Improving Human-Classifier Interaction through Enhanced Highlighting Techniques

by

Ronald Thomas Kneusel

B.S., Valparaiso University, 1988

M.S. Physics, Michigan State University, 1993

M.S. Computer Science, University of Colorado, Boulder, 2012

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

2016

This thesis entitled:
Improving Human-Classifier Interaction through Enhanced Highlighting Techniques
written by Ronald Thomas Kneusel
has been approved for the Department of Computer Science

_____

Michael Mozer

_____

Clayton Lewis

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

IRB protocol #13-0405

Kneusel, Ronald Thomas (Ph.D., Computer Science)

Improving Human-Classifier Interaction through Enhanced Highlighting Techniques

Thesis directed by Prof. Michael Mozer

This dissertation is concerned with developing techniques to improve detection of target objects in digital imagery, e.g., satellite image analysis, airport baggage screening, and medical image diagnosis. Human experts are fairly good at these tasks, but expertise takes years to acquire and human performance is fallible. Computer systems trained through machine learning methods are promising, but in many difficult tasks computer systems have not yet reached the level of performance of human experts. This dissertation proposes an approach to human-computer cooperative analysis to obtain results that are better than either human or computer alone could achieve. The traditional route to improving human performance with results from automatic classifiers is to *highlight* images by drawing boxes around regions of an image that the computer system believes likely to contain a target object. Human experts typically do not like this form of assistance: it's often obvious to the expert that the highlighted region is relevant or irrelevant, and highlighting some regions often causes other regions to be overlooked. This dissertation proposes an alternative to the *hard highlighting* technique of drawing boxes around candidate targets. The alternative, *soft highlighting*, provides graded saliency cues based on the confidence level of a classifier. For example, with grey scale satellite imagery, soft highlighting might take the form of varying the saturation level of a particular hue. The dissertation describes a series of 8 experiments to evaluate the costs and benefits of soft highlighting versus hard highlighting versus a control condition of no highlighting.

In Experiments 1-5, subjects search an array of handprinted digits for a given target digit identity. The elements of the array are highlighted according to the output of a classifier. The quality of the classifier was manipulated using a *stochastic, oracle-based* classifier that simulates classification to achieve a specified degree of discriminability between targets and nontargets. The

experiments measured the time to locate targets in the array. Soft highlighting allows subjects to find targets faster than hard highlighting or the no-highlight control, even for *weak* classifiers, i.e., classifiers which had little disrcriminative power. The experiments found that highlighting affects search slopes (the time to process each element in the display), meaning that search becomes more efficient with highlighting. Not only was search more efficient, but fewer targets are missed with soft highlighting versus hard highlighting.

Experiments 6-8 used actual satellite imagery. Subjects searched images for a particular target, a McDonald's restaurant. Highlights were obtained from a state-of-the-art convolutional neural net classifier which output a continuous confidence level. Experiment 6-8 also found that subjects could locate a target more quickly and with fewer misses with soft highlighting than with either hard highlighting or the no-highlight control. However, highlighting—both soft and hard—yielded more false alarms (nontarget locations identified as potential targets). We argue that while false alarms are a problem for novices who do not yet have the skill to verify the presence of a target, experts should not suffer from this same problem.

Dedication

To my family, in gratitude for their patience with this endeavor.

Acknowledgements

I would like to thank Robert Lindsey and Brett Roads for helpful discussions over the course of this work. I would also like to thank my family for their sacrifice of my time with them during the past six years. Maria, David, Peter, Paul, Monica, Joseph, and Francis, I love you all and again, thank you for your support. Lastly, I would like to thank my advisor, Michael Mozer, for his willingness to take on the risk of a student who was already far too busy for his own good.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

This dissertation examines the effect of various image highlighting techniques on viewer's perception of targets (objects) in digital images. In particular, it focuses on ways in which classification output, ie, the output from a machine learning classifier, can be presented to a viewer along with the original imagery so that the viewer can more quickly assess the presence of a target without becoming distracted or misled by errors made by the classifier.

We begin in Chapter 1 with a definition of what it means to highlight an image. We then develop a taxonomy of image highlighting and describe existing techniques in each area. From this taxonomy we will discover that highlighting techniques preserving image information while still guiding the viewer's attention to recognized targets in unexplored.

In Chapter 2 we describe how hard agent-based highlighting techniques (see Section 1.2) are currently used and analyze in detail research that demonstrates the limitation of these hard techniques to serve as motivation for experiments involving soft agent-based highlighting.

Chapter 3 presents the results of experiments involving highlighting of synthetic imagery. Chapter 4 details the construction of the machine learning classifier used in the experiments of Chapter 5 while Chapter 5 itself presents the results of experiments investigating the effectiveness of soft agent-based highlighting when applied to satellite imagery. Lastly, Chapter 6 discusses the results of the experiments and offers directions for future research in this area.

1.1 What is Highlighting and What Value Does It Offer?

As primates, humans rely heavily on visual input to understand the world. When looking at images our attention is drawn to regions that are in some way more salient than other regions. For example, an image of a calm lake is relatively uniform and uninteresting but if there is a white swan in the lake it is highly salient and our attention is immediately drawn to it. Why? In part it is because the swan is highlighted by being different in color and shape than its surroundings and our well-tuned visual system perceives this and pays attention to it.

Highlighting is the process of manipulating an image to alter the saliency of objects in the image. This is done in order to draw the viewer's attention to specific objects of interest. To be specific, in this article "highlighting" means any manipulation of an image which increases the viewer's ability to detect objects of interest. A range of highlighting techniques are found in the literature, and we describe a natural taxonomy of the techniques to help organize the otherwise scattered and diverse array of highlighting research.

Many review surveys have been written about specific aspects of image highlighting. These surveys cover everything from basic techniques [103] to more specific techniques such as segmentation [156] [125], edge detection [189], classification [57], and domain-specific processing [19] [220].

This article proceeds as follows. In Section 1.2 we describe the taxonomy which forms the framework for our survey. Sections 1.4 through 1.5 review each of the static image categories defined in Section 1.2. Dynamic highlights are reviewed in Section 1.6. We end with a dicussion and look forward to areas for possible future research in Section 1.7.

1.2 A Taxonomy of Image Highlighting Techniques

We divide the space of image highlighting techniques along the following dimensions: *soft* versus *hard*, *image-based* versus *agent-based*, and *static* versus *dynamic*. These dimensions form a taxonomy by which we can organize the space of highlighting techniques. This taxonomy can be visualized most easily as in Figure 1.1. For most of this article we will, because of the relative lack

of examples, ignore the *static, dynamic* axis in Figure 1.1 and focus instead on the combinations leading to *hard-image-based*, *soft-image-based*, *hard-agent-based*, and *soft-agent-based* highlighting techniques with the understanding that any of these approaches may be *static* or *dynamic*. Dynamic highlighting will be discussed separately.

Why these dimensions? The *soft* versus *hard* dimension is largely intuitive though it does form a continuum and a judgement call was sometimes necessary to assign a technique to one or the other. It is also recognized that many highlighting techniques are the result of outside information which in some way attempts to explain the content of the image. This leads to the the introduction of the *image-based* versus *agent-based* dimension to separate these techniques from those that are not attempting to understand the content of the image. The *static* versus *dynamic* dimension expresses the fact that highlighting can incorporate time-varying changes to the image itself in order to change the saliency of objects. Each of these dimensions will be described in turn.

1.2.1 Soft Highlighting Versus Hard

A *soft highlighting technique* will modulate the image in some manner, continuously or nearly so, in order to draw attention to the region of interest. The hallmark of this technique is that the modulation is not one that actively masks or removes other image features but instead causes those features to still be visible. A classic example of a soft highlighting technique is contrast adjustment.

A *hard highlighting technique*, on the other hand, will either add new features to the image, features that were not originally present, or remove image information in order to emphasize other parts of the image. For our purposes both the adding to and removing from are considered a hard highlight because they alter the information in the image. Adding a box around a portion of the image is a hard highlight, for example. We will present many examples of each of these techniques in the analysis below.

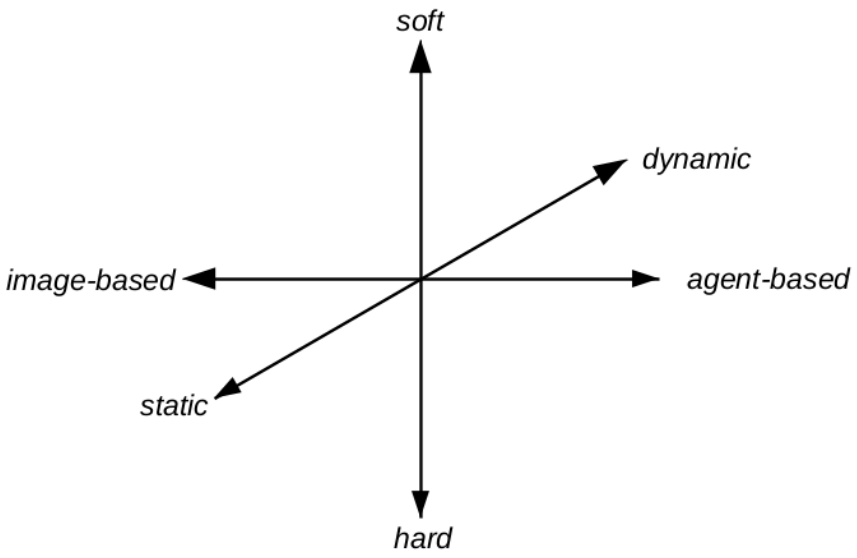


Figure 1.1: Taxonomy of image highlighting techniques. We categorize image highlighting techniques along three dimensions: *soft* or *hard*, *image-based* or *agent-based*, and *static* or *dynamic*. While these dimensions separate the space of highlighting techniques into eight octants in practice we often ignore the *static, dynamic* dimension because of the paucity of dynamic highlighting techniques. This reduces the space from 3D to 2D and involves only the quadrants spanned by *soft, hard* and *image-based, agent-based*.

1.2.2 Image-based Highlighting Versus Agent-based

An *image-based* highlighting technique is defined to be one in which the highlighting technique, while definitely a function of the pixels in an image, is not concerned with the *content* of the image, *per se*. That one image contains a scene with cattle in a field and another contains a star field is not taken into account by the highlighting technique. In this sense, an *image-based* highlight can be though of functionally as,

$$I' = f(I)$$

where I is the input image, I' is the output highlighted image, and $f()$ is a mapping function that depends only upon the pixels of the input image.

An *image-based* highlight does not rely upon any outside oracle to inform it of the content of the image. Just as contrast adjustment, for example, is a *soft* highlight technique, it is also an *image-based* highlighting technique as well since it does not depend upon the content (objects) in the image.

An *agent-based* highlighting technique, however, does depend upon the content of the image. How this understanding of the content of the image comes about is not specified. It is simply treated as an oracle which knows without specifying how it knows. In this case, we can express an *agent-based* highlighting technique functionally as,

$$I' = f(I, \theta_I)$$

where I is the input image (pixels), I' is the output highlighted image, and θ_I is the input from the oracle informing the function, $f()$, of the content of the image I.

1.2.3 Static Highlighting Versus Dynamic

The highlighting technqiues categorized in Figure 1.1 can be further categorized along another axis as either *static* or *dynamic*. The operating assumption in this article is that the input image

itself is static, the pixel content does not change with time. With this assumption, then, it is easy to see that a *static* highlight is one that generates another image which is also constant in time. The vast majority of image highlighting techniques fall into this category.

A *dynamic* highlighting technique produces a time-varying visual pattern. This pattern has the goal of changing the saliency of relevant objects in the image in order to direct the viewer's gaze to the highlighted region. This is accomplished through motion of some kind, either in space (jittering) or in some other image characteristic such as contrast, sharpness or color.

We expand upon this natural definition here to also include any highlighting technique that produces any sort of output that varies in time. For example, an image which contains regions that change appearance over time is considered to be *dynamic*. Additionally, an image where a varying sound is produced depending upon where the viewer is looking or moving the mouse pointer is also categorized as *dynamic*.

With our taxonomy in place we now proceed to describe each combined category and to offer examples spanning the range of image highlighting techniques found in the literature. We freely admit that our operational definition of "highlight" admits to a legion of possible techniques and we accept that some important techniques may be inadvertently omitted from our discussion. Nevertheless, we hope to have been thorough and to include in each category below the prime examples of these techniques. Where possible, we refer to foundational papers and then give other examples in the same genre to illustrate ongoing research in that area.

1.3 Hard Image-Based Highlighting

The first category we consider is that of *hard image-based* highlighting. The *hard image-based* techniques of this section produce output images which are often very different than the input images on which they are based. For example, they often discard information in order to draw the viewer towards the image portions to be highlighted. However, these techniques are blind in that they operate on the pixels of an image without attempting to "understand" the content. There are a myriad of hard techniques to choose from. Here we attempt to cover the most widely

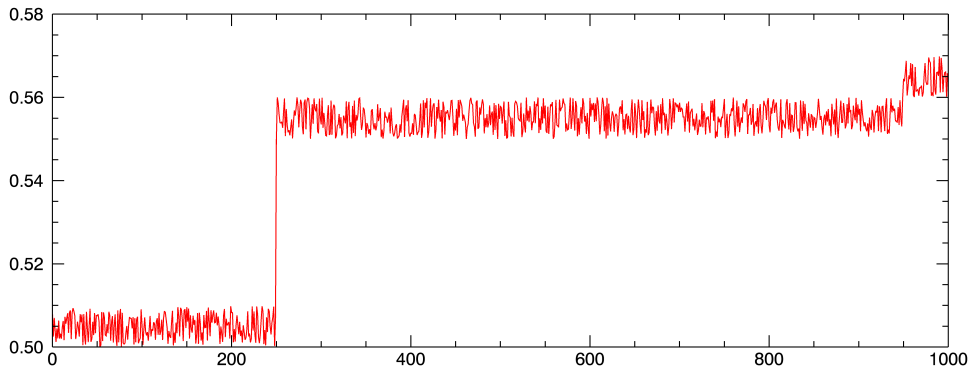


Figure 1.2: A one-dimensional edge. This plot, showing pixel intensity along a single row of a hypothetical image has an obvious edge on the left side and another, smaller, edge on the right.

used. We describe in turn: edge detection, thresholding, morphological filtering, color quantization and color tables.

1.3.1 Edge Detection Techniques

The human eye has evolved to be very sensitive to the presence of edges. Edge detection is important because it detects these edges in images and highlights them at the expense of other image content. In a one-dimensional case an edge is a transition from one relatively constant region to another region. For example, in Figure 1.2 the plot has an obvious edge on the left and, depending upon how sensitive one wishes to be, may have what could be called a second edge towards the right side. Viewed as an image where intensity represents the level of the pixel it becomes clear that transitions in intensity correspond to edges in an image.

Edge detection is a very old image processing technique [135] [47] [178] [12] which is still an active area of research [54] [139] . In this section we will examine three fundamental edge detection techniques: Roberts ([176]), Sobel ([201]) and Prewitt ([167]). All of these techniques are implemented as convolutions over an input image.

The Roberts detector combines the output of two 2x2 convolution operators,

$$E_{\text{Roberts}} = \sqrt{\left(\begin{pmatrix} 0 & -1 \\ \mathbf{1} & 0 \end{pmatrix} * I\right)^2 + \left(\begin{pmatrix} 1 & 0 \\ \mathbf{0} & -1 \end{pmatrix} * I\right)^2}$$

where $*$ represents convolution. As the kernel is 2x2 the output pixel is not clearly the center pixel as would be the case with a kernel of odd dimensions. In this case, the pixel location updated in the output image is marked in bold in the equation above. This creates an edge image (E) from the input image (I) by estimating gradient magnitudes formed by transitions between object regions. For the Roberts detector the gradients are diagonal as can be seen from the $+1$ to -1 transition in the kernels above.

The Sobel operator is similar but uses a different definition of gradients with a 3x3 convolution kernel. In this case each kernel computes the $\hat{x}$ or $\hat{y}$ gradient image which is combined to generate the final output magnitude image. Mathematically,

$$G_X = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} * I, \quad G_Y = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix} * I$$

with the magnitude edge image as,

$$E_{\text{Sobel}} = \sqrt{G_X^2 + G_Y^2}$$

The convolution kernels for the Sobel detector estimate the gradient along the $\hat{x}$ and $\hat{y}$ directions respectively as can be seen from the kernel values themselves. The Prewitt edge detector is very similar to the Sobel but does not place emphasis on the center line of the kernel. For the Prewitt detector the $\hat{x}$ and $\hat{y}$ kernels are,

$$P_X = \begin{pmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix} * I, \quad P_Y = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{pmatrix} * I$$

so that the final edge image is,

$$E_{\text{Prewitt}} = \sqrt{P_X^2 + P_Y^2}$$

We illustrate each of these detectors in Figure 1.3. The original image is in the upper left then going clockwise we have the Roberts image, Sobel image, and Prewitt image. Clearly each technique highlights edges at the expense of the majority of the image data and each does so in a subtly different way. Edge detection can be used as a first step in segmentation and draws the viewer's gaze to the outlines of image objects but still without any understanding of the image contents.

1.3.2 Thresholding Techniques

In its simplest form, thresholding involves making a binary image from an input image by picking a particular intensity value and setting all pixels with a lower intensity to zero and all pixels with a greater intensity to 255 (for a byte image). Naturally, it will be immediately suggested that ranges of thresholds can be used, etc.

As a technique for highlighting features in images thresholding is as aggressive a technique as possible. The resulting output image will be strictly binary. As a quick example, consider Figure 1.4 which shows the impact thresholding has on an input image. In this case a byte valued image of the Pentagon (left) has been thresholded at gray value 212. This means that pixels with gray values less than 212 have been set to zero and pixels with gray values at 212 or above have been set to 255. The thresholded image (right) has highlighted the landing pad on the left side so that it stands out clearly in the thresholded image.

