks. Numerous automatic processes can be carried on at one time; walking while whistling is an example.

Conscious processes, on the other hand, require attention. They are either too unfamiliar or too complex to be automatic, and generally consume most of the available resources in the cognitive system. Multiple conscious processes can only occur simultaneously if they are very simple, although one can perform a number of automatic processes along with a conscious one (listening to the news on the radio while operating the directional signals, perhaps chewing gum and keeping your car on the road).

If all processes were automatic, there would be substantially less cognitive limitation to human processing. There are ways to make conscious processes automatic, however. This is obvious, since walking was once a conscious process for us all as children, yet now it is automatic. Practice is what transforms conscious processes into automatic ones. Humans can be taught extremely complex activities with practice (e.g. air traffic control). Many of the activities of
the controller are highly automated, one reason why controllers are fiercely protective of their procedures and displays.

1.4 Memory

1.4.1 Short-term memory

Miller (1956) is one of many researchers who showed that the span of immediate memory for a single stimulus dimension is about 7 items in length. For multiple stimulus dimensions the number is much larger. We can perceive large quantities of sensations, and we can hold vast amounts of information in our long-term memory, but immediate memory is the bottleneck in this information processing system.

Items can be grouped or "chunked" to stretch the limitation. In fact, given practice and a clever encoding scheme, people have been able to recall large numbers in order – 82 digits in work done by Chase and Ericsson (1982). This type of encoding is done with reference to very familiar (automatic) knowledge accrued in long-term memory. This concept is central to the experiments that are the subject of this thesis, and will be discussed in greater detail in a succeeding section.

Given enough time, almost anyone could memorize an 82-digit number. Often there is insufficient time or attention to apply a scheme to transfer the item to long-term memory. The natural assumption was that time, in the form of memory decay, was the significant factor. Waugh and Norman (1965), however, showed that interference was the primary factor, although the decay theory is difficult to test. Forgetting is strongly influenced by intervening items prior to transference to long-term memory.

Interestingly, Keppel and Underwood (1962) and Wickens (1972) also showed that previous trials in a remembering task influenced recall in current trials, a kind of proactive interference. New learning was disrupted by old knowledge. Another example of this is being taught that in Italian bathrooms “C” means “caldo” which means “hot”, but forgetting and turning the wrong tap. If it were a completely different letter it is unlikely one would make the same mistake with the same frequency.

Recall that attentive processes occur after the accrual of information in memory. The ability to process this information is then affected by redundancy, interference, activation, and automaticity. Understanding the expertise of the controller depends on our understanding of these concepts. For example,
checking in on a new frequency and having the aircraft “ident” their data tag on
the scope is likely to have other benefits than just confirming identification and
functionality. There are redundancy effects as well. The continuous interference
of the numerous call signs, all of similar form, and of other information (altitude,
speed, heading) may explain why it is difficult for controllers to recall that
information (as seen in Endsley & Rodgers, 1996). Alternatively, it may explain
why the form of the task has been developed to ensure that recall of this
information is not necessary to its successful completion. Activation and the
automaticity of many of the highly practiced tasks explain the speed with which
controllers perform.

This does not clarify the purpose of short-term memory, however. To
understand this, we need a broader model of short-term memory. Such a
concept was refined by Baddeley and Hitch (1974).

1.4.2 Working memory

The work by Baddeley and Hitch refined a concept that has been studied since
about 1905, although the term “working memory” was not introduced until the
late 1960s. Baddeley and Hitch attributed to working memory three systems: a
“central executive”, a “phonemic loop”, and a “visuospatial scratch pad.”

The evidence for the phonemic loop (or “articulatory rehearsal loop” or
“phonological loop”) was discovered when they observed that a concurrent
memory load of three items did not impair free recall, comprehension, or
reasoning tasks in subjects. The suggestion was that this process could be
shunted to a separate process, a rehearsal loop that could store limited amounts
of speech-like information. Note that the information does not have to be
auditory, just speech-like. Other studies showed that visually presented verbal
material could not be rehearsed (and hence not remembered) if the subject
repeated out loud a word such as “the the the” or “hiya hiya hiya.” It was also
argued that the capacity was limited by how much a subject could verbalize in 2
seconds, seemingly the rate of decay of information from this system (Baddeley,
Thomson, and Buchanan, 1975).

Salame and Baddeley (1982) proposed a further segmenting of this feature
into a phonological store and a subvocal rehearsal loop. This explained the two-
second verbalization limitation. Auditory stimuli would receive privileged access
to the store, while visually presented verbal stimuli would be accessed through
the rehearsal loop. Subsequent research by Baddeley (Baddeley et al., 1984;
Baddeley, 1986; Vallar & Baddeley, 1984) supported these results. The two-
second limitation was a limitation of the phonological store, but could be
refreshed from the articulatory loop, if available. This loop, however, is
unavailable when suppressed by articulation. So while we can hear and see while
we are talking, we cannot rehearse the verbal information and therefore cannot retain it for long.

The visuospatial system within working memory has been shown to have a similar structure. Experiments done by numerous researchers (Baddeley, Grant, Wight, & Thomson, 1975; Baddeley and Lieberman, 1980; Morris, 1987; Quinn & Ralston, 1986; Smyth, Pearson, & Pendleton, 1988; Beech, 1984; Logie, 1986; Quinn, 1988) confirmed the idea put forward by Baddeley and Hitch (1974). Overall, this body of work appears to demonstrate that the visuospatial system in working memory consists of a short-term store that is linked directly to the processes of visual perception, but that can be refreshed by spatial rehearsal. Irrelevant physical movements, such as waving arms or walking in place can suppress this rehearsal.

The central executive is the last component of this model of working memory. This component has been shown to have no storage capacity of its own, but instead serves as a “general attentional resource that (coordinates) the contributions of different storage subsystems” (Richardson, p. 22).

Some of the experiments to show this had subjects generate a random sequence of letters. The results of these are described by Baddeley (1990):

After about the first 15 or 20 letters, most people find the task becoming increasingly difficult, with the same few letters tending to crop up, and with a tendency for sequences to follow stereotyped patterns such as the alphabet, or familiar acronyms such as CIA, VD and BBC … If one systematically varies the rate at which the subject is required to generate letters, then a very lawful pattern emerges, with the randomness increasing with the logarithm of the time available … Another way of manipulating the task is to vary the number of alternatives, requiring the subject to generate on the basis of two, four, sixteen, or twenty-six letters. This leads to a systematic decrease in the rate at which letters are produced that levels of off about eight alternatives. This suggests that subjects can cope with up to about eight alternatives simultaneously, with smaller numbers of items allowing more attention and faster selection; once the system’s capacity has been reached, adding further alternatives will not affect performance since the system will still be operating on its maximum of seven or eight options.

This task requires some organization and planning, which was the suggested function of the central executive. Other experiments have used this result to occupy the central executive and show its contribution in certain tasks.
1.4.3 Long-term memory

From the previous discussion of working memory, it seems obvious that our long-term memory system is significantly different from working or short-term memory. Its duration is much longer, and its capacity is much higher, both essentially infinite. Yet often recall of this information is time consuming or impossible. The strongest evidence for separate systems, however, is that in studies from brain-damaged patients long-term and short-term memory systems are affected separately from one another.

Two main theories of long-term memory have arisen. The first distinguishes “episodic” memory and “semantic” memory. Episodic memory is simple knowledge, events and facts. Semantic memory is knowledge that relates concepts and information, rules about language and general world knowledge. A second distinguishes implicit and explicit knowledge. The latter is more recent, and the distinction between implicit and explicit is of little value to the application of long-term memory to air traffic control.

1.4.4 Episodic memory

One important feature of episodic memory is the effect of rehearsal. While the specifics of the rehearsal mechanism have been debated, the basic concepts remain, and are fairly intuitive. Items rehearsed repeatedly, and elaboratively (using the meaning of an item rather than its form), are recalled better than those not repeated, or repeated in a maintenance-type manner. This latter type of rehearsal is characterized by the item usually not being recalled once rehearsal ceases, such as repeating a phone number until it is dialed. Elaborative rehearsal establishes context for the item and relationships between it and other items in long-term memory.

Another aspect of episodic memory is retrieval. It has been shown that items in long-term memory do not suffer decay over time. Affecting recall, however, are congruity, interference, and distinctiveness. These things can make items inaccessible, but not lost from memory. The current view holds that items in long-term memory cannot be lost, they remain there permanently, except in cases of physical damage to, or incapacitation of, the brain.

Congruity refers to the heightened ability of people to recall things, if told to recall them under the same conditions they remembered them. A simple example is when someone is asked to list the alphabet. If done in order, remembering is fast and easy since that is the way they were taught. Recall is much slower and more prone to errors if done in any other order. Other experiments have shown that when shown items as pictures, those items are
recalled more easily when presented with pictorial cues, rather than word cues or other types of cues. This is also referred to as the effect of “mode”.