Most automatic thresholding algorithms base their decisions on the image histogram. The most classic algorithm is the Otsu method [155] which has been cited over 17,000 times. This algorithm will be detailed below as representative of the entire volume of literature in this area. Some others include [105], [182], [2], and [156].

The Otsu method seeks the threshold (grayscale) value that maximizes the variance between

Figure 1.3: Convolutional edge detectors. Original image (upper left), Roberts edges (upper right), Sobel edges (lower right), and Prewitt edges (lower left). The various edge detectors highlight the image by preserving the outlines of the main objects in the image at the expense of lower frequency (smoother) image features.

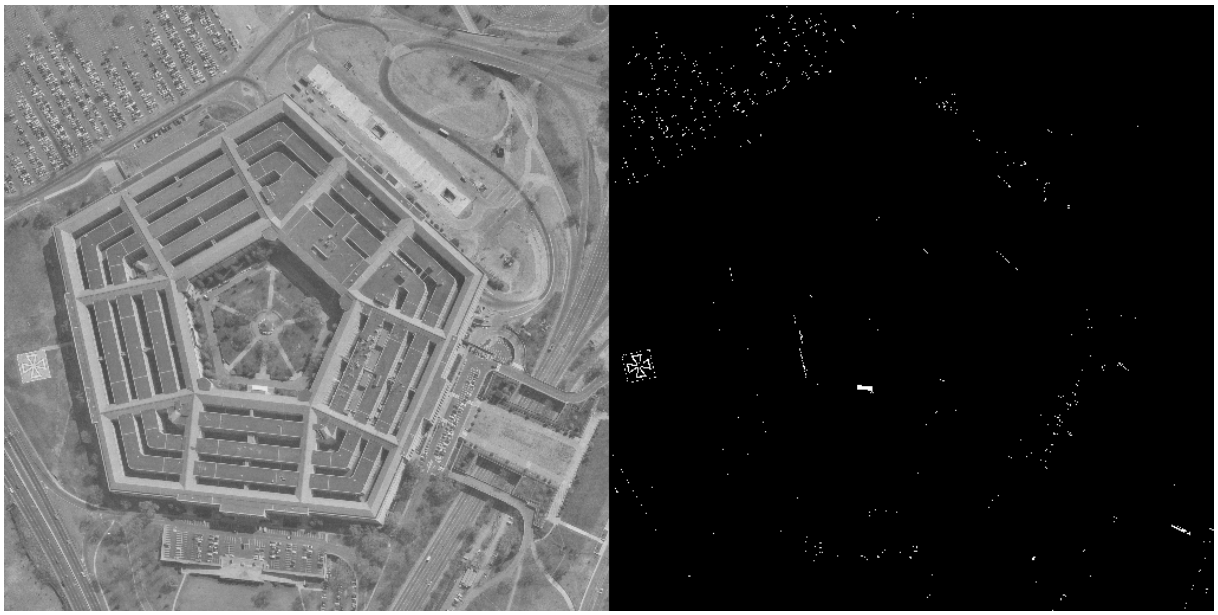


Figure 1.4: Simple thresholding. The Pentagon image (left, byte valued) has been thresholded (right) at gray value 212 so that pixels with gray values less than 212 are set to 0 and those at 212 or above are set to 255. Notice how thresholding has immediately highlighted the landing pad on the left side so that it stands out clearly in the thresholded image.

the histogram values below the threshold and those above. This variance, for a given threshold value t, can be written as,

$$\sigma^2(t) = \omega_L(t)\omega_H(t)(\mu_L(t) - \mu_H(t))^2$$

where,

$$\begin{aligned}
\omega_L(t) &= \sum_{i=0}^{t} p_i \\
\omega_H(t) &= \sum_{i=t+1}^{N} p_i \\
\mu_L(t) &= \frac{1}{\omega_L(t)} \sum_{i=0}^{t} p_i g_i \\
\mu_H(t) &= \frac{1}{\omega_H(t)} \sum_{i=t+1}^{N} p_i g_i
\end{aligned}$$

for gray level g_i which appears in the image with probability p_i as computed from the histogram. The maximum gray level is N. The threshold value $t_{\max}$ which maximizes $\sigma(t)$ is the value at which the image will be thresholded.

Figure 1.5 is an example of the Otsu algorithm in action. On the left is an image of a copy of the Magna Carta where the left side is the input image and the right side is the output after applying Otsu thresholding. The threshold removes background noise while still preserving, in most places, the text itself. On the right is a histogram of the Magna Carta image with the threshold selected by the Otsu algorithm indicated by a vertical line.

Thresholding as a highlighting technique is generally most effective in highlighting image objects that are either brighter or dimmer than the selected threshold. This is true for simple single-value thresholding as was shown in the examples above where the Pentagon landing pad and text of the Magna Carta were highlighted.

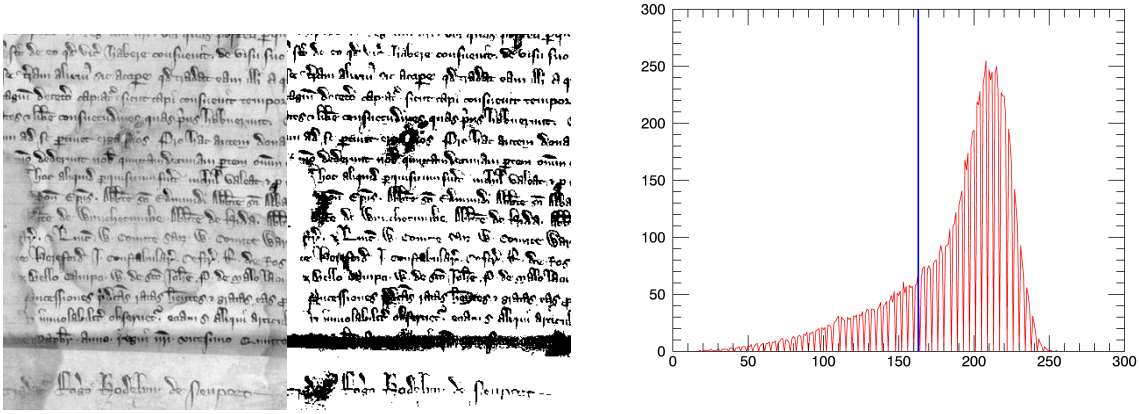


Figure 1.5: Otsu thresholding. The original Magna Carta image (left) is thresholded according to the Otsu algorithm to produce a binary image (center) which has removed much of the extraneous image information and highlighted the text in most areas. A histogram of the original image (right) along with a vertical line marking the threshold gray value determined by the Otsu algorithm. This is a byte image so gray levels range from [0,255].

Figure 1.6: Dilation. The small gaps in the top image are filled in by dilation to produce a more connected image on the bottom.

1.3.3 Highlighting with Morphological Filtering

Morphological filtering convolves a structuring element over an input binary image. The core operations in morphological filtering are *dilation*, *erosion*, *opening* and *closing* [191]. Each of these will be illustrated below. A typical use case is to take a thresholded image, operate on it with one or more of these filters, and then use the output of that operation to update the original image.

The effect of dilation is to fill in gaps less than the dimension of the structuring element. For example, consider Figure 1.6 where dilation with a 3x3 structuring element,

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

fills in the small gaps between the dots in the characters but not the larger gaps between the characters where the original image is on the top and the dilated image is on the bottom.

Erosion has the effect of removing objects that are smaller than the stucturing element in addition to thinning larger objects. Here objects refers to parts of the image that are fully connected. For example, Figure 1.7 shows erosion of the thick letters present in the top image to produce the thinned out bottom image. This is accomplished by eroding with a 5x5 structuring element of all ones.

Dilation and erosion are primitive morphological operations. From them are built the more

Figure 1.7: Erosion. The thick letters in the top image are eroded to produce the thinned letters of the image on the bottom.

standard operations of opening and closing. It is these operations that fall into our category of *hard image-based* techniques. Opening consists of erosion followed by dilation and closing consists of dilation followed by erosion. The effect of these operations is illustrated in Figure 1.8 which shows an original thresholded image of cells on the left which has been opened by a 3x3 structuring element of ones (center) and then closed with the same structuring element (right).

The net effect of opening followed by closing is to remove objects in the image that are not of the same size or larger than the structuring element and to preserve connections that are of the order of the structuring element in size. So, in Figure 1.8 we see that the right image, the final image, has kept large objects and preserved strong connections (thicker than the structuring element) but has filled in small gaps and removed isolated small objects. Morphological techniques highlight objects according to their scale by removing smaller objects.

Research in morphological image highlighting was more active in past decades (see [215] [211] [132] [45]) but is still being pursued today as well, for example, see [209] [160] [141] and [40].

1.3.4 Color Quantization Techniques

Color quantization is the process of reducing the color space of an image. It falls under the *hard image-based* highlighting category because the quantization involves the removal of image information (color depth) in exchange for possibly enhanced object disernability by which we mean the grouping of similar objects by moving their colors to a similar, common value. Color quan-

Figure 1.8: Opening followed by closing. The original image (left) is opened (center) and then closed (right). The structuring element was a 3x3 matrix of ones. Notice how the image on the right has rejoined sections in the upper middle that were connected originally (left) and then separated when the open operation was applied. Also notice how small objects in the original image have now been removed.

tization is used in this way to group similar colors together in [39] as a prelude to an automatic saliency map algorithm.

The most classic algorithm for color quantization is the median-cut algorithm [82]. This algorithm is recursive and involves setting a bounding box over the colors of the image as represented in RGB space (ie, a 3D space where each axis is one of the color components red, green or blue). This box is then split into two boxes by selecting a plane aligned with one of the color axes such that the number of pixels in the two new boxes formed is roughly equal and the plane is aligned parallel to the axis with the greatest variation. These new boxes are then subdivided recursively until the desired number of boxes matching the desired number of final colors is found. The last step averages the pixels in each box to determine the mean color of the box. These are the output colors.

For example, the median-cut algorithm applied to a 24-bit color image to reduce it to four colors results in Figure 1.9 where the original image is on the left and the four color quantized image is on the right. Naturally best viewed in color. In this example the original true color image contained over 230,000 unique RGB color values. The resulting four color image uses only four RGB triplets: $(79, 83, 63)$, $(142, 148, 136)$, $(194, 114, 65)$ and $(143, 182, 210)$. These are indexed by the output of the color quantization step thereby automatically partitioning the image into regions. A region labeling step can then be used to build segmentation masks to pull out regions in the original image which share visually consistent colors. This allows highlighting as a segmentation step or on its own.

Color quantization was widely studied in the past in order to efficiently map images to limited color depth displays which was the original motivation for the median-cut algorithm. However, along the way many other algorithms and modifications of the median-cut algorithm were developed. For early work see [69] [153] [52] [36] [91] and [89]. For more recent work in this area consider [34] [225] and [221]. In [34] a new, fast algorithm for performing quantization is introduced as an aid, in part, to content-based retrieval. The goal in [225] is similar.

Figure 1.9: Color quantization. The original 24-bit RGB color image (left) is reduced to four representative colors (right). This reduction highlights similar image regions and assigns them to the same color value while trying to preserve as much similarity between the quantized image and the original.

1.3.5 Highlighting with Color Tables

Many images are acquired in a single band and represented in gray scale. Sometimes these images can be hard to interpret as pure gray scale and simple highlighting can be applied by changing the color table [199]. If the acquisition process results in images where there is a definite and consistent mapping between the gray level and some other quantity then color tables can be built which will consistently highlight specific regions when present while also, if desired, diminish regions that are not of immediate interest (hence including color table manipulations in this section).

For example, in medical imaging CT images are single band but are mapped during the acquisition and reconstruction process to Hounsfield units (HU),

$$\mathrm{HU} = 1000 \times \frac{\mu - \mu_{\mathrm{water}}}{\mu_{\mathrm{water}} - \mu_{\mathrm{air}}}$$

where μ is the linear attenuation coeffient representative of the density of the tissue or bone through which the x-ray photons are traveling.

Because Hounsfield units are consistently defined they can be mapped to a specific set of gray level ranges which in turn allows for the creation of color tables to highlight tissue types in various ways. For example, consider Figure 1.10 which is difficult to interpret structurally until a color table is applied that highlights different tissue types from their HU values as in Figure 1.11 or even a compressed color table to highlight bone regions as in Figure 1.12.

Lookup table manipulations are very basic and so widely used that they are not considered a particularly interesting research area. For example, the expected window level and center controls ubiquitous to all modern PACS (Picture Archiving and Communication System) medical display stations [11] are in fact color table manipulations done on the fly to highlight some image regions at the expense of other regions.

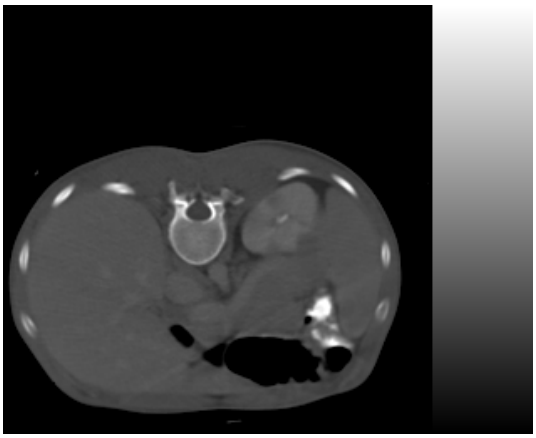


Figure 1.10: Original CT image.

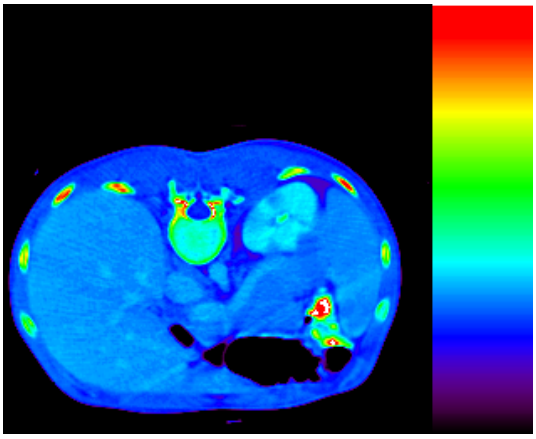


Figure 1.11: CT image with applied color table.

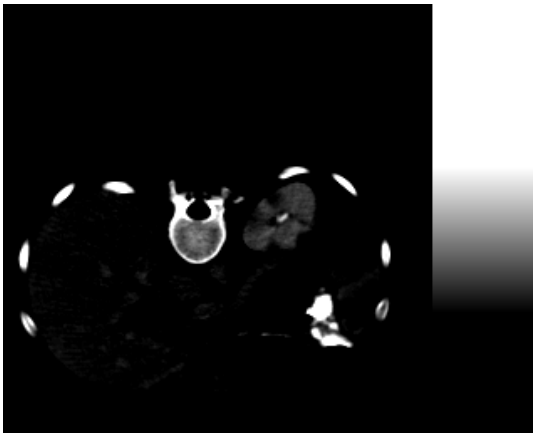


Figure 1.12: CT image with applied color table that compresses the range to highlight mostly bone.

1.3.6 Summary of Hard Image-Based Highlighting

In this section we explored image highlighting techniques that fall into the *hard image-based* category. We looked at edge detection and thresholding, two fundamental image highlighting techniques, and followed that with morphological techniques. Next we considered color quantization and finished the section with some examples of color table manipulations for highlighting. We move now from *hard image-based* techniques to *soft image-based* which highlight images in a way that preserves image content but without understanding that content.

1.4 Soft Image-Based Highlighting

Soft image-based highlighting techniques involve continuous image highlighting which is not based on knowledge of the content or meaning of the image. Many of the techniques in this category are those which are foundational to modern image processing. Specifically, we look at each of the following classes of techniques: contrast enhancement, histogram-based methods, smoothing and sharpening in the spatial domain, smoothing and sharpening in the frequency domain, unsharp masking, and homomorphic filtering.

1.4.1 Contrast Enhancement Techniques

Contrast adjustment is the modification of grayscale image values according to some function, $I' = T(I)$, where the function $T()$ is often based on the histogram of the pixel values in the input image, I. In this section we explore contrast enhancement techniques which are some of the oldest techniques applied to grayscale images. These will make the generic $T()$ specific. Many of these techniques apply to color imagery, indeed, multiband imagery (such as many satellite images), by application per color channel. Here we consider only grayscale images with this fact in mind.

Many imaging systems collect image values as intensities in some integer range, often 8 or 16 bits, as the output of an A/D data acquisition system. For image analysis the full data range is typically used. For display purposes, however, grayscale images, with few exceptions, are displayed

using 8-bit values as this corresponds to more closely to the range of gray values discernable by the human visual system [104]. Therefore, we will content ourselves to examples involving 8-bit data with pixel values in the range $[0, 255]$. For example, this is exactly the situation frequently encountered in remote imaging analysis. Note that this also includes medical images which are, by strict definition, also remote sensing images.

In order to discuss grayscale images and the manipulation of their values for contrast enhancement a quick review is in order. An image is represented with unsigned byte data under the assumption that a 0 pixel is the darkest value and a 255 pixel is the highest value. Typical display conditions map 0 to black and 255 to white though some medical display modes reverse this to make digital x-ray images appear in a manner that more closely matches a film x-ray with bones as white. A raw image is displayed, first by mapping its values to the range $[0, 255]$ and then by showing these digital values on the computer screen as an image. Since virtually all modern displays are 24-bit RGB this is accomplished by setting each channel to the same color values preserving the 255 grayscale range.