Distinctiveness has been shown to aid the recall of events as well. This refers to not only how remarkable a particular stimulus is, but to real-life dramatic events as well. Memory experts suggest that when we want to remember something, to think of a bizarre or distorted image of it, creating distinctiveness. Dramatic events, such as the Kennedy assassination and Challenger accident, have been shown to improve recall of information associated with the events.

1.4.5 Semantic memory

Episodic memory is insufficient, however, to accomplish human activity and understanding. If it were, computers would be able to perform all human function. The interrelation of information, the application of principles to situations, the extrapolation of meaning, in short the sum total of useful human function, requires semantic memory.

One attempt to understand semantic memory started back in the late 1960s, when Quillian (1969) constructed a semantic network, initially interested in teaching a computer to read a sentence. This idea of semantic memory as a network associating concepts to one another in memory has since been refined considerably, and has a great deal of merit in cognitive science. Baddeley (1990) suggests that it may be more of a modeling language than actually descriptive of physical memory. Despite concerns such as these, however, network techniques in cognitive modeling are extremely powerful and descriptive, and in many ways match the physical characteristics of the brain.

A competing model to Quillian’s network is referred to as Bartlett’s concept of schema. In 1932 Sir Frederic Bartlett proposed an interpretation of memory in which subjects used schemas to remember information. Little else was done with this until computers were developed to the point that they could help test Bartlett’s theory. Marvin Minsky (1975), Rumelhart (1975), and Schank (1975) developed this theory along similar lines.

Schemas are packets of information, larger in scope than the nodes in the network model discussed above. They also can contain a wide range of information, from concrete items to very abstract concepts. A schema represents knowledge rather than definitions. Schemas are actively applied to perception to accomplish understanding.

This is by no means a complete description of episodic nor semantic memory. Several important theories and features have been omitted. I have attempted to
concentrate on those theories and features of direct relevance to air traffic control, which will be discussed shortly.

1.5 Situational awareness and mental models

So far I have described the disparate elements of how memory operates. These elements can be readily seen when dealing with relatively simple recall. Memory is of course used in accomplishing more complicated activities as well. Normal activity such as carrying on conversations, driving a car, playing sports or games, flying and controlling aircraft are all examples.

In analyzing how humans successfully accomplish these activities, the terms “situational awareness” and “mental models” have arisen. The most commonly accepted definition of situational awareness comes from Endsley (1988), who says it is the “perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future.” A mental model is the representation of the task or object, including its properties, states and interrelationships with other objects or events. This model is within the mind of the subject, and is sometimes incomplete or inaccurate when compared with the actual task or object.

Situational awareness (SA) is a term that is used frequently in aviation. When a pilot or controller has good situational awareness, they have knowledge of all (or almost all) of the information required to make successful choices. One criticism of this concept is that the set of information required to make good choices can easily be considered nearly infinite. SA can be expanded to include nearly every element involved in a task, causing every error to be attributed to “poor SA”. This approach does not lead to fruitful solutions.

Mental models have been used longer, and for a broader class of phenomena, than SA. Subjects develop a mental representation of the task, and use this to understand the data being received from the task, and to formulate action. Everyone has a mental model of their home, and could probably walk through most of it with their eyes closed. Positions of furniture are generally (perhaps not exactly) known. If we hear a sound from another room we can generally guess what it might be. Sometimes we may hear a sound we don’t recognize, and upon finding it is coming from an open window, we add that sound (and its cause) to our mental model.

The two terms, then, overlap to some extent. Aviation researchers tend to use situational awareness, particularly when discussing a subject’s recall of the environment. Cognitive researchers tend to use mental models, with recall being separate from the model. My approach will be the latter, since the task
accomplished during the experiment lends itself more naturally to mental models and recall.
Chapter 2

PERCEPTUAL GROUPING CONCEPTS

2.1 Introduction

Miller demonstrated that the span of immediate memory for one-dimensional stimuli is $7\pm2$ items. One method to skirt that limitation is through clustering. Subjects in a study by Bousfield (1953) were asked to recall words from a long list. Rather than recall them in the order in which they were presented (or randomly), subjects recalled words in categories, such as “doctor, lawyer, fireman; fly, mosquito, spider; dog, cat, horse”. Numerous experiments confirmed that this tendency is a necessary condition for long-term storage of information, and better organization led to better recall. Even in experiments using words apparently unrelated to one another, subjects would group items in some fashion to aid in remembering the items.

This grouping of unrelated items can be done in numerous ways. Making up sentences that include the words, forming images including the items and other mnemonics are all useful. These strategies are presented as methods to improve memory and recall, and have demonstrated startling results. Subjects trained in such methods can perform remarkable feats of memory, such as memorizing the contents of a magazine in 10 minutes. In such a magazine-remembering demonstration, the subject could recall all the details, including content and location, of any page called out at random.

Chunking is not limited to a conscious application of grouping schemes. If asked to recall a 22-letter sequence given only 2 seconds, most people would doubt their ability to accomplish this task. However, if that sequence is “TAKEMEOUTTOTHEBALLGAME”, the task is a simple one. This type of chunking is automatic, using schemes that we have already internalized. The same can be done with any other type of information – a large set of lines that form the image of a house would be easier to recall than a random configuration.

It is not surprising, then, that grouping schemes have been found in a number of complex activities. Waiters and waitresses have been shown to have a specialized, automatic memorization scheme that helps them attribute orders to
spatial locations. Links were formed between orders and the spatial locations of patrons, allowing the waitperson to recall up to 8 orders (about 20 menu items). Chess players have also shown sophisticated yet internalized and automatic grouping strategies involving spatial relationships.

2.2 Object vs. space-based theories of attention

In order to understand perceptual organization, psychologists have examined how a person’s attention is focused when viewing an object. Two main views of the process of attention have developed.

One view holds that attention is like a spotlight. A person’s gaze tracks and holds on certain parts of a scene. Attention parses the perceived scene, and features within a certain radius of this focus are attended. Those outside are not. This locus of attention need not be “all-or-nothing”. Generally comprehension of features decreases as the distance from the focus of attention increases, rather than sharply dropping off at some predefined radius. However, any items within the locus of attention are all attended. Figure 1 shows a configuration of three letters. Generally it would be impossible (under object-based theories of attention) to attend to “A” and “C” without also attending to “B”.

![Figure 1. Spotlight locus of attention.](image)

Support for space-based theories comes from numerous experiments. Downing and Pinker (1985) showed that detection time for luminance changes increased as the distance from the focus of attention increased. Hoffman and
Nelson (1981) presented subjects with a four-letter display, with an open box superimposed on the image. The subjects were asked to determine which of two letters were not in the display, as well as the orientation of the box. Subjects were more accurate when the box was adjacent to the target letter. Similar support for space-based theories comes from LaBerge (1983), Podgorny and Shepard (1983) and Posner, Snyder and Davidson (1980).

In contrast to space-based attention is object-based attention. This theory assumes that attention is focused on one or more objects in the scene, regardless of location within the scene. Duncan (1984) superimposed two images, then asked subjects in one case to report two attributes of one object, or one attribute of each object in another case. Subjects were more accurate reporting two attributes of one object, despite both objects being within the locus of attention. Neisser and Becklen overlapped two sequences on videotape, then asked subjects to view the tapes and report any anomalous events in one or the other of the sequences. For example, if the video showed a pair of hands clapping superimposed on a group of people playing catch, an anomalous event would be one of the subjects leaving and returning. Subjects performed very well in this task, but rarely noticed anomalous events in the unattended sequence.

Ongoing research by Kanwisher (1999, personal conversation) has identified specific regions of the brain that activate when a face is presented. This region is activated only for faces, not for other objects such as houses. During this experiment subjects were presented with independent images in either eye. The “faciform” region of the brain would activate and deactivate as the subject’s attention switched between the two images, both of which were being received by the brain. Subjects reported the same phenomenon, that of one image switching with the other, similar to the effect of a Necker cube.

While these results do not completely contradict space-based theories, they do show that there is at least some object-based influence to attention.

2.3 Bottom-up vs. top-down

If perceptual objects influence our attention, what determines the form of the perceptual object? Perceptual organization has long been considered a purely stimulus-driven, bottom-up process. This means that organization is influenced by the image properties alone, regardless of the goals of the task or the relative importance of the features. Typical features regarded as important are proximity, similarity, common motion, symmetry and good form (“Prägnanz”).
2.3.1 Gëstalt psychology

One of the original Gëstalt psychologists was Max Wertheimer. He published a very influential paper in 1923 that laid the framework for the Gëstalt movement. Gëstaltists believe that we do not experience a number of things, but rather larger, interrelated wholes.

Wertheimer's original examples used dots. When we see a picture such as that shown below, it is natural to see the dots as being grouped. In fact, it can be difficult to see them any other way. This figure demonstrates what Wertheimer called the “Factor of Proximity.” In this case the dots close to one another are grouped.

![Factor of Proximity](image)

Figure 2. Factor of proximity (Wertheimer, 1923).

The second factor is the “Factor of Similarity”. If we take the same dots, space them equally, but change their color or shape, we again see grouping. One tends to group like objects together. This similarity factor is affected by the magnitude of dissimilarity as well. So if some of the dots were white, some black, and some shades of red, the shades of red would tend to be grouped together. If they were all shades of red, however, the different shades would be grouped separately.

![Factor of Similarity](image)

Figure 3. Factor of similarity (Wertheimer, 1923).

If we take a group of dots, and shift several of them suddenly, the grouping of those dots may change in response. If they were previously grouped by proximity, the dots may subsequently be grouped according to shift in position. This is referred to as the “Factor of Common Fate” or “Uniform Density”.

There are other factors as well. They all have one thing in common: certain arrangements are more compelling than other arrangements. Our perception of a scene is powerfully influenced by Gëstalt factors.
Kanizsa figures provide further evidence for the power of this grouping mechanism. We can see triangles and circles in figures 4 and 5 where none exist, because our brains fill in the missing information. When creating a sphere on a computer screen, the three-dimensional look is supplied by subtle and continuous shading. The human viewing this sphere would not see the continuous color shading that actually exists, but rather small polygons of equal shading separated by dark lines. These dark lines are our mind inserting edges where none exist; this illusion is well known within the discipline of computer graphics.

![Kanizsa figure](image)

Figure 4. Kanisza figure.

![More Kanizsa figures](image)

Figure 5. More Kanisza figures.

Gestalt factors are, therefore, very commonly used in imparting structure on a visual task. In many cases they are automatic, and difficult or impossible to alter.

Some of the factors that regulate grouping are top-down or imposed somehow on the stimulus. Wertheimer included “past experience or habit” as one of the factors involved in grouping. Through training or experience, subjects could come to see patterns contrary to the Gestalt factors mentioned above. Proshansky and Murphy (1942) showed that subjects could change their perception of lines when the punishment/reward system was changed. There has been a large body of work following along the lines of the famous Pavlov’s dog
experiment that showed subjects, when shown a pair of stimuli together frequently enough, would actually see both stimuli when presented with only one.

Bruner (1947) showed that social and cultural factors, and even age, impacts perception of organization. Bruner (1949) also showed that while incongruity factors (such as playing cards where the suit colors were reversed) caused an increase in reaction time, once the subject expected incongruity this delay was eliminated.

The fact that one can intentionally switch attention when viewing a Necker cube is another example of this. Given a goal and proper training, perceptual organization can be influenced.

2.4 Perceptual group formation

Pylyshyn and Storm (1988) had subjects visually track a small number of target objects moving about a display among a number of identical non-target objects. At various times an object would flash and the subject would have to identify it as either target or non-target. Despite this task being highly demanding, subjects correctly identified over 90% of the objects. Further experiments showed that this result was inconsistent with the “spotlight” theory of attention. The conclusion was that the targets were being cognitively “indexed”, allowing tracking independently of the visual task.

Yantis (1992) proposed that the targets are attended, but that they are grouped into a virtual object. It is this virtual object which is tracked. A similar experiment was performed, but the ability to group the targets into a virtual object was varied.

Previous work had determined that formation of virtual objects is a two-step process. The initial formation of the virtual object is stimulus-driven, controlled by the same bottom-up factors mentioned earlier (proximity, similarity, common motion, symmetry, Pragnanz). Once the object is formed, it must be maintained by reviewing the configuration of elements and updating the object if necessary.

Yantis first varied the factors that influence formation. Significant improvement in the task occurred, but only in the early trials. Subjects eventually learned to efficiently group targets, even when they were configured randomly. A next set of experiments varied the factors that control maintenance of the virtual object. Performance improved when maintenance was simplified. These results validate the earlier suggestion that perceptual grouping can be influenced by the objectives of the task. Subjects were able to track a virtual object from within a random configuration of identical elements.
ATC is in some respects a target-tracking task. As such perceptual grouping should play a significant role. Means et al. (1988) found evidence that aircraft are grouped in memory with related aircraft. Other research has suggested aircraft are recalled with respect to the sector, indicating a space-based attentional scheme.

Dougherty et al. (1997) conducted experiments to test the source of groupings in ATC. Controllers were run through an air traffic simulator, which was stopped at random intervals. Subjects then attempted to recreate the position and information of the targets on a map.

To demonstrate the presence of groupings, the subjects were videotaped performing the recall task to obtain timing information. Previous cognitive research has shown that if items are perceptually grouped, recall of one item facilitates recall of the other members of the group, resulting in decreased recall times. An example of this concept is if one were asked to list all possible first names. It is likely that the names of friends and acquaintances would be recalled in rapid succession, followed by a pause as other groups of people are recalled.

No such clusters appeared for the controllers in this task, suggesting that no aircraft-to-aircraft links exist. Although accuracy of recall was not reported for this experiment, it has been shown to be excellent in previous experiments. Controllers placed 95% of the aircraft in their sector within 10 nautical miles of their actual position in an experiment by Means et al. (1988). Recall was 90% in a study by Gronlund, Dougherty et al. (1997), most placed within 2.5 cm of their actual position. Gronlund, Ohrt, et al. (1997) found a 79.6% recall rate, with an average miss distance of about 9.6 nautical miles. Controllers reported an average of 8 aircraft (out of an average of 12.8 visible) with a mean distance error of 9.6 miles in an experiment by Endsley and Rogers (1996).

Since there were on average 14 aircraft in the sector in most of the aforementioned experiments, it is unlikely that this recall is due to strict memorization. Furthermore, the normal workloads of a controller would hinder memorization. An alternate suggestion is that there are aircraft to sector links that guide retrieval, although no suggestions for how this mechanism might work have been forwarded as yet.

One difficulty in using recall information to determine perceptual grouping in ATC tasks is the inability of subjects to remember the details being asked. In Means et al. (1988), controllers could associate aircraft types to call sign with a 28% accuracy. For ground speeds, recall was only 6%. Subjects in the Gronlund,
Ohrt, et al. (1998) study recalled 56.5% of the altitudes and 20.6% of the speeds of traffic in their sector. Endsley and Rogers' (1996) subjects identified the numerical part of the callsign with 38.4% accuracy, the correct altitude for 59.7% of the aircraft, groundspeed for 28%, correct heading for 48.4% and reported whether the aircraft was in a turn with only 35.1% accuracy. However, controllers reported correctly 73.8% of the initial alphabetical part of the aircraft callsign (which indicates the operating company) and reported the control level (whether the aircraft was in their sector, about to be in their sector, etc.) also with 73.8% accuracy.

The better accuracy for altitude, control level and at least part of the call sign suggests that there are some details controllers recall better than others. Yet these details are all found in the data tag, at or near the iconic representation of the aircraft. If spotlight theories of attention are correct, there should be little or no difference in recall between these pieces of information. There is, therefore, some object-based attentional selection occurring.
Chapter 3

EYE MOVEMENTS

3.1 Introduction

From the results of Dougherty at al. (1997), it was found that recall-timing tests normally used to determine perceptual groups were ineffective. The finding that certain pieces of information were recalled fairly well, while others were not recalled suggests, however, that there is some clustering of the targets. An alternative method of determining conflicts is to examine eye movements. As will be shown, fixations on targets are, in the words of Stark & Choi (1996), “repetitive and idiosyncratic to a particular picture and to a particular subject.”

3.2 Eye movement basics

Eye movements can be separated into a sequence of fixations and the movements between those fixations, called saccades. Most information is perceived from foveal vision, high-resolution (approximately one-half arc-minute) vision with a small field of view (approximately one-half degree). Peripheral vision, on the other hand, has 120 to 180 degrees field of view, but is sensitive only to flickering lights and motion.

Each aspect of vision can transmit about 40K bits per second. Foveal vision has fewer pixels but a (relatively) high bits per pixel, while peripheral vision has many pixels but generally only one bit per pixel. Given these limits, in order for a person to perceive a complex scene and extract information from it, a sequence of fixations and saccades must occur. A normal scene contains items of interest areas of little interest or content. Some items of interest must be viewed for a period of time to process the level of detail required.

Each of these fixations imparts a piece of the scene to the brain. This is a rather sparse sampling of the information contained in the entire visual scene. In the lightning example given earlier, a very brief exposure to a scene still provides a sense of visual completeness to the subject. The subject, and each of us in everyday life, sees the “whole” world before us, even though only a small fraction of the information available reaches our brain. Stark & Choi (1996) refers to this as the “illusion of completeness”.

31
3.3 Scanpath

If only a small part of the scene is actually viewed, there must be a mechanism for selecting what parts of the scene to view. One would expect to find the eye fixating on some area of interest, move to another, then another, and so on until as much as possible of the scene is viewed. This turns out not to be the case.