Contrast enhancement depends heavily on the histogram of the image values. In this case the histogram is simple to compute unambiguously, we use 255 bins, always, and simply count the number of image pixels with that value in the image. Modification of the mapping function between old pixel values and new pixel values then makes use of thresholds derived from this histogram.

A very common image enhancement technique is that of mapping specified percentiles of the input values to 0 or 255 [66]. For example, a 5% stretch is found from the original histogram by calculating the gray values corresponding to the 5^{th} and 95^{th} percentiles. Call these g_5 and g_{95}, respectively. New gray levels (g') are calculated from the existing pixels (g) using,

$$g' = (g - g_5)\left(\frac{255}{g_{95} - g_5}\right)$$

where g' values less than zero are set to zero. For example, consider Figure 1.13 where the original image pixel values are remapped using a 5% stretch.

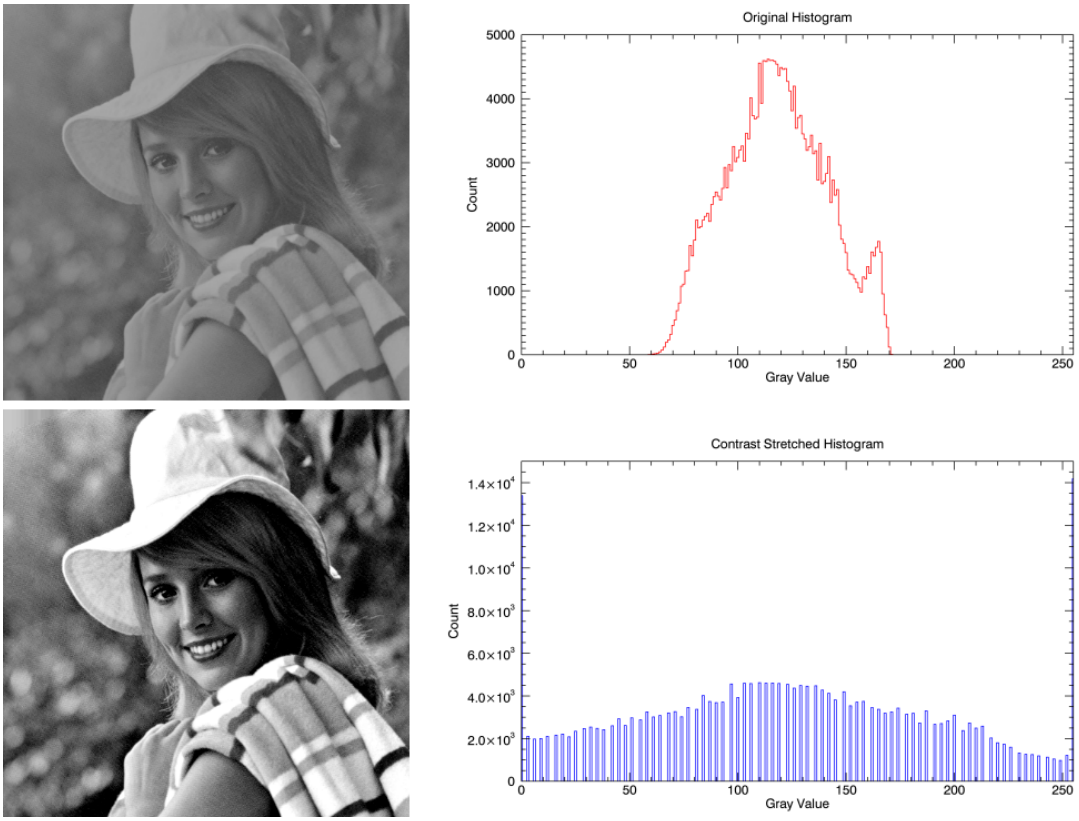


Figure 1.13: Contrast enhancement. The original image is on the top left and the 5% contrast stretch image is on the bottom left. Their respective histograms are on the right. The limited range of the original image is improved by remapping the gray level values from the 5-th percentile to the 95-th percentile along a linear ramp. Values below the 5-th percentile are set to zero and values above the 95-th percentile are set to 255.

The original image is on the top along with its histogram on the right. From the histogram it is easy to see that the gray levels are only using a fraction of the possible range. After the 5% contrast stretch the entire range of allowed gray levels is used producing an improved image. In this case the 5^{th} percentile gray value was 80 and the 95^{th} percentile gray value was 159 so pixels were remapped using,

$$g' = (g - 80)\left(\frac{255}{159 - 80}\right)$$

Remapping of graylevel values based on the original image histogram is the fundamental technique in a wide range of soft image-based highlighting techniques [75] [117] among others. Of particular importance is that of histogram equalization. This equalization may be global or adaptive to the local image environment, as described below. The goal is to adjust the grayscale values so that the histogram makes maximal use (however defined) of the range of grayscale values available to it. This will adjust the contrast increasing it when viewed by the human eye. Histogram equalization is related to histogram matching where the goal is to either match the image histogram to a specified distribution or to that of another image as shown in the example below. Histogram equalization and matching techniques appear often in the literature, for example, see [223] [92] [71] and [163].

As an example of histogram equalization consider the images in Figure 1.14. The original image is in the upper left, then clockwise we have global histogram equalization, generalized histogram equalization which adds non-integer values to the bins based on a region around the pixel under consideration, and locally adaptive histogram equalization. Clearly, each of these images alters the contrast of the original allowing other image areas to be highlighted, albeit sometimes subtly so.

Global histogram equalization seeks to redistribute the image gray levels over the entire allowed range, $[0, 255]$. For locally adaptive histogram equalization each pixel in the image is assigned a new gray value based on its ranking in a region around the pixel. This can be considered

a global equalization applied region by region. Lastly, a generalized histogram equalization acts in the manner of the locally adaptive but uses non-integer values when building the histogram.

When compared to the original image on the top-left, each of the histogram equalization techniques has highlighted regions of the image that were not clear before. For example, the top-right image (global histogram equalization) has has highlighted features in the cameraman's coat, though sometimes excessively so. The lower-left image (locally adaptive histogram equalization) has also highlighted features in the coat but at the expense of producing an unnatural look to the image. Generalized histogram equalization (lower-right) highlighted features in the coat as well but without producing an exaggerated image. Note, the contrast between the coat sleeve and the glove or the appearance of buttons on the coat that were difficult to discern in the original.

Histogram matching was mentioned above along with some references. In Figure 1.15 we show the effect of histogram matching. In this case the Cameraman image (top) is matched to the histogram of the Lena image (middle) to give a new image (bottom). The histograms for each of these images is also shown on the right. Clearly, matching the histogram has highlighted details on the Cameraman's coat that were not visible in the original image. Also, the shape of the matched image histogram closely resembles that of the reference (Lena) image.

Histogram matching calculates the histograms of the input image, i_1, and a reference image, i_2. From these histograms the cumulative histogram is calculated by setting, for each gray level, the output value to the sum of all previous histogram values (finally normalized to 1.0). Call these C_1 and C_2 respectively. Then, for each gray level value, g_1, find the gray value g_2 such that $C_1(g_1) = C_2(g_2)$. This is then the mapping from g_1 in the original image to the new output gray value, g_2.

Figure 1.16 illustrates the remapping process. The cummulative histogram of the original image (right) is paired with that of the reference image (left). For each gray level value, here 100, the cummulative frequency is found (vertical blue arrow up). From this, the corresponding value is found in the reference image (horizontal blue arrow). From this, the new gray level value is read (vertical blue arrow down). This forms the output mapping.

Figure 1.14: Histogram equalization. The original image (upper-left), global histogram equalization (upper-right), generalized histogram equalization (lower-right) and finally locally adaptive histogram equalization (lower-left).

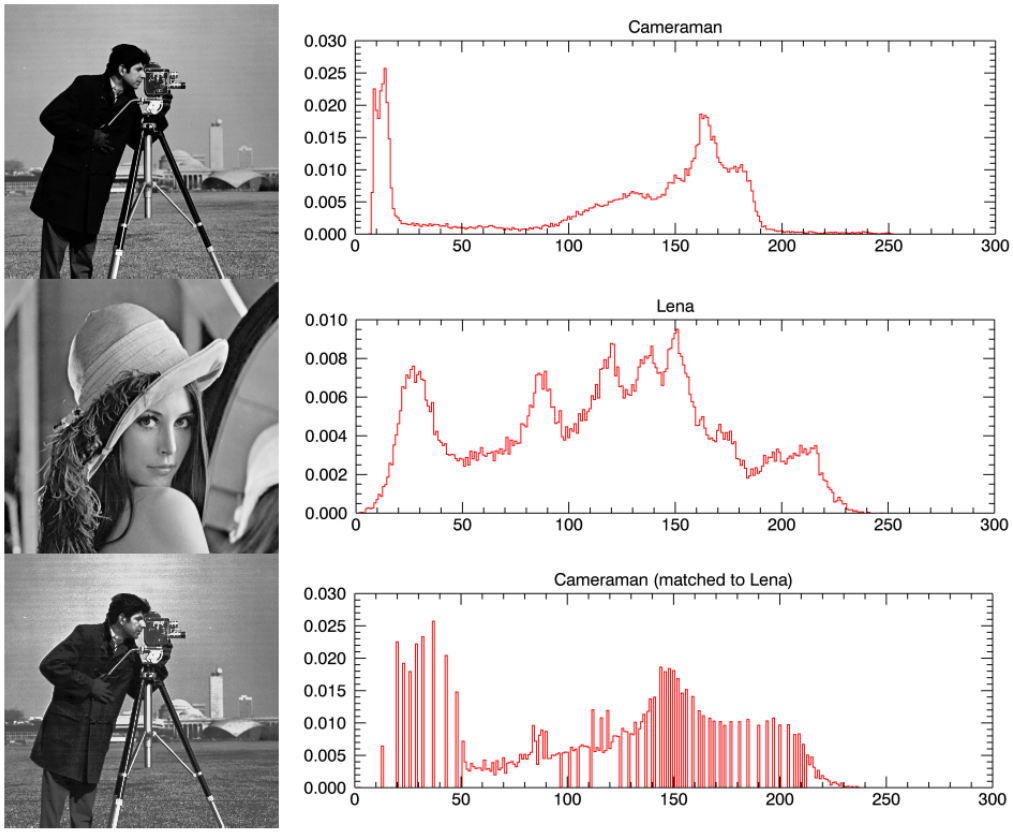


Figure 1.15: Contrast enhancement by histogram matching. The original image (top) is matched to the reference image (middle) resulting in the new output image (bottom). The histograms for each image are given on the right. Notice that the final output image histogram closely matches the histogram of the reference image.

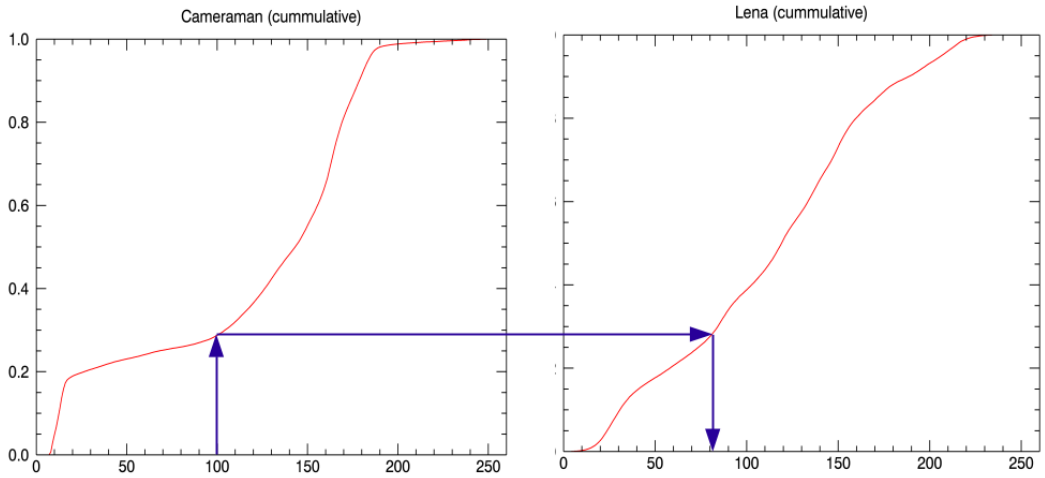


Figure 1.16: The histogram matching mapping process. A gray level value in the original image is mapped to a new gray level in the output image through the cummulative histograms of both the original image (left) and reference image (right) by following the blue arrows.

Histogram equalization and histogram matching are specific instances of a general process of remapping gray levels in an image according to some transformation. Following [75], we can write this transformation as,

$$y(x) = T(x)$$

where a set of input gray levels, x, is mapped to a new set of output gray levels, y. In [75] this transformation is restricted to be monotonic and increasing to preserve black to white gradations but there is no reason for the mapping to be restricted in this way. All histogram techniques generate an effective $T()$ in order to make the mapping.

For example, histogram equalization can be defined mathematically as,

$$y(x) = T(x) = \sum_{j=0}^{x} p_r(g_j) = \sum_{j=0}^{x} \frac{n_j}{n}, \quad x = 0, 1, 2, \ldots, L-1$$

for a set of L gray levels where $p_r(g_j)$ is the probability that gray level g_j is present in the original image. This is estimated from the histogram as n_j/n where n_j is the number of pixels in the input image with gray level g_j and n is the total number of pixels in the image. The output image is created by changing all pixels with gray level x in the input to gray level $y(x)$ in the output.

One wonders if mappings might be found automatically by casting them as optimization problems to minimize or maximize an objective function related to some desired property such as monotinicity or some desired distribution of number of gray values in specific regions of the histogram, etc.

1.4.2 Smoothing and Sharpening Techniques

A particularly common technique is to smooth (blur) an image, perhaps to reduce the visual effect of noise, or to do the opposite and enhance edges to give the image a sharper appearance. For example, see Chapter 5 of [174] for a description of smoothing and sharpening for remote sensing imagery. Noise in an image can be thought of as any artifact introduced by the image acquisition process. This might be cold or hot pixels in a CCD camera which lead to salt and pepper noise, or

thermal effects in the CCD camera itself which adds a background to each pixel that is typically assumed to be Gaussian in nature. The effect of an optics system may also be considered noise through its point-spread function which blurs out image detail.

Why is smoothing or sharpening considered a highlighting technique? For smoothing noise is typically reduced, along with other fine image detail, and this highlights large image features that might be less noticeable without the reduction in fine detail. For example, in a remote sensing image the viewer may be most interested in large-scale features of the land itself. A smoothing operation will reduce the visual impact of smaller scale features thereby making the larger features easier to see. For sharpening the opposite happens. Edges are enhanced which aids the viewer by more clearly highlighting the boundaries between image regions.

Smoothing and sharpening in the spatial domain are specific operations involving convolution of a kernel over the image. These operations may also be performed in the Fourier domain via multiplication but we will consider only convolutional operations in the spatial domain as the net highlighting result is quite similar.

Convolution involving digital images is summarized as follows (ignoring edge effects),

(1) Define a kernel which typically is square and has an odd length, eg, 3x3 pixels.

(2) Starting in the upper left corner of the image, take a 3x3 image patch.

(3) Multiply each pixel in the image patch by the corresponding value in the kernel and sum the results.

(4) Replace the center pixel of the 3x3 image chip in the output image with the value computed in Step 3.

(5) Shift over one pixel and repeat Steps 2 and 3. At the end of the row, move down one pixel and repeat.

With this definition a smoothing operation, also called a lowpass filter, is accomplished by an averaging kernel (see [117]) which can be defined as,

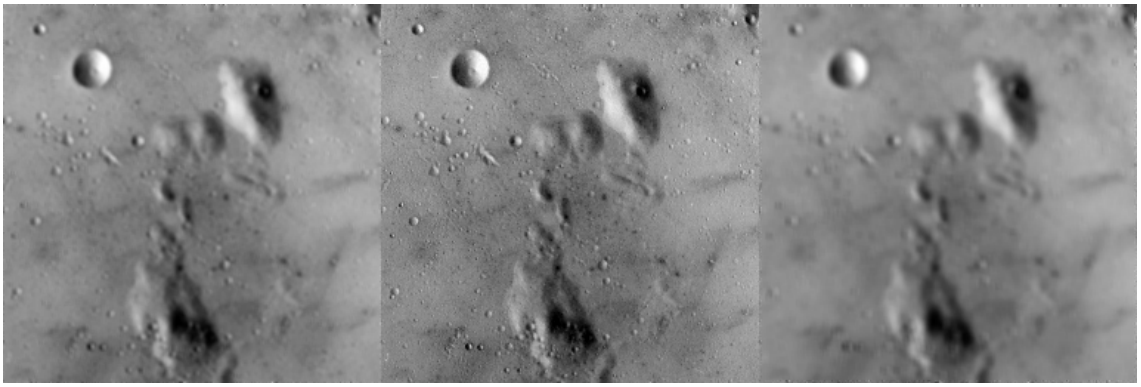


Figure 1.17: Smoothing to highlight large image features. The original image (center) is smoothed using a 3x3 kernel (left) and a 5x5 kernel (right).

$$k_{3x3} = \frac{1}{9}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

for a 3x3 kernel or as,

$$k_{5x5} = \frac{1}{25}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

for a 5x5 kernel.

As might be expected, averaging over ever larger regions (kernels) will result in more and more smoothing of the image simply because the average pulls values towards the mean. This effect is illustrated in Figure 1.17 where the original image (center) is smoothed with a 3x3 kernel (left) and 5x5 kernel (right) to highlight the larger features of the image by reducing or eliminating the smaller features.