Noton and Stark (1971) showed that these sequences of fixations were "idiosyncratic" and "repetitive". Subjects' eye movements when viewing a scene showed repeated fixations on particular items of interest, and a pattern for revisiting those items which was different for different subjects, but generally consistent for an individual subject. An example of a scanpath is seen in figure 6. The white ovals are fixations, the white lines are saccades. If the same subject viewed this picture again, or a similar picture, the scanpath would be very similar. If a different subject viewed it, the same features would be viewed, but the scanpath would be different.

Figure 6. Scanpath (Privatera, 1999).
One opinion holds that features of the stimulus control fixations. Our eye is drawn to areas of contrast or some other objective criteria that is distinct from the object as a whole, or the goal of our viewing. The scanpath theory holds that an internal model of the object directs fixations. Fixations are a method of comparing and updating the model. The "illusion of completeness" that we have of the world being viewed is formed from the internal model and the sampling achieved by the scanpath.

Further support for the scanpath theory comes from viewing of ambiguous figures. Ambiguous figures are pictures which "switch" from one picture to another when viewed. Ellis and Stark (1978) showed that when attention shifted from one figure to another (such as from the vase to the opposing faces in figure 7), the scanpath changed as well.

![Figure 7. Ambiguous figure.](image)

The most convincing evidence comes from Brandt et al. (1989). Subjects viewed three different patterns of Xs, then were first asked to visualize the patterns on a blank screen, then draw the patterns on a blank screen. Similarity analysis of the visualized object's scanpath showed strong agreement with the scanpath of the viewed patterns. No stimulus was available to the eye or brain when visualizing, yet nearly the same scanpath resulted, confirming that a top-down cognitive model was the source.
Chapter 4

SUPERVISORY CONTROL CONCEPTS

4.1 Introduction

In a strict sense, air traffic control falls in the realm of manual control. No control loops are closed by automation, so that if the controller walks away from the scope, all control actions stop. There is significant processing of data by the computers, but they perform no actions on the aircraft. In a looser sense, however, the controller does act in a supervisory control role. Sheridan (1992) defined five basic supervisory functions as:

1. planning
2. teaching/programming
3. monitoring/detect failures
4. intervening
5. learning from experience

While these were meant to refer to automation (in the cases of 2, 3, and 4), they can equally apply to a process. Figure 8 is taken from a report on air traffic control complexity (Wyndemere, 1996). Note the similarities between this figure and Sheridan's five basic functions.
An analogy can be made between the teaching/programming function and the controller's sometimes-subtle interaction with other agents in the air traffic control system. In performing his/her duties, the controller informs (often implicitly) the other agents what the plan is for each aircraft to arrive at their goal state (landing at an airport, climbing/descending, traversing a sector to a fix, etc.). The controller automates the process of controlling to some extent by distributing information (either directly through instructions or comments, or via the fact that much of the information is shared through the party-line nature of the communications system).

The controller does not have control-performing automation to monitor. The conflict detection process, and the need for the controller to monitor compliance to instructions, can be substituted for this however. Controllers issue instructions, require the pilot to read back the instructions, and monitor the pilot's compliance to those instructions.
The controller also does not need to intervene in the sense of supplementing or taking over automatic control activities. The controller does need, however, to intervene in the normal sequence of events to prevent conflicts or increase efficiency.

So it is in this sense that we apply supervisory control concepts to air traffic control. Although there is no control-producing automation (at least not yet), the human interacts with the process in much the same way as a human operator interacts with automation. The analyses normally applied to automation can be applied to the human.

4.2 Workload and complexity

In describing the difficulty of a complex task, researchers try to define the terms “complexity” and “workload”. There is some difficulty in ascribing objective values to these terms when evaluating tasks such as air traffic control. Complexity is independent of the subject, and can be viewed as a function of the number of states or possible events attributable to the task. Workload is usually broken into the amount of physical work being accomplished (physical workload) and the amount of cognitive work being accomplished (mental workload).

Mental workload has been very difficult to measure. Objective measures are task dependent, and their accuracy relies on our ability to understand what makes a task difficult mentally. Subjective measures are somewhat more reliable, and different scales have been developed. One such scale commonly used in aviation is a Cooper-Harper scale. This was used to evaluate the controllability of new aircraft designs. The subject would follow a decision tree, answering questions about errors, ability to perform the task, and mental workload to attribute a number to the mental workload. This scale has been modified for use in numerous other experiments. Sheridan and Simpson (1979) developed a three-dimensional scale, which separated out “time load”, “mental effort load” and “stress load”. The subject chooses the description which best fits their experience to come up with a rating in each category. This scale was used in the experiment described in the following sections.

The experiments which follow were intended to test subjects’ ability to monitor conflicts and perform secondary tasks when confronted with high workload situations. Sheridan (1992) showed that as workload increases, performance does not degrade significantly, up to a point. Beyond that, very small increases in workload cause precipitous decreases in performance. Additionally, at very high workload situations, subjects begin to adopt an attention-switching strategy. In this situation they are no longer able to perform the task in an orderly fashion, but tend to “sample” the task, and performance
suffers as a result. Conversely, if workload is too low, vigilance suffers and performance also declines.

4.3 ROC analysis

Of particular interest in this paper is the application of the receiver operating characteristic (ROC) curve from signal-detection theory. This has been discussed and used in the context of human supervisory control for several years (see Sheridan, 1992; Kuchar, 1995).

In this analysis, one normally would crossplot the probability of having a false alarm (an alert issued when none is required), against the probability of having a missed detection (no alert issued when one is required). The result is a receiver operating characteristic (ROC) curve. The curve demonstrates a tradeoff between increasing correct detections and false alarms. For example, in a purely guessing strategy, the ROC would be a straight line connecting (0,0) with (1,1). As the system improves, the center of the curve distends towards the (0,1) point (upper left corner). This point is where we would have perfect correct detections and zero false alarms. The closer one gets to this point, the better the alerting system.

In automated alerting systems, only improving the sensors can modify the curve. In the human context, the curve can be improved through training, altering procedures, using different controllers, or other ways of modifying the conditions of the task or the human operator. The point on the curve at which one operates, however, can generally be changed easily. In an automated system, changing alert thresholds would accomplish this. For a human-based system, altering the rewards/penalties for alerting should cause the operating point to shift.
Chapter 5

TARGET TRACKING EXPERIMENTS

5.1 Introduction

Given the substantial evidence that top-down perceptual organization occurs in accomplishing complex tasks, and that eye movement is indicative of this organization, it was endeavored to examine whether this organization could be captured in an experimental setting. A first experiment manipulated factors which were expected to affect workload to determine if those factors might be a basis for clustering. A second experiment used an eye-tracking device to detect cognitive groupings. If air traffic controllers were perceptually organizing traffic, then the ease with which this organization occurs should impact task performance.

As noted earlier, previous attempts using recall techniques to capture perceptual grouping have been inconclusive. It was surmised that this is due to the generally poor level of recall of details often asked in recall questionnaires. If this is the case, recall may be poor, while task performance should remain good.

5.2 Experiment 1

The first experiment described below was an effort to examine recall and performance levels, and to establish elements of complexity to be examined as potential grouping criteria.

5.2.1 Setup and design

The experiments were run on a 486-based PC using a target-tracking program written by the author in C/C++. A typical screen is shown in figure 9. Each cycle the screen was updated and all targets moved. A cycle was a program timeout of 1000 msec, although the actual event time was slightly less (approximately 990 msec) due to the way Microsoft Windows processes these timeout events.
The subject’s task was to accept targets as they appeared on the screen (by using the mouse to select the target and pressing the “A” key), “handoff” the targets before they left the screen (by using the mouse to select the target and pressing the “H” key), and monitor for conflicts. A conflict was defined as any targets within 10 levels of each other that would touch as they traversed the screen. For instance, if one target was at level 200 and another was at level 190, and the subject believed they were going to collide on the screen eventually, he would select the two targets as a conflict (by using the mouse to select the two targets and pressing “C”). The subject was also asked to deselect pairs no longer in danger of conflicting, but this instruction was routinely ignored and was removed from the analysis.

This latter finding may be of interest if studied further. There may be a distinction between tracking an individual target and a combined pair of targets, and it may be difficult to focus on the two things simultaneously. In other words, if the subject was supposed to track only pairs of potentially conflicting aircraft, it may be easy to do. However here the subject was supposed to track individual aircraft (accepting, handing off, watching for conflicts), and it may be difficult to combine tracking individual targets with pairs of targets.
The experiment consisted of 3 subjects performing the target tracking and monitoring task. The subjects were male M.I.T. graduate students. Two of the subjects performed three sets of 12 trials, and the remaining subject performed two sets of 12 trials. Each trial was five minutes in duration, with the configuration of the trial varying over the set of 12 trials, as shown in table 1. "Speed differences" set to "no" means that all targets had a speed of 3 pixels/cycle (a cycle is approximately one second in this program). If set to "yes", it means that speed for each target was randomly chosen from 1 to 5 pixels/cycle. "Level changes" set to "no" means that all targets remain at the assigned level (randomly set between 100 and 200) throughout the trial. If set to "yes", targets were given a random starting level, a random finishing level and a random time (set for sometime during the traversal of the screen) to descend/climb to that finishing level.

<table>
<thead>
<tr>
<th>Trial</th>
<th>Targets</th>
<th>Speed differences</th>
<th>Level changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>18</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>18</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>11</td>
<td>18</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>12</td>
<td>18</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 1. Experiment matrix.

The subjects ran through at least one long trial to eliminate learning curve effects, and to ensure there were no questions about how to perform the different actions used in the experiment. The delays and errors in accepting and handing off targets, the conflict detection error rate (false alarms, missed detections, correct detections and correct rejections), and the lead times in determining conflicts were all measured.

If a user did not accept a target within the first 60 seconds of it being on screen, an acceptance error was recorded. If the user failed to handoff a target prior to its leaving the screen, a handoff error was recorded. False alarm and missed detection results were normalized to the number of crossing to obtain
percentages. This was done because some targets did not approach other targets, indeed some weren’t even on the screen at the same time. So I attempted to obtain an accurate picture of when a user might reasonably infer a collision potential, and capture the difficulty of the scenario. 150 pixels were used after analyzing the data and finding an outer limit of when a user definitely wouldn’t detect a conflict.

5.2.2 Results

The pertinent results of the experiments are included as figures 11 to 15. A definition page precedes this (figure 10).

<table>
<thead>
<tr>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Errors:</td>
</tr>
<tr>
<td>- Missed acceptances (&gt;60 sec)</td>
</tr>
<tr>
<td>- Missed handoffs</td>
</tr>
<tr>
<td>FA = false alarms</td>
</tr>
<tr>
<td>- not conflicting due to altitude</td>
</tr>
<tr>
<td>- not conflicting due to distance</td>
</tr>
<tr>
<td>MD = missed detection</td>
</tr>
<tr>
<td>CD = correct detection</td>
</tr>
<tr>
<td>Crossings: total number of pairs that pass within 150 pixels</td>
</tr>
</tbody>
</table>

Figures 11 and 12 show the correct detection and false alarm rates, respectively. The correct detection rate is perfect for 6 targets, and drops off to a low of about 38% for one of the 18 target trials. The false alarm rate, interestingly, is higher for the six target trials than for the more difficult trials. The initial suspicion was that this was due to inattention caused by low workload, or perhaps the subject was merely “creating work” for himself. Figure 13 suggests another reason, however.
Correct detection rate

![Correct detection rate graph]

Figure 11. Correct detection rate

False alarm rate

![False alarm rate graph]

Figure 12. False alarm rate
Figure 13 shows the detection lead times, which is the amount of time prior to the closest point of approach that the subject chose two targets as potentially conflicting. As the number of targets increases, this time gets shorter. This helps explain why the number of false alarms increased at lower numbers of targets. The subjects were selecting targets for conflicts earlier, and hence their accuracy was less.

It was then undertaken to show the correlation between false alarm rate and prediction time. Figure 14 shows this correlation. As the lead times increase, the false alarm rate also increases.
Figure 14. False alarm rate vs. detection lead time.

The ROC curve is shown as figure 15. The effect of increasing complexity is to increase the number of missed detections, while decreasing complexity increases false alarms. The detection performance is perfect for 6 targets, with missed detections increasing rapidly as the number of targets increases to 18. The number of false alarms averages about 4% for the combined subjects and 18 targets, and increases to about 10% for 6 targets.
5.2.3 Experiment 1 – discussion

The number of aircraft in the scenario was the largest contributor to complexity. At 18 targets, subjects indicated they began to adopt a switching strategy, where targets are randomly scanned. This is typical of subjects overloaded in a target-tracking task. The 12 target scenarios were more easily managed. The fact that a typical controller workload is also in this range could be a coincidence, but there also may be an underlying reason for this number.

The main result of this experiment was to show that as the subject had more time to perform the task (lower workload), the number of missed detections went down at the expense of higher false alarms. This was due to a longer detection lead time. As the lead time increases, the accuracy of the projection of trajectories decreases, resulting in more false alarms. Increasing workload, thereby reducing spare time, had the opposite effect of raising missed detections and lowering false alarms.
Discussions with the subjects also showed that recall for data tag information is very poor. In most cases the information was not recalled at all. Subjects reported that they had seen the information, but had no ability to recall it.

5.3 Experiment 2

5.3.1 Setup and design

Since the contributions of complexity were proving difficult to identify using recall and performance measures, an eye tracker device was purchased. This would provide input into software developed by Yufik (1996) that would return grouping information. The experimental software was also replaced. A Visual Basic program was developed which was very similar to the C program used in the first experiment. This allowed the use of database information for scenario generation and data collection, greatly simplifying both. The experimental set up is shown in Appendix 1.

The purpose of experiment 2 was to attempt to correlate grouping information with some objective criteria, such as recall, performance, or task objectives. Nine new scenarios were developed. The number of targets would remain constant (approximately 14) to remove this effect from performance. The interface was simplified to eliminate the need to use the keyboard, since this would be outside of the range of the eye tracker.

The variables in the nine scenarios were configuration and number of potential conflicts. The three configurations used were canonical, sequenced and random. The canonical configurations used are geometric, and have the “good form” described by Gestaltists. Three scenarios were developed which contained canonical configurations of traffic. In one case a diamond shape cluster of traffic moved from the upper right corner of the screen to the lower left. In a second scenario one diamond-shaped formation and one triangular formation were used. In a third a triangle formation was used. Sequenced configurations have “common fate”, appearing to move in the same direction. Three scenarios had sequenced configurations. In these scenarios there was a predominant flow of traffic along a predefined path, with about 50% of the total traffic volume moving along the sequenced route. This route begins (and ends) at a geographic point and continues in either direction. In one scenario two of these flows were apparent. In random scenarios care was taken to avoid any canonical or sequenced configurations while randomly dispersing targets across the screen,
with each moving in a semi-random direction. Directions were not completely random so as to avoid the configurations mentioned above, and to keep the targets on the screen for the desired amount of time.

The number of conflicts was varied between high (around 5), low (around 1) and none. Targets were randomly assigned an altitude from either flight level (FL) 310 (31,000 feet), FL330, FL350, FL370 or FL390. A wind speed and direction was randomly assigned to the scenario and groundspeeds were calculated (approximately) by adding wind velocity to a true airspeed (TAS) velocity of 400 knots. Groundspeeds ranged from 340 to 480 knots.

<table>
<thead>
<tr>
<th>Scenario number</th>
<th>Configuration</th>
<th>Coaltitude crossings</th>
<th>Actual conflicts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sequenced</td>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>Random</td>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>Sequenced</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>Canonical</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Random</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>Random</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>Canonical</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>Sequenced</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>Canonical</td>
<td>11</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2. Experiment matrix.

All acceptances, handoffs, and detections of conflicts were recorded, including the scenario time for each action. From this data were extracted lag times for acceptances, missed acceptances, missed handoffs, lead times for detections, missed detection counts and false alarm counts.

5.3.1.1 Eye movement monitoring system

An Iscan, Inc. Eye Movement Monitoring System was used to record eye movement data. This system consists of a standard video camera attached to a pan and tilt unit which images the subject's eye, an IL-400 Infrared Illuminator, an RK-464 Remote Eye Imaging System, an RK-520 Autocalibration System and an RK-426 Pupil/Corneal Reflection Tracking System. Two monitors provide feedback to the experimenter. The experimenter can view the output of the eye-imaging camera on one monitor (eye monitor). The other monitor redispays video output from the monitor that the subject is viewing (scene monitor). The scene monitor was output to a video recorder for playback and analysis.
The IL-400 Illuminator is a 20 watt incandescent light source mounted behind a glass infrared pass filter, housed in a cylindrical assembly. Both the intensity and direction of the light source can be adjusted. This unit provides an infrared (IR) light source for the RK-426, without being visible to the subject.

The RK-426 is a real-time digital image processor that tracks the center of a subject’s pupil and a reflection of IR illumination off of the cornea. Input is received from the video camera centered on the subject’s eye. Areas of the eye except the pupil reflect the IR, while the pupil acts as a sink, resulting in a dark pupil image. It is this image which is tracked by the unit. The cornea, the curved surface of the eye behind the pupil, then reflects the IR. The unit automatically tracks the positions of the pupil and corneal reflection (CR) over a two-dimensional matrix of the eye-imaging camera. Changes in the subject’s gaze angle are calculated from these positions.

The RK-426 divides the subject eye video into a 512 x 256 pixel field, and the coordinates of the center of the pupil and CR are updated every 16.7 msec and recorded. Accuracy of the system is dependent on the size of the eye image. Tighter zooming results in a larger eye image, more pixels per millimeter, and better resolution. This is done at a cost of diminished range of head movements. Typically, eye movements of ±15° can be accommodated. Subjects wearing glasses and contacts were successfully calibrated during the course of the trials.

The RK-520PC is a real-time computation and display unit that calculates a subject’s point-of-regard with respect to the viewed scene. A cursor is overlaid on the operator scene monitor, and the output can be saved and analyzed by software included with the system.