A special case which can be placed under the smoothing catagory is a filter that removes noise from the image. As an example, a median filter (see [37] for an advanced version and [146]

Figure 1.18: Median filtering. The original image (center) is corrupted with 10% salt and pepper noise (10% of the pixels have been randomly set to 0 or 255). The image on the right is the result of applying a 3x3 smoothing filter. The image on the left is the result from a 3x3 median filter.

for a separable version) is quite similar to the averaging filters given above but instead of replacing each pixel in the output with the mean of the kernel region it uses the median. The effect of this filter is to remove discrete salt and pepper noise which can come from poorly acquired image data. For example, the center image of Figure 1.18 is corrupted with 10% salt and pepper noise, ie, 10% of the pixels have been randomly set to 0 or 255. The image on the right is the result of applying a 3x3 smoothing filter and the image on the left is the result of applying a 3x3 median filter. Clearly, median filtering highlights the image by removing distracting noise values.

Convolution with a kernel can also be used to sharpen the image by emphasis of the edges. A key tool to enhance edges is the Laplacian (2nd derivative) kernel ([131]) which on its own locates edges but can be combined with the original image to highlight edges as in Figure 1.19 where the original image (center) is sharpened by adding back the Laplacian image (left) to produce, with proper scaling, the sharpened image (right). In this case the Laplacian is defined as,

$$Laplacian_{3x3} = \begin{pmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{pmatrix}$$

which will cause the center pixel of each region to be strongly enhanced while suppressing the impact of nearby pixels. Adding this edge enhanced image back into the original will highlight

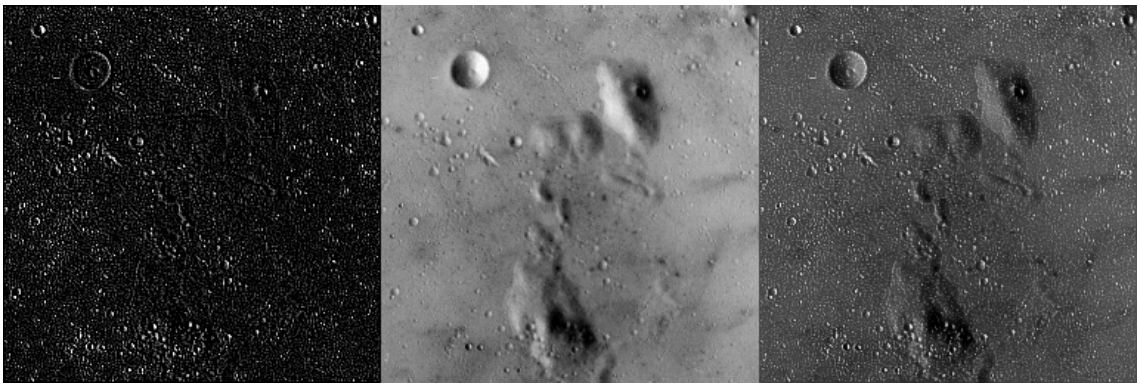


Figure 1.19: Laplacian filtering. The original image (center) is convolved with a Laplacian kernel (see text) to produce the image on the left. This image is added back to the original image and with proper scaling gives the sharpened image on the right. Notice how the edges of the craters are now more pronounced, for example.

edges while preserving low frequency information as seen above.

1.4.3 Highlighting with Unsharp Masking

Unsharp masking, originally developed for film processing, is a highlighting technique which enhances the clarity of digital images ([15], [171], [170], [165]) by making use of a blurred version of the input image. The essence of the unsharp masking algorithm is,

(1) Blur the input image. For example by convolving with an averaging kernel.

(2) Subtract the blurred image from the input to create a highpass image.

(3) Multiply the highpass image by some amount (λ) and add it to the input image.

(4) For computer representation, clip the resulting image to the expected data range.

which can be written mathematically for input image i as,

$$\begin{aligned} d &= i - \text{smooth}(i) \\ i' &= i + \lambda d \end{aligned}$$

where λ controls the amount of sharpening. Application to color images is by band.

The effect of unsharp masking can be seen in Figure 1.20 where the original image is on the left and the sharpened image is on the right. In this case $\lambda = 1$ which highlights image edges throughout the image increasing the visual clarity and making fine details stand out.

One difficulty with unsharp masking, however, is that strong enhancement has the undesirable effect of exaggerating noise in the image as well. Mitigating this effect is the focus of much of the research in this area as evidenced by the papers cited above.

The typical way in which this issue is addressed involves making the λ value locally adaptive. When λ is a constant it is applied to each pixel of the input image equally. It is desired to make λ more intelligent and aware of the local region around the pixel currently being considered. If the local region is smooth and without strong edges then the enhancement can be greater than regions which vary wildly. A simple approach is to consider the standard deviation of pixel values in a local

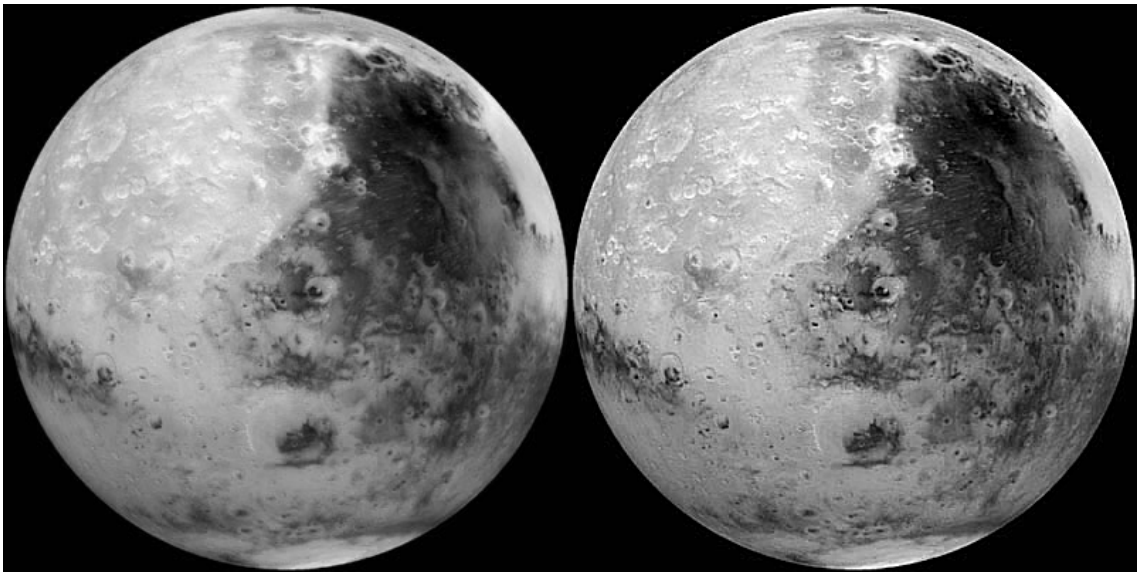


Figure 1.20: Unsharp masking. The original image (left) is sharpened by unsharp masking with $\lambda = 1$ (see text) to produce the sharpened image on the right. Notice the distinctness of the features, especially those of the upper left part of the right-most image.

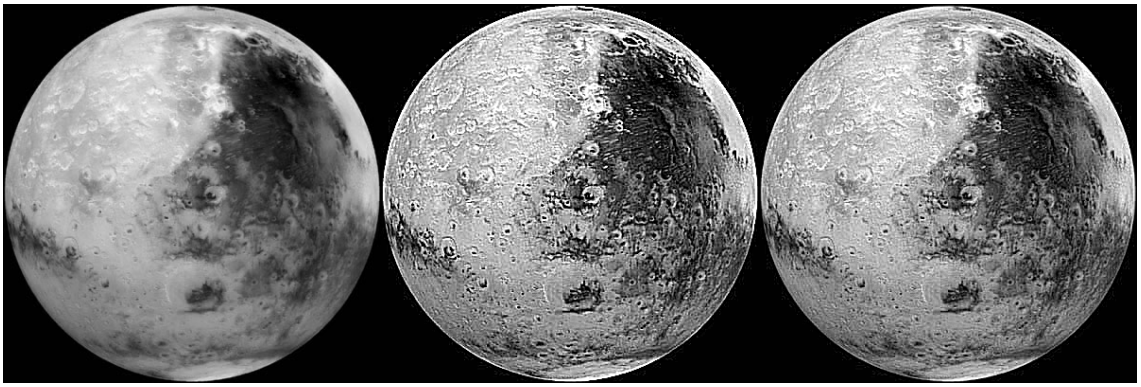


Figure 1.21: Locally adaptive unsharp masking. The original image (left) is sharpened with standard unsharp masking (center) and locally adaptive unsharp masking (right) which sets λ according to $\lambda = \lambda_0(1 - \log(\sigma)/5.6$ where σ is the standard deviation of a 5x5 kernel convolved over the input image. Notice how the image on the right highlights small image features without exaggerating them as in the center image.

region and use this to determine the fraction of the desired amount that is used. In this case, λ is now a function of a 5x5 region around the current pixel where the constant desired amount (λ_0) is modified by a fraction from 0 to 1 based on the standard deviation of the 5x5 region,

$$\lambda = \lambda_0(1 - \log(\sigma)/5.6)$$

where the factor of 5.6 is specific to byte valued pixels.

The effect of locally adaptive unsharp masking can be seen in Figure 1.21 where the original is on the left, standard unsharp masking is in the center, and locally adaptive unsharp masking is on the right. In this case $\lambda_0 = 6$ which results in strong artifacts in the center image but fewer in the image on the right. This image is best viewed on a computer.

1.4.4 Highlighting with Homomorphic Filtering

An image may be viewed as the product of two terms typically called "illumination" and "reflectance" so that image I is defined by $I = ir$ where i is the illumination component and r is the reflectance component. If an image is considered in this way then it becomes possible to operate on these components individually in the frequency domain through the log of the original image,

$$\log(I) = \log(ir) = \log(i) + \log(r)$$

giving,

$$\begin{aligned} \mathcal{F}\{\log(I)\} &= \mathcal{F}\{\log(i)\} + \mathcal{F}\{\log(r)\} \\ &= F_i + F_r \end{aligned}$$

as the Fourier transform of the input image. This separation allows for the independent filtering of illumnation and reflectance components. To recover the filtered image one need only apply the inverse transform, sum the components, and apply the inverse log,

$$\begin{aligned} s &= \mathcal{F}^{-1}\{HF_i\} + \mathcal{F}^{-1}\{HF_r\} \\ I' &= e^s \end{aligned}$$

where H is a filter function applied in the frequency domain. This approach is known as homomorphic filtering [205].

The equations above specify the illumination and reflectance components of the input image. In order to use homomorphic filtering as a highlighting technique it is not necessary, nor often possible, to actually know the separation between these two components. However, in general, low frequency components, corresponding to smooth regions of the image, are associated with illumination while high frequency components, corresponding to edges and boundaries, are associated with reflectance. Therefore, if the filter function H is well parameterized it can operate simultaneously on the illumination and reflectance components without actually determining these components. This is the key observation that makes homomorphic filtering a useful technique in practice.

For example, consider a two-parameter highpass Gaussian filter for H,

$$H(u,v) = (\gamma_H - \gamma_L)[1 - e^{-cD^2(u,v)/D_0^2}] + \gamma_L$$

where $D(u,v)$ is the distance from the center frequency in the 2D Fourier space to any location u, v and D_0 is a user-selectable cutoff frequency. The parameter c controls the slope of the transition region. Lastly, γ_L and γ_H are two additional frequencies bounding the transition region. Consider

Figure 1.22 where the original image is on the left and the filtered image is on the right. The filter reduced the dynamic range of the image and also highlighted details on the cameraman's coat that were not visible in the original image. Homomorphic filtering remains an area of active research, for example, see [3], [58], and [193].

1.4.5 Summary of Soft Image-Based Highlighting

In this section *soft image-based* highlighting techniques were presented. Many of these techniques are fundamental to the field of image processing and are represented by a large number of variations on basic themes. The hallmark of these techniques was hopefully evident in what was presented, namely, that viewer attention is drawn to objects in images through soft or continuous changes to the pixels which preserve image content while highlighting some of that content. However, these techniques are blind in that they are not based on any judgement about the actual content (objects) in the image. We now move on to look at *hard agent-based* techniques that are not blind but instead base their highlighting on an understanding of the image content that has come from an outside agent.

1.5 Hard Agent-Based Highlighting

Highlighting applied to an image as the result of object understanding from an outside agent or oracle falls into the *hard agent-based* category. In this category, the agent has, using unspecified means, some understanding of the content of the image and the hard highlighting applied is based on that understanding. The agent may be another person who, as an expert, has already viewed the image and determined where highlights should be placed. Or, the agent may be the output of a software classifier which has determined locations of objects of interest with enough certainty that the highlight is placed at those locations. This section considers agent-based highlighting in two areas as representative of this approach. These areas are medical imaging and remote sensing.

Figure 1.22: Homomorphic filtering. The original image (left) is simultaneously highlighted while the dynamic range is reduced via a homomorphic filtering operation to produce a new image (right). Notice that the image on the right has a reduced dynamic range while simultaneously increased detail as can be seen in the cameraman's coat.

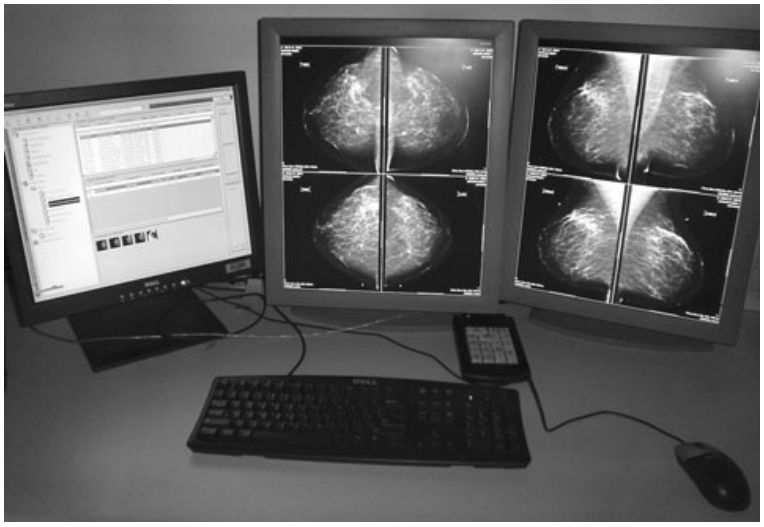


Figure 1.23: Image presentation for a typical mammography workstation. The two view mammograms (MLO and CC views) are shown for the current and previous visit.

1.5.1 Hard Agent-Based Highlighting in Medical Imaging

Hard agent-based highlighting schemes have been in use in medical imaging for some time. With the large demands placed on radiologists to view images, and these demands are increasing, there is a desire for an ability to either prescreen or pre-detect lesions in images. Hard agent-based highlighting has been most prominent in mammography and lung nodule detection so we will focus on those two areas.

In mammography, a radiologist will typically view eight images as a time: the current two-view mammogram of the left and right breast and the previous two-view mammogram. A two-view mammogram typically consists of a cranio-caudal view (CC) (head looking down to the feet) and a mediolateral oblique (MLO) view (looking from the center of the chest down to the outside edge of the chest). For example, a typical mammography viewing station might lay out the images as in Figure 1.23 (image from [169]) so that the radiologist can quickly scan the previous and current mammograms looking for changes.

The most common agent used in hard highlighting of medical images is the output of a classifier applied to the images. This is known as computer-aided detection or CAD. These classi-

fiers, typically neural networks or support vector machines, exact algorithms are often proprietary, generate a series of decision locations on the image where a human should look.

The highlights (or prompts) used by CAD systems consists of hard circles around a suspicious area, symbols placed over the area, or arrows pointing to the area that should be examined. Figure 1.24 illustrates several such highlighting techniques with images taken from the literature. These images represent actual prompts from actual CAD systems.

In Figure 1.24 CAD prompts from as early as 1993 to a current state-of-the-art system are shown. In essence, these prompts have not changed in that over twenty year span. A hard prompt is used to outline the suspicious area or a marker is placed over the area with the symbol used a function of the type of lesion suspected, either a mass or a microcalcification. CAD is also used frequently in the detection of lung nodules in both CT and chest x-rays. The prompts used for lung lesion detection are very much the same as those used in mammography.

CAD systems are in wide-spread use today. The efficacy of CAD has been demonstrated repeatedly in multiple studies (see [65] [37] [18] [181] and [5]) but is not without an effect on the radiologist (see [63] [73] [227] and [162]). In particular, as [227] points out, highlights, especially false positive highlights, affect the radiologist and reduce the likelihood that missed lesions will be detected.

Some CAD systems, especially those used in research, are interactive and allow users to query locations in order to be told what the likelihood is that a lesion is in that region [186]. However, even with these systems the highlight is hard and consists of an outline, perhaps generated by region growing, indicating a likely location for further examination.