The system requires a calibration procedure to associate the relative position of the eye camera with the subject’s eye. The calibration requires the user to fixate on five known points on the screen, one at a time. The operator indicates when the user is fixating on the points, and correlating factors are calculated. These correlating factors are used to determine point-of-regard for subsequent eye movements. Calibrations were performed before each trial. Once calibration was complete, subjects were asked to fixate on the calibration points again to manually check the calibration.

The video scene being viewed by the subject is divided into 512 x 256 pixels. The raw pupil and CR positions are correlated to a pixel position on the viewed scene.
The RK-464 allows the operator to control the direction, focus, magnification and iris of the eye-imaging camera. Once the eye image is manually acquired, it is operated in an automatic mode where the camera automatically keeps the pupil centered in the field of view. The automatic mode is effective within ±12 inches horizontally by ±4 inches vertically of head movement.

5.3.1.2 Yufik software

An efficient method for calculating groupings was developed by Yufik (1996). This software initially was developed to speed computation of target-allocation problems, which have shown clustering characteristics. The software defines nodes (in our case targets), and forms links between these nodes based on some defined criteria. In this experiment the links were formed by incrementing links between two targets when transitions between those two targets took place, and weighting the link by the fixation durations.

A “virtual network” is thereby formed, with nodes linked to other nodes with different weights. Clusters are formed when the sum of the weights internal to a sequence of nodes exceeds the weights of all external links. Essentially the software returns clusters of targets among which repetitive patterns of fixations have been demonstrated.

5.3.1.3 Method

Eighteen subjects were recruited to participate in the experiment. Subjects were undergraduate or graduate students at M.I.T. Four subjects could not be calibrated and were dismissed from the experiment. The remaining fourteen subjects were generally untrained in air traffic control, although one subject had 140 flying hours, one subject had 70 flying hours, and those two subjects as well has three others reported some experience with air traffic control concepts. Subjects were paid $15.00 for their participation. No incentives for performance were used.

Screen shots are shown in Appendix 2. When the trial was started, a number of targets would appear on the screen. Any targets at the edge of the screen, or those coming onto the screen after the trial was begun, needed to be “accepted”. Targets that needed to be accepted were indicated with a circle as the target icon. Those already accepted were crosses. To accept a target the subject had to click on the target, at which time the icon would change to a cross. Once targets had traversed the screen and approached the edge, subjects were required to “hand off” the targets. Handing off would be done by clicking on the target, which would cause the target icon to be highlighted. The subject must then select a menu item, “Handoff”, which could be found under the “Controls” item. If the
target were close enough to the edge (approximately 5 NM) and were leaving the screen through that edge, the target would then be removed from the screen.

The primary task of the subjects was to monitor the screen for conflicts. They were instructed that all other tasks were secondary to detecting conflicting pairs of targets. A conflict was any time two targets at the same altitude passed within 5 NM of each other. Double clicking on any target would bring up a 5 NM range ring (called a “J-ball” by air traffic controllers). A conflict was identified by clicking on the two (or more) targets, which would highlight them. Selecting “Conflict” from the “Controls” menu would remove the highlight and add the target pair to the “Conflict Pairs” menu. If more than two targets were selected, the targets would be broken down into all possible pairs, with those pairs being added to the “Conflict Pairs” menu.

Subjects were briefed to select conflicts as soon as they thought there might be a chance of conflict, and to deselect them after the conflict potential (or actual conflict) had passed. This was accomplished by selecting the pair to be deselected from the “Conflict Pairs” menu. As in the first experiment, deselection was almost universally ignored and was removed from the analysis.

Once the trial time limit was reached, target details would be erased and two questionnaires appeared. The first questionnaire asked questions about three targets. The same questions were asked for each trial and each subject, and the same targets were asked for the same scenario for each subject. For each trial three different targets were selected. Generally, one target was selected from a conflict pair (except where there were no conflicts), one target was selected from the canonical or sequenced configuration (if such a configuration existed for that trial) and one was selected which was somewhat of an outlier. The questions were specific to the target, and asked about speed, altitude and with which other targets the subject target conflicted. The speed and altitude questions were multiple choice, selected by check box. The conflict question was “fill-in-the-blank”. Subjects could choose any of the five possible altitudes as an answer, and could choose either “below 400 knots”, “400 knots” or “above 400 knots” as an answer regarding speed.

The second questionnaire contained three questions, which were selected pseudo-randomly from a list of 15 questions. The list of questions is shown in table 3. The last three questions were the questions that appeared at the end of the sample trial. Each question was chosen to be somewhat relevant to the scenario, but no question was asked more than three times. The questions were designed to be general questions about the scenario, rather than specific to any one target.
<table>
<thead>
<tr>
<th>Number</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>How many conflict pairs were there?</td>
</tr>
<tr>
<td>2</td>
<td>How many targets did you accept?</td>
</tr>
<tr>
<td>3</td>
<td>How many targets did you hand off?</td>
</tr>
<tr>
<td>4</td>
<td>What was the predominant direction of flight?</td>
</tr>
<tr>
<td>5</td>
<td>What was the most common altitude?</td>
</tr>
<tr>
<td>6</td>
<td>In what direction were the targets moving the fastest?</td>
</tr>
<tr>
<td>7</td>
<td>In what portion of the screen were there conflicts?</td>
</tr>
<tr>
<td>8</td>
<td>How many targets were travelling south, including southeast and southwest?</td>
</tr>
<tr>
<td>9</td>
<td>How many targets were travelling north, including northeast and northwest?</td>
</tr>
<tr>
<td>10</td>
<td>How many targets were travelling east, including southeast and northeast?</td>
</tr>
<tr>
<td>11</td>
<td>How many targets were travelling west, including northwest and southwest?</td>
</tr>
<tr>
<td>12</td>
<td>How many targets were at FL390?</td>
</tr>
<tr>
<td>13</td>
<td>How many targets were at FL310?</td>
</tr>
<tr>
<td>14</td>
<td>How many targets were travelling above 440 knots?</td>
</tr>
<tr>
<td>15</td>
<td>How many targets were travelling below 360 knots?</td>
</tr>
<tr>
<td>16</td>
<td>Test Question 1?</td>
</tr>
<tr>
<td>17</td>
<td>Test Question 2?</td>
</tr>
<tr>
<td>18</td>
<td>Test Question 3?</td>
</tr>
</tbody>
</table>

Table 3. General questions.

Subjects were told to answer the target specific questions only if they were sure. However, they could choose more than one answer. For instance, if they knew that the target was at either FL310 or FL330, they could choose both answers. For the question that asked about the conflicting targets, if they were sure that no conflicts existed, they were told to answer "none".

Each subject signed a consent agreement and read instructions that gave a general description of the research, although mentioned nothing about perceptual grouping. The instructions indicated their tasks and how they would perform them. Once any questions about the task were answered, subjects began a five-minute sample trial to familiarize themselves with the task. During the trial they were shown how to perform all required tasks and were allowed to ask any questions they might have. If subjects had further questions, or demonstrated some uncertainty about the task or controls, a second sample trial was run. At the end of the sample trial the two questionnaire windows would open. The subject was briefed on what to expect from the questionnaires and how to answer them.
5.3.2 Results

The timing, recall and fixation data were collected, evaluated and correlated. Specific recall questions were graded by how accurate the answer was, while general recall questions were graded by how closely to the right answer the subject answered. Fixation data was compared with target trajectories to correlate a fixation with a target. Two methods were used to establish which, if any, target was being observed. One method was to select the nearest target to the fixation. A second method applied a weighting based on the fixation's proximity to the target. The former method was used for input into the Yufik software. The latter was used for all other analysis.

5.3.2.1 Raw data

As seen in the previous experiment, acceptance misses and lags are an indicator of workload. In this task there was little correlation to acceptance lags, although in general acceptance misses do seem to indicate workload.

![Missed Acceptance Counts](image)

Figure 16. Missed acceptance counts.

The results are somewhat consistent with expected workload. However, scenario 3 had no conflicts, was sequenced and had only 10 coaltitude crossings. These factors should have made it one of the easiest scenarios, and yet there were a large number of missed acceptances. Likewise scenario 1 should have been difficult (5 conflicts, 25 coaltitude crossings, sequenced), but showed a surprisingly low number of missed acceptances. Acceptance lags had large
standard deviations, making it difficult to draw conclusions from the data. There were only eight targets accepted in each scenario. This small number probably resulted in the times seen below and the missed acceptances seen above being due more to exactly where on the screen the targets appeared than on workload-related factors.

![Acceptance Lag Times](image)

Figure 17. Acceptance lag times.

Conflict detection performance was poor in the task. Experiment 1 demonstrated that 12 targets were difficult for untrained subjects to track, and 18 were nearly impossible. These subjects appeared to have difficulty at 14 targets per scenario. The number of missed detections was high, and the number of false alarms low.
Table 4. Correct detections and false alarms.

<table>
<thead>
<tr>
<th>Run ID</th>
<th>CD Rate</th>
<th>FA Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 5</td>
<td>NA</td>
<td>0.94%</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>NA</td>
<td>4.11%</td>
</tr>
<tr>
<td>Scenario 4</td>
<td>78.6%</td>
<td>1.32%</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>57.1%</td>
<td>1.87%</td>
</tr>
<tr>
<td>Scenario 8</td>
<td>50.0%</td>
<td>2.15%</td>
</tr>
<tr>
<td>Scenario 7</td>
<td>32.9%</td>
<td>1.91%</td>
</tr>
<tr>
<td>Scenario 6</td>
<td>26.5%</td>
<td>3.75%</td>
</tr>
<tr>
<td>Scenario 1</td>
<td>15.7%</td>
<td>3.45%</td>
</tr>
<tr>
<td>Scenario 9</td>
<td>14.3%</td>
<td>2.78%</td>
</tr>
</tbody>
</table>

Figure 18. ROC curve.
When crossplotted, the ROC curve seen above is generated. Although the curve indicates high performance in general, it is where on the curve the subject is operating that is of concern. Subjects are valuing correct rejections above correct detections. This causes there to be few false alarms at the cost of high missed detections. Most of the points should have a near perfect CD rate, at the expense of higher false alarms. In a real ATC task, false alarms are much less costly than a missed detection.

Recall was also very poor, in both the general and specific questions. In the specific questions, most were not answered. When they were answered, there were more right answers than wrong, most distinctly for the conflict question. Only 3% of the questions were answered wrong for the conflict question (compared to 15% for the other questions). So although conflicts are not recalled more often than other pieces of information, they are recalled more accurately.

![Specific recall questions](image)

Figure 19. Specific recall performance.

General recall was better for certain types of questions. It was still very poor for questions about direction of flight, speed and altitude. The only questions answered perfectly above 50% of the time asked about the number of conflict pairs, the number of targets accepted and in what direction the targets were moving the fastest.

Interestingly, subjects correctly answered the number of conflict pairs 68% of the time, but only answered where on the screen those conflicts were 33% of the time. Air traffic controllers have been shown to have excellent spatial memory.
for target position. It may be that controllers obtain this ability during training, or that this is a necessary talent to be a successful air traffic controller.

Another interesting thing to note is that subjects knew how many targets were handed off 64% of the time, but only recalled the number of targets accepted at 21%. This is probably because acceptances were a relatively simple task, requiring only that the subject see the unaccepted target icon and click on it. Handoffs required the subjects to monitor a target until they approached the edge of the screen, highlight the target, then select a menu item, making sure that the target then disappeared. The additional time spent dealing with the target, is what caused better recall.

Workload was considered somewhat low by the subjects, although their performance suggests otherwise. The order of the workload rating does seem to have good correlation to expected workload factors (number of conflicts, number of coaltitude crossings). Against number of conflicts $\rho=.72$ for time load, .75 for mental load and .60 for stress load. Against number of coaltitude conflicts, $\rho=.61$ for time load, .82 for mental load and .83 for stress load. There was very little correlation to performance ($\rho=.23/.22/.15$ for time load, mental load and stress load respectively).

Figure 20. Raw workload averages.
5.3.2.2 Correlations

An attempt was then made to correlate the various statistics. Primary task performance is indicated by CD/FA/MD/CR data. The subjective workload scale and secondary task performance indicate workload. The remaining items, recall and clustering (or transitions), have essentially an unknown effect on primary task performance and workload.

One might speculate that the reason for poor recall is that subjects do not fixate on targets sufficiently. Figure 21 indicates otherwise. There appears to be no correlation ($r = .17$) between the number of fixations on a target and the ability to recall. Subjects fixate as frequently on targets for which they had poor or no recall as those on which they had good recall.

Number of fixations does appear to have some correlation with conflicting targets and detections, however. In figure 22, it is seen that targets detected as conflicts have more fixations than those that are not detected, as expected. The third column is missed detections (MDs), which are targets that are actual conflicts but are not detected. These are actually included in the “not detected” column, which includes correct rejections as well as these MDs. While the difference is not statistically significant, it appears that MDs may have more fixations than targets not detected. This would mean that subjects actually fixated
on the conflicting targets they did not detect as conflicts to a greater extent than those targets that were not conflicting. This would reject the interpretation that subjects did not detect a conflict because they did not look at one target or another.

A more interesting statistic when examining conflict pairs is transitions, in particular how many transitions between certain pairs of targets. The average number of transitions between targets detected as conflicts and not detected as conflicts is seen in figure 23. Pairs detected as conflicts had nearly three times as many transitions between them on average compared to pairs not detected.

Figure 22. Fixation categories.

Figure 23. Transitions between pairs.
The Yufik software extends the transition analysis to clusters, and gives a more complete picture of the pattern of fixations. In attempting to account for the elements within the cluster, the clusters were examined for the presence of Pragnanz factors (targets belonging to a canonical or sequenced group), proximity of targets within the cluster, number of coaltitude pairs within the cluster, and the prevalence of detected conflict pairs within the cluster. The first two factors would indicate Gestalt-type groupings, the latter two would indicate goal-oriented clustering.

For each subject, scenario and cluster the percentage of targets within that cluster that belong to a sequence or canonical form was calculated, and the average of this percentage is the “Actual” column in figure 26. In several scenarios there was more than one sequence or more than one canonical form, and the targets had to be of the same sequence or form to be counted. The expected percentage was calculated by taking the percentage of targets for the scenario that belonged to the canonical or sequenced group, multiplying it by the size of the cluster. This provided the number of targets within each cluster one would expect to find based on random chance. The percentage was found by summing over all clusters and dividing by the total number of targets within the clusters.

As an example of the calculations, the actual canonical 1 percentage is 66.9%. Only scenarios 4, 7 and 9 had sequenced targets. The clusters were compared to
a list of the targets belonging to the first sequence, and a percentage of those targets within the clusters was calculated. For scenario 4, sequenced targets represented 60.3% of the clusters on average. For scenario 7 the percentage was 63.6%, and for scenario 9 the percentage was 76.8%. Averaging these percentages yields the 66.9% figure.

To obtain the expected percentage of targets, it was determined that 25% of the targets in scenarios 4 and 9 were in the sequenced group, and 18.8% of the targets in scenario 7 were sequenced. For each cluster this figure was multiplied by the number of targets within the cluster. The result is the number of sequenced targets within the cluster one would randomly expect to find. The results were summed and divided by the total number of targets within the clusters to obtain the percentage seen above (22.9%).

It is clear from figure 24 that Pragnanz factors are an extremely significant factor in the formation of clusters. On average, Pragnanz factors are found 2.4 times more often than expected. In the canonical forms this figure is much higher.

Next the role of proximity was examined. The average separation between all pairs of targets was calculated and compared with the average separation of pairs within the clusters. Pairs of targets were used since proximity is (at least) between pairs of targets. If proximity were a factor in grouping, then the average separation of pairs within clusters would be lower than the average separation of all pairs. As can be seen in figure 25, this is the case.

![Average Separations of Pairs](image)

Figure 25. Average separations of pairs.
A common technique discussed by controllers is to partition traffic into coaltitude groups of aircraft. This is an example of goal-oriented clustering. To determine if the subjects were grouping the targets in this manner, each cluster was examined for the presence of coaltitude targets. Each target could be at one of five altitudes, with some altitudes more prevalent in some scenarios than others. If this type of clustering were occurring, one would expect to find a higher percentage of coaltitude targets within each cluster than one would find based on random chance. Figure 26 shows that the actual percentage of coaltitude targets within clusters is nearly identical to the expected value of coaltitude targets. The calculations were performed in the same manner as for the Pragnanz factors.

![Bar chart showing percent of coaltitude targets in clusters](image)

Figure 26. Percent of coaltitude targets in clusters.

Another possible goal-oriented clustering technique would be grouping by conflict pairs. Subjects may be grouping together targets that are in conflict. Since conflicting pairs often belong to another group (proximity, sequenced, coaltitude), it is somewhat difficult to separate out this factor. In figure 27, the percentage of correctly-detected conflict pairs that are found within the same cluster is shown. Also shown are incorrectly-detected conflict pairs (false alarms), correctly-rejected pairs and missed detections. Figure 27 shows that there are a higher percentage of detected pairs within clusters than pairs not detected, particularly for correctly-detected pairs. The line in figure 27 is the expected
number of pairs from any particular category one would expect to randomly find. This is the percentage of all pairs found within clusters.

![Percent of Pairs in Clusters](image)

Figure 27. Percent of pairs in clusters.

5.3.3 Discussion

Subject performance indicates workload is high, as the subjects are not making large numbers of false alarms and are making significant numbers of missed detections. Since the subjects were not given incentives to detect all possible conflicts, it is possible that they were simply ignoring the instructions, which emphasized that not missing a detection was the primary goal. Fourteen targets is a common number of targets to use in an enroute-type ATC task, although this particular sector would be much smaller than those commonly used. The density of the aircraft was higher, and that undoubtedly added to the task difficulty. Also, subjects noted that they were continuously busy during the task.