A particularly interesting example of hard agent-based highlighting in medical imaging is found in Litchfield 2010 [123]. It has been demonstrated, for example in [110], that radiologists will fixate for longer periods of time even on lesions in medical images that are missed. This caused Litchfield in [123] to explore the question of what happens when a radiologist, in this case a novice, scans chest x-rays looking for pulmonary nodules by following the path taken by an expert radiologist. The result is that the gaze path of the expert, in this case the agent, when shown to

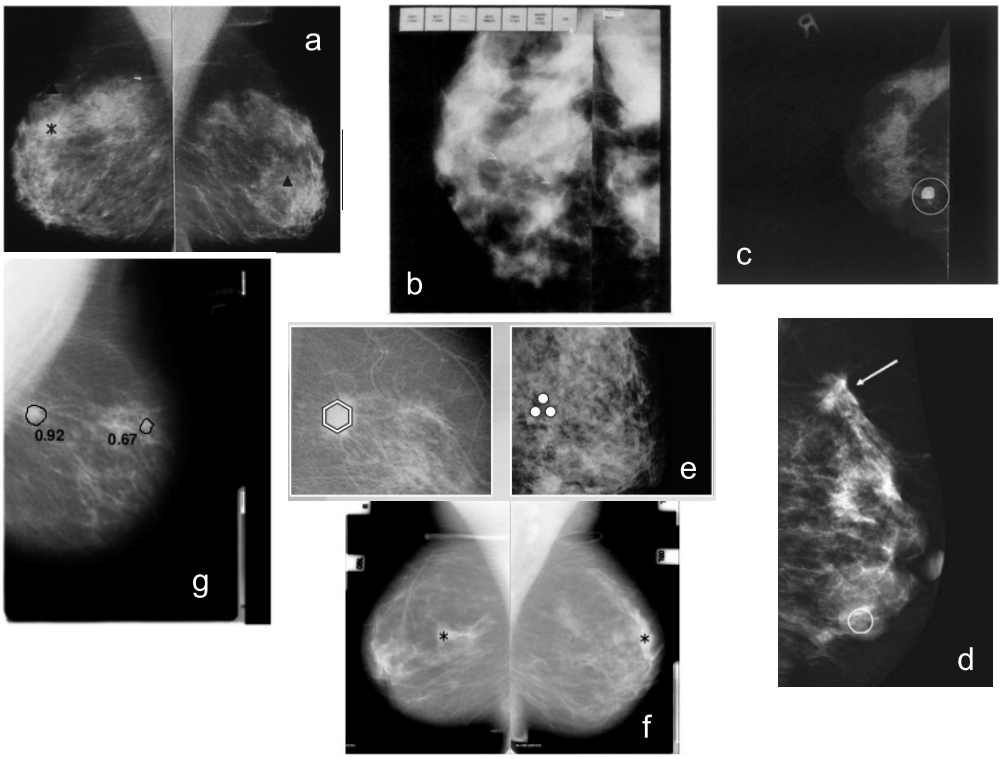


Figure 1.24: Typical prompts used in mammography CAD. Sources: (a) [65], (b) [72], (c) [118], (d) [162], (e) [67], (f) [184], and (g) [186]. Note that (b) represents a system from 1993 while (e) represents a current state-of-the-art system. In essence, the prompts have not changed in twenty years.

the novice increases the novice's ability to locate lesions. Figure 1.25 illustrates such a path and how it is superimposed over the chest x-ray image.

On the left is the chest x-ray image to be searched. In this example there were three lung nodules which are marked with a hard highlight. The gaze path of an expert radiologist was tracked using an eye-tracker and was presented as a hard agent-based dynamic highlight to a novice radiologist (the viewer) in order to guide the viewer to regions that were investigated by the expert while still allowing the viewer the freedom to look elsewhere. The image on the right is the same chest x-ray with the white line marking the gaze path of the viewer and the gray line highlighting the gaze path of the agent (expert). The highlighting is dynamic in that only the path for the last 500 ms of gaze time is shown on the image. This form of hard agent-based highlighting was demonstrated to improve the performance of viewers searching for lung nodules.

1.5.2 Hard Agent-Based Highlighting in Remote Sensing

The term "remote sensing", if applied strictly, encompasses medical imaging, but in common use it refers to imagery of the Earth typically taken from space by a satellite but also includes aerial imagery. The types of imagery used in remote sensing varies considerably from panchromatic (grayscale) single band images to hyperspectral imagery with potentially several hundred bands. Most of the systems mentioned in this section fall into the panchromatic or multispectral category. A multispectral image has more bands than a visible light image and typically includes several infrared bands. Often, a panchromatic image of high resolution is combined with a color image of lower resolution multispectral imagery to create a pansharpened image which retains both color and resolution. Much of the imagery shown on web-based maping sites like Google Maps likely falls into this class.

The agent used in remote sensing imagery is usually a pixel-level classifier, for example [192] [144] and [143]. This classifier attempts to assign a class label to each pixel of the image. Then a thematic map is often created showing in cartoon form the classes by color as is seen in Figure 1.26 where each color indicates the output of a classifier which has assigned a class to each pixel of

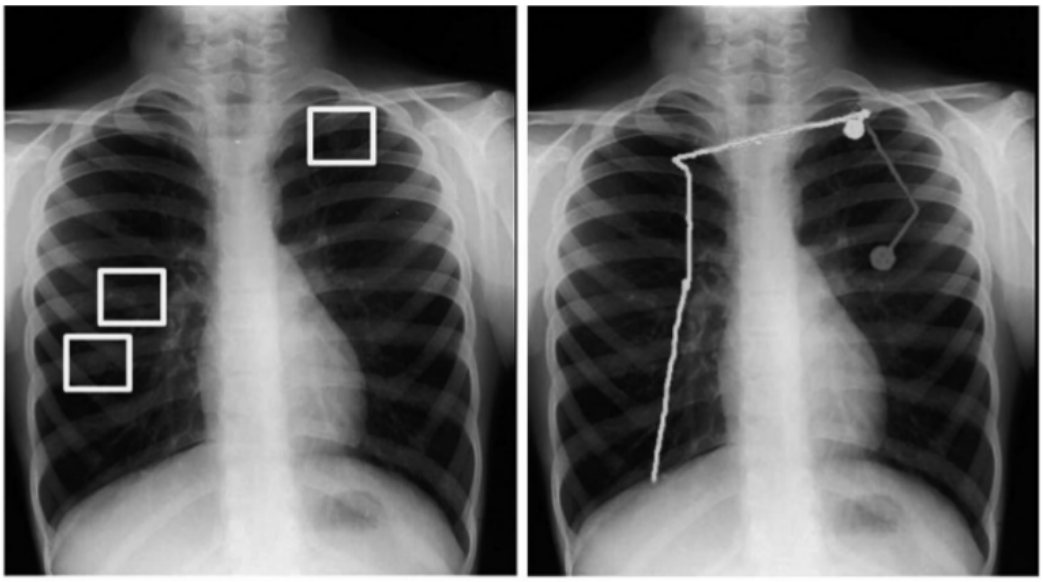


Figure 1.25: Improvement of a viewer's lesion detection by following the gaze path of an outside agent. The chest x-ray on the left has hard highlights over three lung nodules. The same image is on the right along with the gaze path (last 500 ms) of a viewer searching for lung nodules. The white line is the path followed by the viewer while the gray line is the path followed by the outside agent. In this case an expert radiologist. This form of hard agent-based highlighting was demonstrated to improve the detection abilities of novices. From [123].

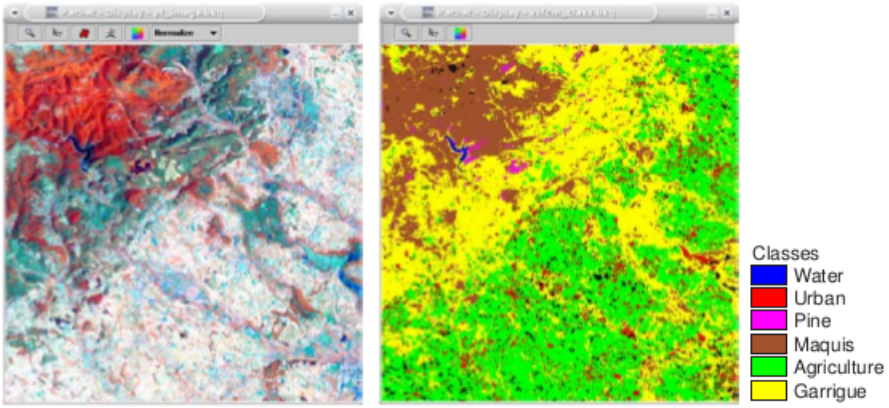


Figure 1.26: A thematic map. The input false color image (left) is classified, pixel by pixel, to assign each pixel to a class. From this the thematic map (right) is created which has assigned each pixel to one of six classes.

the image. This map is from [126] and shows a false color image (multispectral with three bands selected) on the left and the thematic classification image on the right.

Note that a thematic map does not attempt to indicate objects in the image, it is only concerned with labels applied to pixels. While systems that attempt to build object-level classifications of items in remote sensing images have been explored or are in development, either through classifiers or via rule-based engines [202], it is the ubiquitous thematic map that will be our focus here.

Thematic maps indicate the most likely class but they do not provide any notion of classifier uncertainty. Presentation of the thematic map, along with the original image, has been used for some time [8] [214].

The use of thematic maps and the lack of indication of classification certainty they provide has lead to a small literature on ways to use highlights of some type along with the map to give viewers an active indication of uncertainty. These highlighting approaches fall into four general categories: color manipulations, dynamic visualizations/animations, probability map and entropy map displays, and 3D plots of classifier decision boundaries.

In color manipulations the colors of the thematic map are modified to indicate increasing levels of uncertainty in the class assignment. In Hengl [84] colors are increasingly moved towards

white as uncertainty increases producing output as in Figure 1.27 where the example shows three class predictions on the left and the same predictions on the right with whiter colors used to indicate uncertainty. Note that these predictions are for a continuous output (percent of sand, silt and clay in the soil) but do not include any uncertainty in the prediction. When the uncertainty is included by whitening, the images on the right are obtained. The whitened images clearly show the uncertainty in the predictions and in particular that the clay prediction is very uncertain.

Color saturation is also used to indicate classification uncertainty [128]. The saturation of the color changes with the uncertainty (lower probability of class membership). Similarly, hue may also be used [213].

Animations present a time-varying image to the viewer in order to highlight the classification uncertainty. In Van der Wel [213] a probability map is toggled with the thematic classification map so that persistence of vision will make both present to the viewer. In [22] the level of uncertainty is animated through different thematic class labels appearing according to their probability so that highly certain pixels are mostly constant in color while completely uncertain pixels would be changing frequently.

The probability map used in Van der Wel [213] for animation may be displayed on its own. In this case the gray level indicates classification certainty as to class membership as seen in Figure 1.28 which can be viewed as a extreme form of the whitening approach in used by Hengl above [84].

If the classifier is multiclass, which most thematic output maps are, then it becomes possible to compute an entropy over the output class probabilities in order to produce an entropy map [213] which may also, with proper scaling, indicate visually the classifier certainties as in Figure 1.29 where the entropy is calculated from the per pixel probability of class membership vector using,

$$E = -\sum_{i \in C} P(C_i|x) \log_2 P(C_i|x)$$

with $P(C_i|x)$ the probability that the given pixel x belongs to class C_i.

Both the probability map alone or entropy map alone qualify as *hard agent-based* highlights

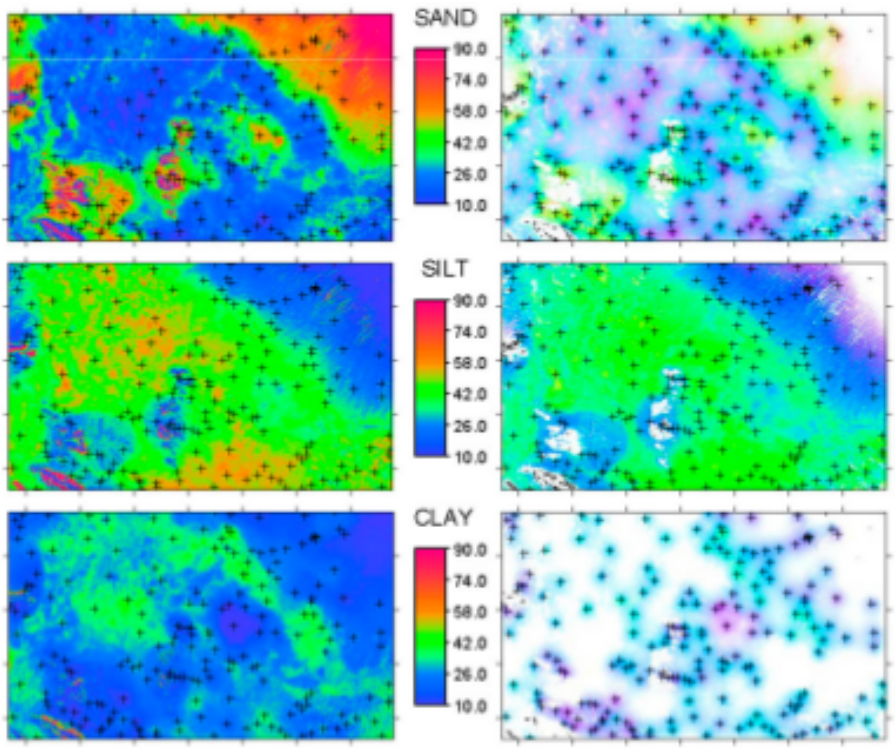


Figure 1.27: Uncertainty display using whitening of the color in the thematic map to indicate increased uncertainty in the pixel classification. The black crosses are markers used in the classification and are not part of the uncertainty display.

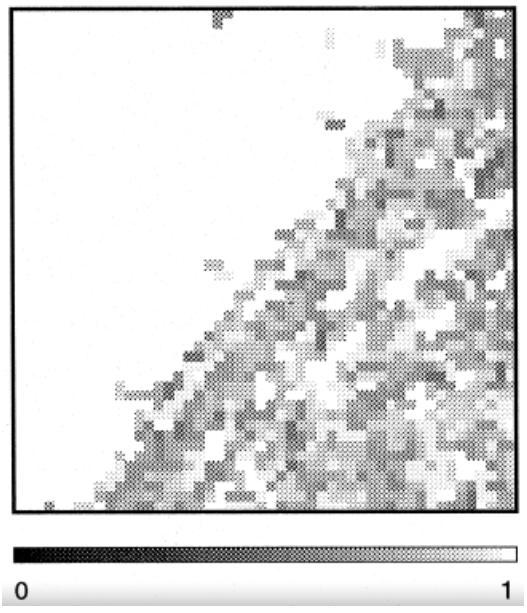


Figure 1.28: Probability map display. In this display the gray level indicates the uncertainty in the pixel level classification at that point. The less certain the classification is the darker the pixel.

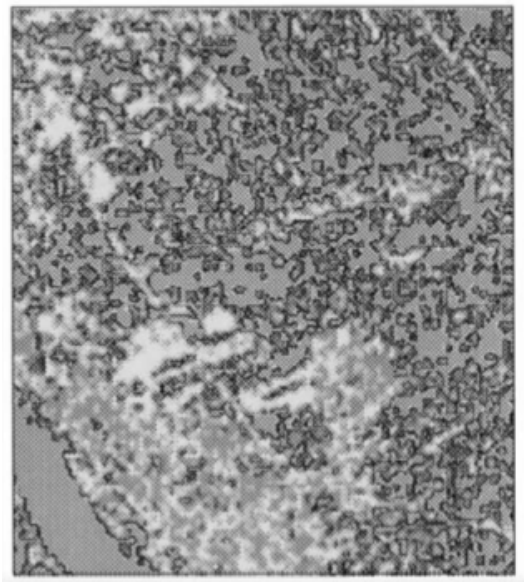


Figure 1.29: Entropy map. The entropy of the classifier outputs per class per pixel are displayed as gray scale values.

according to our taxonomy because they do not preserve actual image data but instead remove that data and show an image which is correlated with classifier uncertainty, not pixel values.

The last highlighting technique is a separate visualization of the classifier decision boundaries in 3D for up to three classes. The boundaries are points or isosurfaces as in [217] and [126]. For example, [126] uses isosurfaces as in Figure 1.30 which can be rotated in 3D to give the image viewer a sense of the overlap between classes. Naturally, this approach would be most useful with the thematic map, itself a highlight, and the original image present in a single interactive display.

Highlighting techniques based on the thematic map are commonly used, when uncertainty is used at all, in remote sensing images. None of these techniques are specific to objects in the image, the agent here has classified each pixel and the uncertainty of class membership is what is used in the highlight itself.

Some of these techniques were dynamic. They were included in this section because of their association with remote sensing images. Other dynamic highlighting techniques are considered below in Section 1.6.

1.6 Dynamic Image Highlighting

In this section we review some attempts at dynamic highlighting beyond those referred to in Section 1.5. These techniques all feature effects that change in time. Some of these techniques are animations while others are not visual but make use of other sensory information. First we consider the area of sonification as it relates to images. Sonification [101] [85] [20] [188] is a small but active research area which seeks to render complex data as sound in order to improve perception of information within the data. When applied to images the sonification can be regarded as an extreme form of highlighting. Second we look at the small literature that deals with animations or other visual cueing, often in response to motions of an expert. This form of dynamic highlighting is most often used in teaching where the goal is to increase the performance of the student by drawing the student's attention to areas of the image deemed important by the expert.

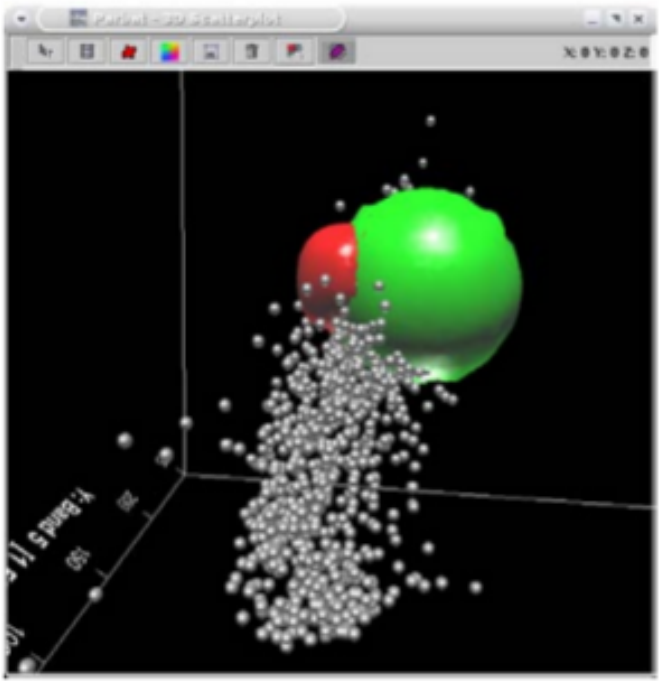


Figure 1.30: Isosurface display of class boundaries. The isosurfaces enclose regions representing the boundaries between classes for a particular classificaion scheme.