The question then becomes why conflicts are missed. A common hypothesis is that it is somehow due to "situational awareness". The results of the recall portion of the experiment, which are in agreement with the recent results of Gronlund, Ohrt, et al. (1998), suggest that this may not be the main reason. Performance on the recall task was generally poor. Specific information was only
recalled about 35% of the time. Of that 35%, speed and altitude information was recalled accurately 21% of the time, while conflict information was better at 30%.

It may seem that the poor recall and poor performance are somehow linked, and the situational awareness hypothesis is correct. However, previous practice trials with better trained subjects, who were familiar with both the software and ATC techniques, showed better performance and equally poor recall. Also, controllers were asked what information they could remember should the screen go blank, and they indicated that they could recall very little information. What information they could remember was approximate locations of targets, particularly conflict pairs, approximate locations of conflict points and some general information. This is also supported by the studies mentioned earlier.

Actual controllers would probably do better than untrained subjects at the recall task in the experiments. To what extent better or worse recall results in better or worse performance is the important question to answer. One of the more revealing results from above is that subjects recalled information about things that required more handling and more physical action than those that only required scanning. In recalling number of conflict pairs and handoffs, but not recalling where the conflict pairs were or the number of acceptances, subjects seem to be reflecting the fact that conflicts and acceptances were “handled” by selecting targets and using the menu to manipulate them. This result is supported by supervisory control concepts that have shown a relationship between physical action and better recall. It also supports controllers’ concerns about replacing physical paper flight strips with electronic ones.

The lack of correlation between number of fixations and recall seems to indicate that fixation time is not used to remember information. The average duration of fixation was around 200 msec. In combination with the very short durations between fixations, this pattern is too short to transfer information into long-term memory. The iconic memory is being constantly erased by subsequent fixations on essentially the same data. This would be like trying to read a list of phone numbers quickly with recall. The pattern of saccades and fixations is not conducive to recall. This suggests that the scan pattern is doing something else.

The question as to whether this “something else” is perceptual grouping, both formation and maintenance, is a difficult hypothesis to test directly. Yet perceptual grouping is difficult to avoid, and Yantis’ research shows that grouping occurs in target-tracking tasks. If the form of the grouping that we can see in this task agrees with expected groupings, then that would be strong evidence that such groupings occur.
The data clearly indicates that while conflict detection was somewhat dependent on number of fixations, there was a stronger relationship between number of transitions between a pair of targets and detection of a conflict between those two targets. The eye must move between the targets in the conflict pair in order to detect the conflict.

This may not seem like a startling conclusion. It does, however, grant more importance to any factors that influence the pattern of fixations. Moreover, targets are grouped perceptually, and this grouping affects the sequence of fixations. The factors that influence the formation of clusters, then, affect the ability to see conflicting target pairs.

It is then left to identify factors that influence grouping. Yantis found that in addition to Gestalt factors, "goal-oriented" factors played a role in perceptual grouping. "Goal-oriented" refers to any feature for which knowledge of the task objectives is required to connote importance. For this task, conflict detection is the primary objective. Only target pairs (or groups) that are coaltitude can conflict, and controllers often use this fact when scanning the scope. So coaltitude pairs or groups, and pairs or groups detected as conflicts should appear frequently if goal-oriented clustering is occurring.

This was (mostly) not the case. Gestalt factors, such as common fate (sequenced targets), Prägnanz (canonical targets) and proximity were found in great abundance within clusters compared to expected values. Other overlapping Gestalt factors which could not be measured with this data, such as crossing paths (regardless of altitude), were subjectively found to be present frequently as well. Coaltitude pairs were found as often than expected.

Detected conflicts were found more often, but the extent to which this is due to the proximity of targets and other factors is not entirely clear. The finding that 50% of the correctly-detected pairs were found within the same cluster, while only 30% of the missed detection pairs were found is important, however. This indicates that the ability to group conflict pairs within a cluster is important to the ability to detect that conflict. Gestalt factors, having an apparently more important role in determining clusters than conflicts, can interfere with the ability of the controller to form "valuable" clusters which contain conflicting pairs.
CONCLUSION

The first experiment failed to find evidence for contributions to complexity by task features such as target speed variation or altitude variation. The number of targets overwhelmed any impact these elements may have. Performance followed classic human supervisory control expectations in two ways. First, as the number of targets increased, subjects' ability to organize their search efforts deteriorated, eventually resulting in an inefficient random search. Secondly, at lower numbers of targets subjects had higher numbers of false alarms, due to the longer conflict detection lead times allowed by the lower workload, or perhaps also due to the subjects “creating work” for themselves.

The second experiment attempted to find evidence of clustering in eye movements, which have been shown to represent internal models of viewed scenes. The clustering was dominated by Gestalt factors. Some evidence of the involvement of clustering based on conflict pairs was seen, although no evidence of any other goal-oriented clustering was found. Furthermore, the presence of both targets in a conflict pair was an important factor in whether that conflict pair would be detected. If conflict pair members are also members of a Gestalt cluster, it is much more likely that the conflict will be detected. Alternately, Gestalt factors can interfere in detecting a conflict if the targets are not members of the same cluster. The extent to which this is also the case for trained controllers needs to be examined. It is expected that to some extent training can overcome this tendency. Gestalt factors, however, are an innate clustering strategy and would be difficult to eliminate altogether.

This result has important implications for future air traffic control improvements. Air traffic is currently highly structured, and undoubtedly this structure improves the ability of the controller to detect potential conflicts. An accepted proposal for future capacity improvements is the adoption of “free flight”, which will eliminate much of this structure. The extent to which controllers rely on this structure to “see” conflicts is not understood, and the lack of this structure could compromise safety.

Another conclusion drawn from the second experiment calls into question the use of recall tests to make complexity or performance conclusions. Recall
performance is generally poor and there does not seem to be a correlation between fixations and recall. The correlation with recall appears to be the method of interaction with the target. The method of accomplishing the primary task of detecting conflicts interferes with the ability of the controller to recall details of individual targets. If the controller interacts with the target in a more physical way, by accepting the target, or handing off the target, or in some other way more significant than providing separation assurance, the target’s details are more susceptible to recall.
Chapter 7

FURTHER WORK

The predominance of Gestalt factors over goal-oriented factors can be overcome through training, as was the case in Yantis' experiments. For this reason it is expected that actual controllers would show more goal-oriented clustering than these untrained subjects did. The extent to which Gestalt factors continue to interfere with detection for trained controllers would be a logical next step in this research.

The structure imposed on the task by the regulated nature of air traffic is jealously guarded by controllers, and can certainly be an aid in detecting conflicts. To some extent Gestalt factors can be imposed on the task, even in a random (free-flight) traffic scenario. Structured traffic can yield more consistency between potentially conflicting pairs and Gestalt groupings, reducing the likelihood of a missed detection due to failure to transition between conflicting targets. An experiment that probes the magnitude of Gestalt interference or assistance would clarify this contradictory effect of natural grouping criteria.

This could be done without significant modification of the software. New scenarios that specifically target certain grouping criteria would have to be developed. Most of the tools to import and analyze fixations already exist in the database, but new queries would have to be written to specifically pinpoint the factors being examined. Additionally, real controllers would have to be recruited. Attempts were made to recruit controllers for this set of experiments, but no responses have been obtained as of this writing.

Should the primacy of Gestalt groupings be confirmed even in the case of controllers, there is a number of uses to which this information could be put. If training allows the controller to overcome the Gestalt tendency and see conflicts not normally clustered, recording eye movements could be a valuable training tool. It could indicate techniques to effectively produce goal-oriented clustering, and could indicate when potential controllers require additional training.

The absence of conflict pairs in clusters is a warning sign of a potential missed detection. Eye tracking systems are probably still too intrusive to be used regularly by controllers, but the technology is improving. Should the technology
become sufficiently unintrusive, passive alerting systems could be developed to point out overlooked conflict pairs. The accuracy of prediction would need to be excellent for this to be useful.

Another use of this information is to simply display in some manner appropriate groupings of aircraft to help overcome the tendency to group aircraft in unproductive clusters. This could be done with colors or actual links incorporated into the display. What to do about climbing and descending aircraft would have to be addressed. Another difficulty is that if groupings are too convincing, it will be difficult to detect conflicts outside those clusters.

The techniques used in these experiments may also have applications to expert systems. If controller decisions can be related to how targets are clustered, and these decisions can be shown to have some predictability to them, then expert rules can be elicited without extensive subjective questionnaires.
EXPERIMENTAL SET UP – EXPERIMENT 2
Scenario 1
Scenario 2
Scenario 3.
Scenario 4.
Scenario 5.
Scenario 6.
Scenario 7.
Scenario 8.
Scenario 9.
REFERENCES