1.6.1 Highlighting via Sonification

As primates, humans rely greatly on our sense of sight, so much so that we sometimes downplay the value that might be gained by our other senses. Human hearing is quite acute and therefore sound is a natural target for explorations of other means by which complex information may be presented. In this section we consider highlighting, taking some liberty with the definition of the word, of images through sound. Some of the work referenced in this section has the aim of making the visual world more accessible to people who are visually impaired while some seeks to expand the range of senses used in data analysis.

In [108] the authors develop a framework to translate imagery into sound and speech. Of interest is the color model that assigns each RGB value to a combination of two sounds of differing intensities. As the user moves over the image with a mouse the color of the area under the mouse pointer is converted to the combination of these tones. The system also supports a speech module which enables the user to ask simple questions using "what" and "where' words and to be told predefined answers to these questions. Current work in deep neural networks includes considerable effort in the automatic generation of textual descriptions from images. This seems a natural fit to such a system should it be updated.

Another approach to sonification of images is to predefine a path [62] or a tour [99] which flows through the image allowing set portions to be mapped to sound. A raster scan of an image enabled users in their tests to perceive spatial relationships between objects in the images. A transfer function (not described in [62]) maps a region of the image around the current path position to a sound. The system described in [99] maps the entire image, pixel by pixel, to a series of sounds which, it is claimed, are able to be interpreted by the listener, with practice, as a mental image. The goal was a wearable system for the visually impaired.

In [136] the textual information in an image is mapped to sound through cepstral features [152] which are based on the inverse Fourier transform of the log of the Fourier transform (or power spectrum) of an input signal. Here the signal is based on the pixel values of the image and hence

the texture present in the image. Through an ad hoc scan pattern through the image the authors develop an output stream of sounds which can be mapped back to the texture. Experiments with basic shapes enabled listeners to detect these shapes and textures in test images.

The examples above give a flavor of the kind of work done in sonification of images in order to highlight specific components or regions of the image. Broadly speaking, these explorations fall into one of two major groups: sonification as an aid to the visually impaired or sonification as an aid in complex data analysis. For the former, see examples such as [224] and [188].

The authors of [61] ask the question: should sound be part of standard data analysis software? They then present means for the sonification of data formats like plots and histograms that are commonly encountered in routine data analysis. In [85] the same approach to complex data is explored at length. However, neither of these present a means for the sonification of images. In [222] the authors present a framework for sonification of images involving preset paths which they refer to as scanning or probing with a pointer over the "inverse spectrogram" by which they mean a 2D image with a time step on the x axis (the direction of "play") and frequencies (spatial?) in the vertical axis. Lastly, in [86] we see a system for the sonification of multi-channel image data where the image data is a stack of fluorescence microscopy images.

1.6.2 Highlighting via Dynamic Image Cueing

Dynamic image cueing is a technique most often used in training where the eye movements of an expert are used to cue novices in order to aid their learning of a difficult visual search task. For example, in [204] radiologists were trained by the application of a visual highlight in a mammography image that corresponded with the location viewed by an expert radiologist. The eye motions of an expert radiologist were recorded during a scan of paired mammography images with an eye tracker. Then, novices viewed the same images also with an eye tracker. As the novices scanned the images the locations viewed by the expert were subtly manipulated, always in the peripheral field of view of the novice, to draw attention towards the location. The highlighting consisted of spatial modulation of the region associated with the expert but within the peripheral

vision of the novice. See [16] for a complete description of the technique.

In [164] and [9] a static mammogram was turned into a video sequence by remapping image pixels frame to frame with the motion implied in the mapping based on the pixel intensity. This means that brighter pixels were shifted further that dimmer pixels with the aim of making the static intensity variation a variation in the amount of motion seen by the viewer. In this approach no outside oracle is used, the effect is simply to add motion to the static image to take advantage of our acute sensitivity to motion within our field of view, a prime evolutionary advantage. In their experiments, viewers were more likely to perceive calcifications in the moving images than in the static images.

1.7 Discussion and Areas for New Research

This survey examined the wide field of image highlighting techniques. It organized this area into several key groups: *soft image-based*, *hard image-based*, and *hard agent-based* as well as into *static* or *dynamic* highlights. Key techniques in all of these areas were mentioned and illustrations were given when possible. The essence of the categories is that highlights may be subtle and preserve image information (*soft*) or be severe and remove image information, even to the point of not leaving any original image information at all (*hard*). These soft and hard highlighting techniques may use the pixels of the image without understanding of the image contents (*image-based*) or make use of an outside agent which does understand the objects present in the image (*agent-based*).

Soft agent-based highlighting is missing from this survey. It is precisely this quadrant of our taxonomy that is available for future research. See Chapter 2 for details motivating new research into this area. See Chapter 3 and Chapter 5 for the results of recent experiments into this area along with a series of proposed experiments related to soft agent-based highlighting.

Lastly, in Section 1.4.1 we discusssed the posibility of creating image grayscale mappings in response to a minimization (or maximization) of some objective function. This would be a generalization of the histogram equalization and matching techniques and it also appears to be an area available for future research.

Given the depth and breadth of image highlighting it is hoped that this survey will have helped to organize the techniques for the reader and to have adequately covered enough techniques to make the distinctions of the taxonomy employed clear.

Chapter 2

Limitations of Hard Agent-Based Highlighting

The previous chapter presented examples of hard agent-based highlighting. In this chapter we examine known limitations of this approach which we believe motivates an alternative approach, *soft agent-based highlighting*, which has not yet been studied in the literature.

The application of hard agent-based highlighting to medical imaging and its effect on human readers of medical images has been extensively studied. Several key papers in this area are here reviewed in detail as they demonstrate important weaknesses and pitfalls associated with the use of hard highlighting.

2.1 Hard Agent-Based Highlights Inhibit Detection of Non-Highlighted Targets

Krupinski [111] was an early study in this area focusing on the effect hard highlighting has on the performance of radiologists viewing chest x-ray images. This study consisted of a series of carefully controlled experiments involving hard circle ROIs placed around a lesion in a chest x-ray. In particular, each experiment made use of 40 chest x-ray images, 20 that were tumor-free and 20 with a solitary simulated tumor at $< 50\%$ detectability (how this was determined was not described). The images were digital and displayed full size on a 17 inch video monitor. The simulated tumors had a Gaussian edge profile (ie, faded into the image without an abrupt transition) and had diameters ranging from 0.8 to 2.0 cm. The viewing distance was 70 cm. A circular cue was placed on each image over the tumor or an anatomically similar region if the image

was tumor free. From the fixed viewing distance the circle (black, 2 pixels wide) covered a 5 degree region. The circle was randomly offset a few pixels from the center of the tumor to avoid placing any tumor in the exact center of the circle. Two complete sets of images were made, one with a circle cue on all images and another with no cues present. Images were presented to viewers in a block design. Test images were assigned randomly to blocks of 10 images, each having 5 tumor-free and 5 tumor-containing images. Each block contained either cued or noncued images without mixing.

All experiments followed the same procedure. Viewers were radiology residents with an average of 2.5 years of experience reading radiographs. Images were viewed in half-hour sessions separated by an average 2.5 weeks each. Each image viewed was preceded by a precue image. This precue image had a constant gray background set to the average pixel intensity of all the test images. The precue itself was a pair of horizontal black lines 1 cm long and placed tangentially to the location where the edge of the circle cue would appear when the test image was displayed. This means that the precue lines covered the top and bottom edges of the place where the circle cue would be seen. The precue was used even in the noncued images and it appeared in the location where the cue would have been placed.

The display sequence was as follows: precue image, test image (cued or noncued), mask image, gray image. In this case the precue image was shown until the viewer pressed a button on the joystick. The viewers were instructed to fixate on the precue before pressing the button. The test image then appeared for 200 ms and was replaced by a random noise pattern image which was also presented for 200 ms. Finally a uniform gray background image (matching the mean intensity of the test images) was shown while the system loaded the next precue and test images.

After viewing the test image the viewer reported whether the location contained a tumor or not by selecting from a five point scale: 5 = tumor, definite, 4 = tumor probable, 3 = suspicious, 2 = no-tumor, probable, 1 = no-turmor, definite. Viewers went through a set of 20 practice images before viewing any actual test images. Performance was measured using the area under the ROC curve (Az) and the differences between Az values were assessed using ANOVA.

Of the five experiments described in Krupinski [111] it is only Experiment 5 that directly

concerns us here so we will quickly summarize Experiments 1 through 4 and then cover Experiment 5 in more detail.

Experiment 1 measured the baseline performance of 5 radiology residents in detecting subtle tumors in chest x-rays and determined whether precuing followed by cuing with the circle ROI was effective in enhancing detection performance. The precue and cue were compared to performance using free search. The results indicated that precuing (P) was effective over free search (FS) with mean Az free search $= 0.502 \pm 0.049$ ($\bar{x} \pm SE$) and mean Az precue $= 0.608 \pm 0.018$. Precue plus cue ($P + C$) was effective over precue alone with mean Az precue + cue $= 0.767 \pm 0.019$. Experiment 2 repeated Experiment 1 but changed the view time from 200 ms to 2000 ms. A similar ordering of mean Az such that $P + C > P > FS$ was found.

Experiment 3 tracked the eye movements of the viewer with and without the circle cue (recall that the precue was always present). The results indicated that the circle has a small but significant effect on reducing the visual area searched for the target with fixations covering, on average, 18% of the region when the circle was present and 24% when it was not.

Experiment 4 considered whether processing of information outside the circle ROI was inhibited by the presence of the ROI. Therefore, regions of the image outside the circle ROI were masked (set to gray) in concentric rings of increasing size covering 25%, 50% and 100% of the region. No significant effect was seen on the precue + cue detection performance.

Experiment 5 examined what happens when viewers were asked to report on a tumor (present, not present) which was placed near to but *outside* of the circle ROI while still reporting on the tumor that may or may not be within the ROI. In this case, half the tumor-containing images ($n = 10$) and half the tumor-free images ($n = 10$) had a simulated distractor tumor placed outside the circle ROI but within 2.5 degrees of the boundary of the 5 degree ROI. No outer circle was used to mark the end of the outside tumor region.

For this study four radiology students participated and images were presented for the precue only and precue followed by cue conditions. Within each block were images with and without the distractor tumor. Viewers were asked to make two decisions, one about the outside tumor and one

Decision	Inside	Outside
TP (P)	0.59	0.35
TP (P+C)	0.81	0.18
FP (P)	0.44	0.16
FP (P+C)	0.36	0.15

Figure 2.1: Proportion of true-positive (TP) and false-positive (FP) reports for the inside and outside ROI zones for Experiment 5 of Krupinski [111].

about the inside tumor. The presence or absence of the outer and inner tumors was independent.

The results of Experiment 5 are in summarized in Figure 2.1 by looking at the proportion of reports that were true-positives versus false-positives for precue-only (P) and precue followed by cue (P+C). A z test for proportions showed a significant difference between the true-positive reports ($p < 0.01$ inside or outside) but no significant difference for false-positives.

The results of Experiment 5 clearly indicate that the presence of the circle ROI inhibited detection of the noncued tumor target. When the circle was present fewer tumor targets in the outside region were detected. This was true whether or not a tumor was reported inside the circle region.

The study reported in Krupinski [111] illustrated several key effects on viewers when hard agent-based highlighting was used. One was the improvement in tumor detection seen in other CAD experiments (see [65]) when tumors were located within the highlight. The other, potentially more harmful, effect was the decrease in tumor detection for nonhighlighted tumors. In the case of medical imaging, missing a target might literally mean the difference between life and death.

2.2 Hard Agent-Based Highlighting Quality Affects Detection of Non-Highlighted Targets

One might argue that the stringent viewing conditions of the Krupinski study are unrealistic in a hospital setting where radiologists are free to view images for as long as they like and where they are free to use viewing tools like zoom and pan. Will the missed nonhighlighted targets effect persist in that environment? For an answer, we now consider the study reported by Zheng in [226]

The purpose of this study was to "assess the performance of radiologists in the detection of masses and microcalcification clusters on digitized mammograms in a CAD environment after modulating cuing sensitivity levels and false-positive rates." [226] This goes beyond the Krupinski study outlined above by looking for an effect on performance as a function of the quality of the CAD system.

Seven board-certified radiologists with a minimum of 3 years of experience in mammography interpretation were the participants in this study. All tumor-positive images in the study were verified by biopsy. Original film mammograms were digitized using 12-bits of resolution and these digital images were displayed.

The study wanted subtle and difficult cases and used a multi-step selection process. First, 200 positive cases were selected for which the CAD output reported a low probability of target present. A set of 80 suspicious negative cases which had a high CAD probability were also selected. This initial set of 280 images was pruned by two experienced radiologists using the same viewing hardware as would be used in the study. This resulted in a final set of 120 cases covering a range of abnormalities (masses and microcalcifications). The final breakdown was 85 images depicting either masses or microcalcifications or both and 35 negative cases with no abnormalities.

Each radiologist viewed the 120 cases five times, once for each of the five viewing modes. The blocking for the viewing is described below. The five viewing modes consist of a noncued mode (no markings), and each combination of a true-positive cuing sensitivity of 90% and 50% with false-positive rate (per image) of 0.5 and 2.0. Specifically, the viewing modes are given in Figure 2.2. During the study, radiologists loaded the cued images into the display and after that were free to pan and zoom but the window leveling was fixed and not able to be altered.

The CAD algorithms used to mark the study images were developed by the authors and fall under the category of neural network-based approaches using custom, hand-derived features. After applying the CAD algorithm to the selected images a random set of cuing locations was selected until the desired sensitivity was reached. This resulted in 51 out of 57 abnormalities cued for masses and 34 out of 38 for microcalcification clusters (for the 90% sensitivity case, a similar approach

Reading Mode	CAD Cuing	Cuing Sensitivity	Cuing False-Positive Rate
1	No	n/a	n/a
2	Yes	0.9	0.5
3	Yes	0.9	2.0
4	Yes	0.5	0.5
5	Yes	0.5	2.0

Figure 2.2: Image viewing modes used by Zheng [226].

selects the cues for the 50% case). The selected images had a false-positive rate of 0.5 and were used as-is for that condition. The false-positive cuing rate was determined from the neural network output probabilities by selecting regions from the output.

Viewers participated in 20 reading sessions of 30 randomly selected cases using one of the five viewing modes. The 20 sessions were divided into four blocks of five sessions each covering each viewing mode. A minimum time delay of 10 days was used between two consecutive readings of the same case.

Viewers were asked to first identify suspicious areas of the image for the presence of an abnormality and to mark it as benign or malignant by scoring it on a sliding confidence-level scale $[0, 1]$. The likelihood scores were used to generate free-response ROC curves which were used to compare average across the viewers for each of the five viewing modes.

Zheng [226] provides FROC curves for the average detection of abnormalities for each viewing mode. These curves will not be reproduced here but their interpretation, despite the small data set size, is consistent. If the noncuded (nonhighlighted) mode is taken as a baseline, then Mode 2 and Mode 3, which feature a cuing sensitivity of 90%, enhance the detection of abnormalities. This result mimics that of Krupinski above. However, Mode 4 and Mode 5, with a sensitivity of 50% and false-positive rates of 0.5 and 2.0 respectively, clearly hinder detection with Mode 5 much worse than Mode 4.

The key result of interest to us here is that reproduced in Figure 2.3. This figure summarizes the number of missed abnormalities in nonhighlighted regions during CAD-cued readings. In Figure 2.3 the number in parentheses is the number of missed regions in that mode that were detected

Reader	Mode 2	Mode 3	Mode 4	Mode 5
1	5 (1)	5 (1)	13 (3)	14 (5)
2	6 (0)	8 (0)	19 (2)	21 (7)
3	5 (1)	5 (0)	11 (2)	15 (3)
4	5 (0)	6 (0)	19 (3)	25 (5)
5	6 (0)	4 (0)	10 (4)	13 (5)
6	7 (1)	7 (2)	14 (4)	20 (9)
7	6 (0)	5 (0)	15 (3)	18 (6)
Average	5.7 (0.4)	5.7 (0.4)	14.4 (3.0)	18.0 (5.7)

Figure 2.3: Number of missed abnormalities in noncued regions by viewer. The number in parentheses is the number of missed regions that were detected in Mode 1 (no highlighting). From Zheng [226].

in Mode 1 which had no CAD cuing. If the number is low, as is the case for Mode 2 and Mode 3 with a highly sensitive CAD detector, this implies that these abnormalities were also missed in Mode 1. However, when the cuing sensitivity was 50% (modes 4 and 5) the average number of missed abnormalities in noncued regions was significantly higher ($P < 0.05$). Also, approximately 30% of these regions were detected in Mode 1. The difference between a 50% sensitivity and 0.5 false-positives per image and 2.0 false-positives per image was not statistically significant in this case. The authors point out that this is likely an effect of the small sample size ($n = 7$).

The result shown in Figure 2.3 reinforces that shown above for Krupinski [111] in that the presence of hard agent-based highlighting has a negative effect, if the highlighting is insensitive, on the ability of viewers to detect *unhighlighted* regions of the images. So, while Krupinski demonstrated an effect, this study by Zheng demonstrates further that the effect is related to the quality of the hard agent-based highlighting system.

2.3 Hard Agent-Based Highlighting False-Negatives Inhibit Non-Highlighted Target Detection

A possible limitation of the Zheng study is the small sample size. In Alberdi [4] the sample size was larger ($n = 20$). This study was an extension of a larger multi-center UK study (HTA [206]) and focused exclusively on viewer responses to CAD false-negatives. Specifically, the authors were interested in estimating the probability of a reader making a wrong decision conditional on

	Correctly marked (N=10)	Incorrectly marked (N=23)	Unmarked (N=27)
Cancer (N=30)	10	11	9
Normal (N=30)	n/a	12	18

Figure 2.4: Data set composition. "Incorrectly marked" means that the CAD system placed a highlight on the image in the wrong location. "Unmarked" means that the CAD system did not place any highlights on the image when it should have done so. From Alberdi [4].

the false-negative reported by the CAD system. Hard agent-based highlighting was used to mark the location of CAD prompts as is typical in this area.

The authors point out that there are two means by which a false-negative can be produced by a CAD (or any other agent-based) system. First, no mark of any kind may be placed on the image when a mark should be. This is called by the authors an *unmarked* mammogram. The second is to place a marker on the image but away from the area of the actual abnormality. The authors call these *incorrectly marked* mammograms.

Alberdi [4] describes two experiments in this paper. For Experiment 1 twenty readers from three different UK screening centers participated. Twelve were radiologists, seven were trained radiographers and one was a breast clinician. The authors do not indicate level of experience but state that all were actively involved in breast screening.

Sixty sets of mammograms were used. Each set consisted of four images, two of each breast. The images were provided by the company that developed the CAD system used in the study. These images had known output, either proven by biopsy or determined as clearly negative. Images for the study were selected as shown in Figure 2.4 which included a disproportionate number of false negatives (a focus of the study). In particular, no normal cases unmarked by the CAD system were included in the data set. This means that any image with no markings on it was an example of an "unmarked" false negative. For the 30 cancer cases 10 were microcalcifications and the other 20 were masses.

All readers viewed all images once in one session with highlights shown in all cases. Each reader viewed the images in the data set in a unique random order. The readers viewed images as actual films on a standard viewing roller and also had a paper version of the digitized image on

	Correctly marked	Incorrectly marked	Unmarked
Cancer	81%	53%	21%
Normal	n/a	92%	94%

Figure 2.5: Percentage of correct human decisions when viewing mammograms with incorrect or unmarked CAD highlights. From Alberdi [4].

which to mark locations. For each location marked readers were to assign a decision: 1 - recall; 2 - discuss but probably recall; 3 - discuss but probably no recall; 4 - no recall. These categories are largely in line with those used by Krupinski [111] above. CAD marks were on the paper copy but not on the film displayed on the roller.

The results of Experiment 1 are shown in Figure 2.5 where the values given are the percentage of correct decisions made by the readers for each of the CAD image groupings. The percentages are from the total number of recall/no recall decisions from all 20 participants. The low overall sensitivity of 52% and especially the low value of 21% for targets completely missed by the CAD system led the authors to conduct Experiment 2.

In Experiment 2 nineteen readers who had not participated in Experiment 1 were used. Six were radiologists, seven were trained radiographers and six were breast clinicians. All had similar levels of experience to the participants of Experiment 1 (this level was not given in the paper). The data set viewed by these readers was identical to that of Experiment 1. The only difference between Experiment 1 and Experiment 2 was that readers in Experiment 2 viewed the images *without* CAD highlights. The viewing and marking process (ie, films and paper prinouts) was the same as in Experiment 1. The best six readers, as determined by the heads of their respective centers, were selected to act as judges to measure the difficulty of the cases. These judges ranked the cases and assigned labels to indicate how many average readers, in their opinion, would catch the cancers in that case. The results from Experiment 2 are given in Figure 2.6.

When comparing the results in Figure 2.5 and Figure 2.6 the authors used ANOVA to analyze the sensitivity and determined with $p < 0.001$ that there was a significant difference between the two sensitivities. In particular, the difference for the unmarked cancers between Experiment 1 and

	Correctly marked	Incorrectly marked	Unmarked
Cancer	90%	66%	46%
Normal	n/a	87%	88%

Figure 2.6: Percentage of correct human decisions when viewing mammograms from Experiment 1 (see Figure 2.5) without CAD highlights. From Alberdi [4].

Experiment 2 was very highly significant ($p < 0.000001$).

Like the previous two studies detailed in this section, this study indicates that false highlighting hurts viewers in their ability to interpret targets in images. In this case, the key information is in the incorrectly marked CAD images which clearly show that wrong hard agent-based highlights can have a detrimental effect on viewer's ability to detect targets. Of interest is the finding that the unmarked images, those where the CAD system made no marks at all when it should have, are even harder for viewers to interpret. There seem to be several factors which might account for this result. One is that the targets in this case are exceptionally difficult and neither humans nor machines would be able to detect them consistently. The selection of the judges by the authors of this study was meant to address this concern. In the end, only one or two images were deemed too hard to interpret. A second possible factor is that viewers are using the CAD system as if the "D" in "CAD" stood for "diagnosis" and not "detection". It might be that unconsciously viewers are deciding that the CAD system is actually determining malignancy (or target location) and even if it makes many false positives (or, because it makes many false positives) it must be that if it makes no marks there is no cancer (no target to find). The thinking might be "well, it the CAD system makes lots of mistakes and it didn't make any on this image so there really must be nothing to find." Clearly, this is not how the CAD system was designed but if correct, future systems must take this bias into account in some way.

One limitation of the Alberdi [4] study is that it relied on a mixed method for presenting images to the participants. The participants viewed unmarked films on a viewing roller while attempting to mentally map markings on a printout on paper to what was viewed on the roller. This is very different from the controlled digital image display of Zheng [226] and Krupinski [111].

It is unclear what sort of errors or biases this approach might have introduced into the experiments.

2.4 Hard Agent-Based Highlighting Strongly Affects Image Search Patterns

The final study we examine in detail in this section involves a simulated CAD environment of targets and distractors. In Drew [55] we are told in the introduction that the hypothetical benefit of CAD from a signal detection perspective is unrealized in part because of an interaction between the radiologist (viewer) and the CAD itself (an agent-based system using hard highlighting techniques). In order to investigate this interaction the authors performed this simulation study.

The study consists of two experiments both of which involved tracking the eye movements of observers half of whom searched the simulated images without any CAD markings and half who searched with CAD markings. The "CAD" system (artifical here, just a simulation) marked 75% of all targets and 10% of nontargets. Experiment 1 had the goal of making the targets difficult to find and simulated a pure detection exercise. This is the historic and primary use of CAD in medical imaging. Experiment 2 changed the appearance of targets and distractors (here "T" and "L") so that while it was easier to find a target or distractor it was harder to decide which was which. This was to simulate a diagnosis situation.

In particular, observers were instructed to locate the target letter ("T") among distractors ("L") all of which were embedded in $1/f^{2.4}$ cloudlike noise. This noise was selected as being consistent with spatial frequencies encountered in medical images, particularly mammography. The letters were oriented randomly and placed within one of a 4x4 grid of locations with jittering applied within the location to keep the letters from lining up in an easy to identify pattern. Each trial contained an average of five "L" distractors and at most one target ("T"). Images were shown initially with the CAD markings on. Observers viewed the images with half the observers assigned to a CAD condition and the others to a no CAD condition. Observers started with a 50 trial practice block of images without CAD markings. The total number of images viewed by all observers was 150. The gaze pattern of each observer was kept with an eye tracker.

There were 23 observers in Experiment 1 and 24 in Experiment 2. The observers were not

radiologists nor image analysis experts. They ranged in age from 18 to 54 ($\bar{x} = 24.3$). Eleven were male and none were color blind.

Experiments 1 and 2 were different in the opacity of the $1/f^{2.4}$ noise and the similarity between targets and distractors. Experiment 1 had high noise and low similarity between targets and distractors while Experiment 2 had lower noise and higher similarity. Experiment 1 made it harder to find targets but easier to identify them when found. Experiment 2 made if easier to find a target (or distractor) but harder to identify once found.

Experiment 1 compared two groups, those using CAD and those that did not (for the same images). Like the other studies in this section, Drew [55] found an increase in sensitivity when using CAD (80% to 87%, $p < 0.001$). And, again similar to other work in this chapter, the authors found that the highest increase in sensitivity was for targets marked by CAD (81% no-CAD to 97% with CAD, $p < 0.001$). They also found that unmarked targets in the CAD condition had a much lower sensitivity than the same targets in the no CAD condition ($p < 0.001$). This mirrors the findings of Krupinski [111], Zheng [226], and Alberdi [4] above. Experiment 2, with lower image noise and harder to discern targets, did not, in the case of unmarked CAD targets versus no CAD targets, lead to a statistically significant difference between the two groups ($p > 0.5$). As the authors point out, "CAD seems to have changed the way observers spend their time; not the amount of time that they spend".

Because the experiments captured eye movements it is possible to see where observers were looking in the images and for how long. The authors include figures with heat maps showing (averaged over all observers) how much time was spent in different conditions. One of these is reproduced in Figure 2.7 where it is plain to see that the observers spent a lot of time fixated on the incorrectly marked hard highlight (circle) while the observers without CAD assistance searched and located the target.

In Figure 2.8 we see an example of an image with no target present. Here the difference between the unmarked image and the incorrectly marked image (in this case with two CAD false positives) is equally dramatic. For the unmarked image it is clear that observers are searching for

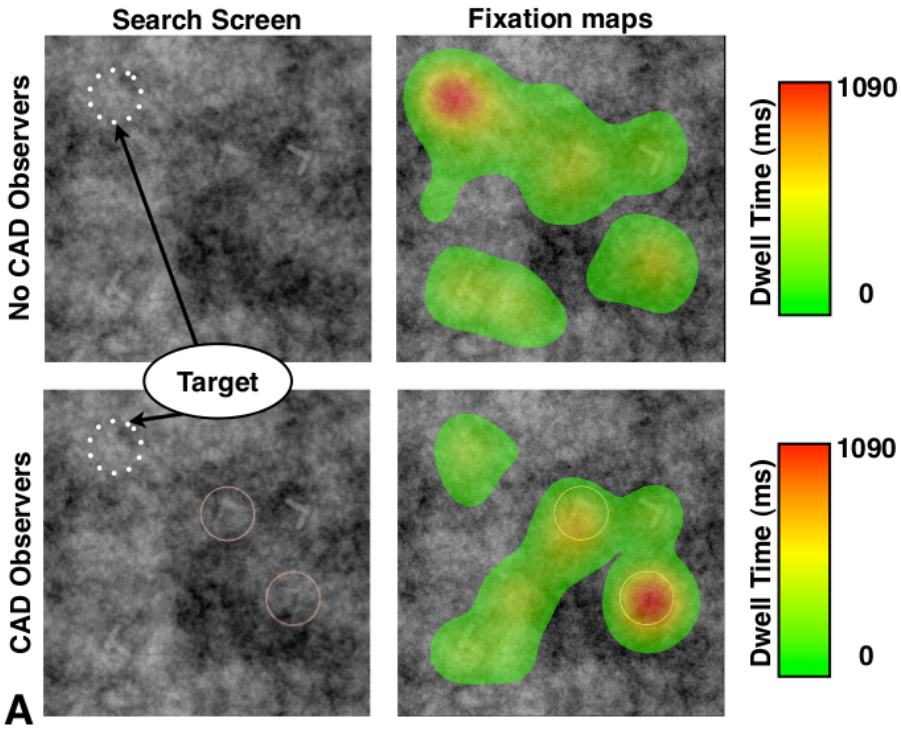


Figure 2.7: Eye gaze duration averaged over all observers for a sample image with an incorrectly marked target location. Observers with the hard highlighting from the erroneous CAD marks spent most of their time looking at the false positive and very little searching as those without highlighting did. From Drew [55], figure 3.

the target while in the CAD case observers are fixed on the false positives instead and do not even search the full extent of the image.

Drew [55] found, like other studies, that when hard agent-base highlighting is used for the purposes of target detection false positives are quite harmful because they lead to the missing of unhighlighted targets. Experiment 2 was meant to simulate a diagnostic situation where locating the target was not as difficult but determining its type was harder. In this case, the influence of the false positives disappeared quite possibly because it was so much easier to locate the target that the observer's gaze was not drawn to becoming fixed on it. Ie, the task was simply too easy for the observer to care what the CAD system did or did not find.

A limitation of this study, besides the fact it used completely simulated data, is that the observers are naive. They suggest further work to see if the tendency to search less in the presence of false positive markings would persist with experts. The previous studies in this section suggest that it might.

2.5 Summary of the Weaknesses and Pitfalls of Hard Agent-Based Image Highlighting

The studies detailed in this section clearly demonstrate that hard agent-based highlighting can come with a cost. Krupinski [111] showed that hard agent-based highlighting, even with direct precuing of where to look in the x-ray image, inhibits the detection of non-highlighted targets in the image. Zheng [226] showed that this effect persists when the viewer is allowed to freely search the image for targets. Additionally, [226] also showed that the quality of the agent has a strong impact on how many non-highlighted targets are missed. Alberdi [4] examined the effect of the two kinds of false-negatives that an agent-based highlighting system can make and demonstrated that these effects lead to erroneous judgements. Lastly, Drew [55] found results similar to the others reviewed in this chapter but also demonstrated the strong influence hard agent-based highlights have on viewer's search patterns.

The effects of hard agent-based highlighting are clear but the causes less so. They could

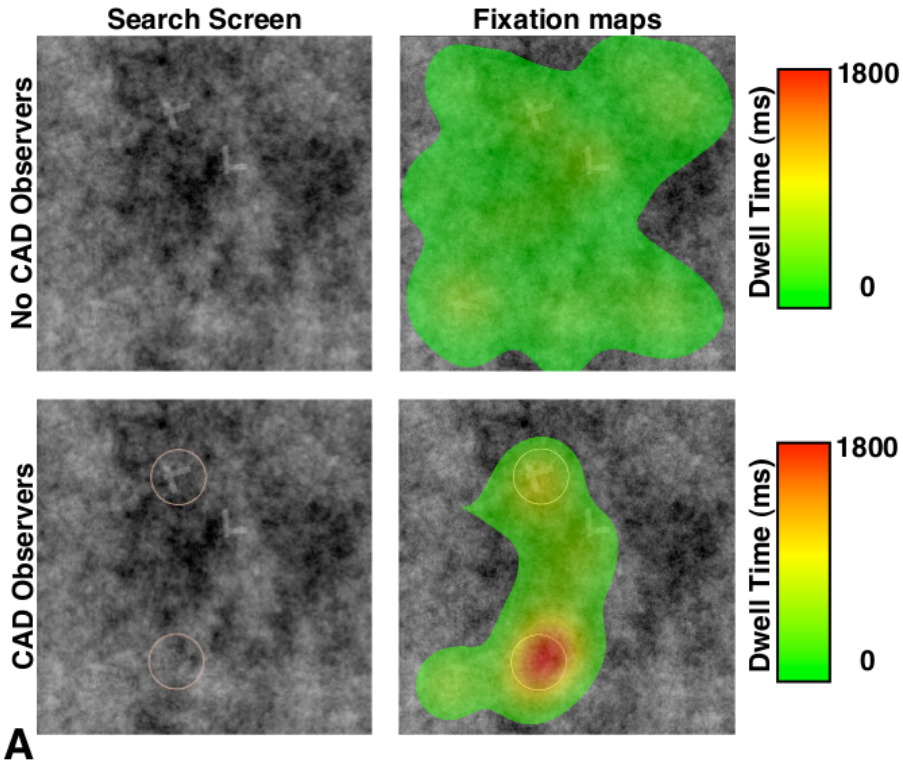


Figure 2.8: Eye gaze duration averaged over all observers for a sample image with two false positives and no target present. From Drew [55], figure 4.

be based on assumptions made by users of these systems as to the quality and reliability of the system. They could also be more primal and due to the way in which the highlights themselves are implemented. All of these studies used a simple circle as a marker for where the human observer should look. Could there be a component of these effects that is due to the way in which the human visual system, a product of over 500 million years of evolution, has primed itself to work in a world full of threats and dangers? If so, then it would be sensible to experiment with the manner in which the highlights themselves are implemented to see if any of these issues persist, diminish, or if new ones appear.

Chapter 3

Experiments with Synthetic Imagery

In this chapter, we explore the effect of soft and hard highlighting in a visual search task. Stimulus displays consist of an array of handprinted digits. Participants are asked to search for and click on all instances of the digit "2". Highlighting is performed by modulating the grey level of the pixels of a display element (one of the handprinted digits). With a light background, the dark elements are salient, and light elements fade into the background as illustrated in Figure 3.1.

Figure 3.1: A grid of display elements of the sort used in the experiments.

The highlights are determined by a classifier which assigns to each display element a probability of being a target. This probability, in $[0, 1]$, determines the grey level shading of the element:

elements with low probability are shaded light, elements with high probability are shaded dark.

To control for the quality of the classifier, instead of training an actual classifier, we generate a series of *stochastic oracle-based classifiers* that systematically vary in quality. High quality classifiers reliably discriminate targets from nontargets; low quality classifiers do not.

Our stochastic oracle-based classifier determines an output level for every display element in $[0, 1]$, conditioned on whether or not the element is a target. (This procedure assumes an oracle that knows if each element is a target; hence the term "oracle-based".) We specify one probability distribution over output levels for targets, and another distribution for nontargets. The less overlap there is between these distributions, the higher is the quality of the stochastic classifier. We use a symmetric pair of beta densities to model these distributions, with $\text{Beta}(\theta, 1)$ to draw output levels for targets and $\text{Beta}(1, \theta)$ to draw output levels for nontargets.

Figure 3.2 shows a plot of several pairs of beta distributions of the above form for different θ values. This shows the complementarity of the two distributions for a single θ value which can be thought of as a proxy for the quality of the classifier.

Although the quality of a classifier can be specified by θ, there are equivalent, more intuitive measures than can be used. A common intuitive measure is the *equal error rate* (*EER*). The equal error rate is the point on the ROC curve where the false positive rate is equal to the false negative rate, or to use a different terminology, when sensitivity equals specificity. See [6] for definitions of sensitivity and specificity. The better the classifier is, the smaller this value will be. A classifier that simply guesses target or nontarget will give an equal error rate of 0.5.

For those familiar with signal-detection theory [78], another intuitive measure of classifier quality is the dimensionless quantity known as as d'. Larger d' indicate that the classifier is more effective at discriminating targets from nontargets. To calculate d' for our paired beta distributions we need to calculate the normalized incomplete beta function to determine the cummulative probability density below a threshold of 0.5 and that above 0.5. The density above 0.5 are true positives (hits) while the density below 0.5 represents false positives for any given θ. If we define,

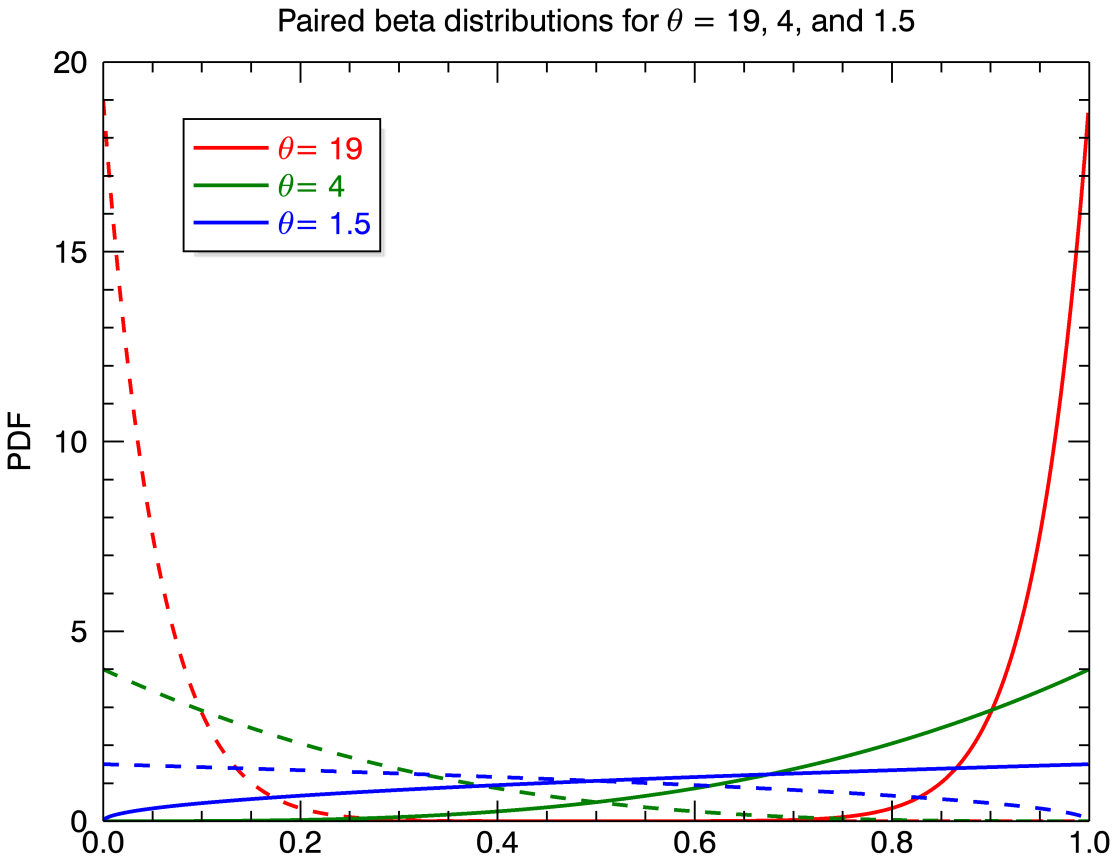


Figure 3.2: A plot of pairs of beta distributions for different θ values. For each θ two distributions are plotted: beta$(\theta, 1)$ for targets (solid) and beta$(1, \theta)$ for nontarget digits (dashed). For large $\theta = 19$ (red) the target draws will almost always be near 1 while nontarget draws will be close to 0. This simulates a highly confident classifier. As θ decreases to 4 (green) and then 1.5 (blue) the pair of curves become flatter meaning that the two probability distributions are becoming increasingly similar. At $\theta = 1$ (not shown) the two curves overlap completely and the distribution collapses to a uniform distribution. This simulates a classifier that simply guesses whether the digit is a target or not.

θ	EER	d'
1	0.5	0.0
1.5	0.3559	0.75
2.333	0.2054	1.69
4	0.0646	3.07
9	0.0021	5.77
19	≈ 0.0	9.24
1.737	0.30	1.00
1.907	0.27	1.25

Table 3.1: The relationship between θ, equal error rate (EER), and d' for values used in the experiments described in this chapter. In some cases, a specific d' value was desired which led to the very specific θ values in the table.

$$B_{<0.5} \equiv \frac{\int_0^{0.5} t^{\theta-1}(1-t)^{1-1}dt}{\int_0^1 t^{\theta-1}(1-t)^{1-1}dt} \tag{3.1}$$

$$B_{>0.5} \equiv 1 - B_{<0.5} \tag{3.2}$$

for $\alpha = \theta$ and $\beta = 1$ then d' is calculated from these densities using the cumulative Gaussian distribution,

$$d' = Z(B_{>0.5}) - Z(B_{<0.5})$$

where $B_{>0.5}$ is the cumulative beta distribution density above 0.5 and $B_{<0.5}$ is the density below 0.5.

In the experiments that follow, we constructed stochastic classifiers with a range of θ values. Larger θ values simulate classifiers with better performance, lower EERs, and higher d' values. Table 3.1 shows the relationship between θ, the equal error rate (EER), and d' for the stochastic classifiers used in the experiments reported in this chapter. Figure 3.3 shows sample displays for a range of θ, along with the beta densities that define the corresponding stochastic classifier.

In terms of the taxonomy of highlighting techniques presented in Chapter 2, the scheme described in this chapter is considered soft and agent-based. It is soft because the continuous

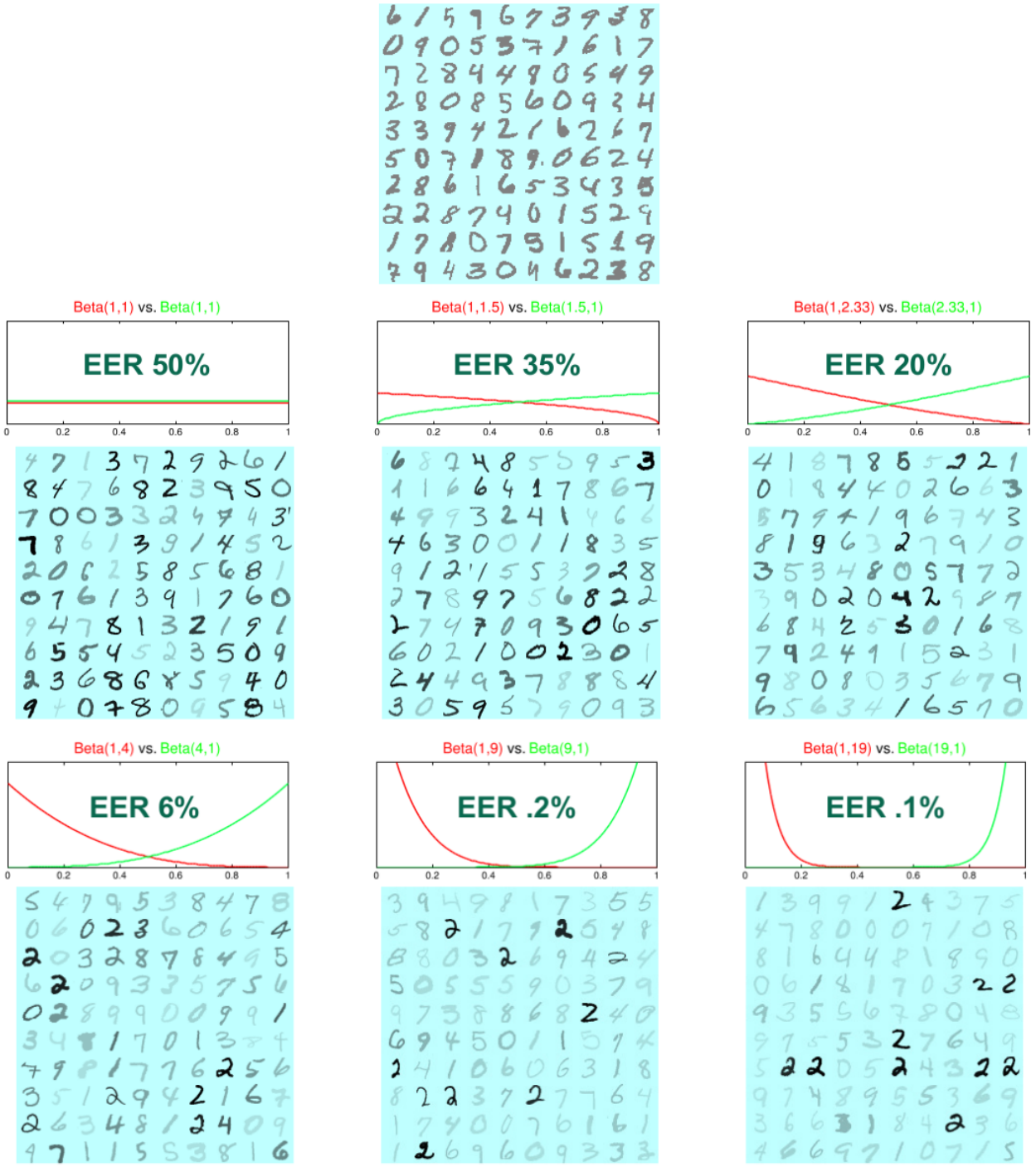


Figure 3.3: Examples of the type of digit grids used in the experiments. The grid in the top center is the control condition with all digits displayed with intensity 0.5. The remaining six grids, from middle row to bottom, left to right, are for $\theta = 1.0, 1.5, 2.333, 4.0, 9.0, 19.0$. See Table 3.1 for the relationship between θ and other measures of classification strength like EER and d'. The main item to note is that larger θ leads to increased contrast between targets and the background along with decreased contrast between nontargets and the background. The paired Beta distributions associated with each display are also shown along with the EER. The smaller the EER the more likely the stochastic classifier is correct.

classifier output is used to determine the display element intensities, and it is based on an agent, the stochastic classifier that we have simulated.

3.1 Experiment 1: Time to Locate a Fixed Number of Targets

In Experiment 1, participants searched for a fixed number of target digits (the digit "2"). We examine the time taken to find the targets based on the nature of display highlighting. We compare a control condition with no highlighting (see top array in Figure 3.3 to soft-highlighting conditions with varying degrees of classifier quality.

3.1.1 Methods

Subjects viewed grids of handwritten digits in which the grey-level intensity of each digit was drawn from beta probability distribution of the form beta$(\theta, 1)$ for targets or beta$(1, \theta)$ for nontargets. Experimental conditions consisted of $\theta = \{1.0, 1.5, 2.3333, 4.0, 9.0, 19.0\}$, corresponding to d' of $\{0.0, 0.75, 1.69, 3.07, 5.77, 9.24\}$. These values simulate a range of classifiers from those that guess by flipping a coin ($\theta = 1$) to a very competent classifier ($\theta = 19$). In addition to these six conditions we included a no-highlighting control condition in which each digit was rendered with a constant intensity of 0.5.

On each trial, the subject was presented a grid of digits which contained, in random arrangement, ten of each digit (0-9), for 100 digits total. The subject was asked to click on each of the ten instances of the digit "2" which would then be replaced by the background color. The trial would end when all ten targets were found or after 15 seconds had elapsed.

The digits themselves, taken from the MNIST data set [115], were drawn in 28x28 pixel blocks in a 10x10 grid. The background color was (red, green, blue) = $(0.79, 1, 1)$. The digit intensity was an index into a 256 element Lab grey scale ramp where the index was $\lfloor (255)(1 - b) \rfloor$ with,

$$b \sim \begin{cases} \text{Beta}(1, \theta), \text{if nontarget} \\ \text{Beta}(\theta, 1), \text{if target} \end{cases} \tag{3.3}$$

The Lab grey scale ramp is generated by fixing $a = 0$, $b = 0$ and varying L from 0 to 100 in even increments.

All subjects were run on Amazon's Mechanical Turk inside of a web browser on the subject's machine. The browser component was implemented in JavaScript using HTML5 canvas for drawing, the server-side component was implemented in Python (CGI), and the actual trial data was generated using IDL (*Interactive Data Language*, Exelis Visual Information Solutions, Boulder, CO). All trials for a particular subject were generated on the fly and passed down to the subject's browser in the JavaScript source code. This included the data to plot each pixel of each digit of each trial. This enabled embedding all experiment data in the single page running on the browser and eliminated any communication with the web server until the experiment was completed. This removed any delays due to network latency.

For each experiment the subject was presented with a consent form (IRB approved) and a browser check was performed to make sure the subject's browser supported HTML5 canvas. For each subject all data pertaining to digit tokens used, their selected intensity, and arrangement on the screen were stored to be able to reconstruct the exact set of trials presented, if necessary.

All subjects were adults who were not color blind. While the IP address of the subject's web browser was recorded in case approximate geographic location was deemed helpful in understanding the results of the experiments this information was not used in any analyses.

Out of 35 subjects who started the experiment 25 completed all trials. The other 10 either quit the experiment on their own or were rejected. The experiment was unable to differentiate between subjects who quit and who were rejected. Subjects were only allowed to participate once so all subjects were unique.

Each subject was presented with 42 trials in six blocks of seven conditions (six θ values and one control of medium intensity for all digits). Trials within a block were generated at random with each digit token used only once in the experiment. The ordering of conditions within a block was also randomized.

All conditions in a block were presented before the next block started with no break or other

indication to the subject that the next block was starting. If the subject failed to find all "2" digits within 15 seconds they were told how many they found and moved to the next trial when they clicked the "Next" button. The "Next" button was located in the middle of the screen and therefore ensured that the mouse pointer was in a known position at the start of each trial. Timing started immediately after the trial digits were displayed. If subjects failed to find at least two "2" digits or clicked on more than six non-"2" digits in a trial they were rejected and the experiment ended.

For each trial, each digit token clicked and the time of the click were stored. When a target digit token was clicked it was removed from the display and replaced with the background color.

After the 42 trials were complete the recorded per trial information was sent back to the server for analysis and the subject was told that the experiment was finished and that they would be paid. If the experiment was terminated early because the subject clicked too many nontarget digits or found too few targets during a trial the subject was told that they would not be paid and their results were rejected. It was felt necessary to add the possibility of rejection to prevent subjects from simply clicking randomly and to keep them focused on the task.

3.1.2 Results

Figure 3.4 shows the percentage of the ten targets detected as a function of time for the control condition (black) and the six experimental conditions (shaded from blue to purple in increasing order of classifier quality). For each time point, the mean number of targets found by that time for each subject in each condition was calculated and the mean across subjects was plotted. A curve that rises quickly and then asymptotes at 100% is indicative of an easy search; a curve that rises slowly and doesn't reach 100% by the end of the 15-second trial is indicative of a difficult search. Error bars in the Figure are corrected for between-subject variability according to the procedure in [138].

A qualitative examination of the curves reveals that search with soft highlighting is more efficient than for the control condition except for the $d' = 0$ classifier, i.e., the classifier that

generates highlights by drawing from a uniform distribution and has no ability to discriminate targets from nontargets. The curves do not cross over one another, except perhaps the control condition and the $d' = 0$ classifier, indicating that the superiority of one method of highlighting over another is consistent across time. One might have expected a different result, e.g., one method of highlighting leads to an early advantage but a late cost. The interpretation of our results is simplified by the consistency of the ordering of conditions over time.

The surprising result of this experiment is that even a weak classifier, i.e., a classifier with $d' = 0.75$, produces highlights that support the subjects in visual search. It was non-obvious from the outset that we would find such a benefit of a weak classifier, given that the weak classifier often highlights nontargets and fails to highlight targets.

3.2 Experiment 2: Searching For a Single Target in Variable Sized Displays

Experiment 1 was atypical of psychological studies of visual search in two respects. First, targets were present on every trial. Second, every display contained 10 targets. The canonical visual search study involves searching for a single target, and often the number of display elements varies in order to determine a search *slope*—the increase in response time for each additional element in a display. For Experiment 2, we performed a more traditional study involving a single target with two display sizes. The response latencies for the two display sizes in a given condition can be recast in terms of a search slope and a search intercept. In Experiment 2, every display contained exactly one target. Subjects were given up to 45 seconds per trial to locate the target.

The speed up due to highlighting that we observed in Experiment 1 could have one of two effects in Experiment 2: it could decrease search slopes or search intercepts. The slope reflects additional time to process each element of the display. The intercept reflects fixed preprocessing time or fixed motor preparation time. If the quality of highlights affects the search slope, highlighting makes it easier for subjects to reject elements in the display (the digits), which also indicates a guidance of attention. We hypothesized that highlighting would have this affect on response latencies.

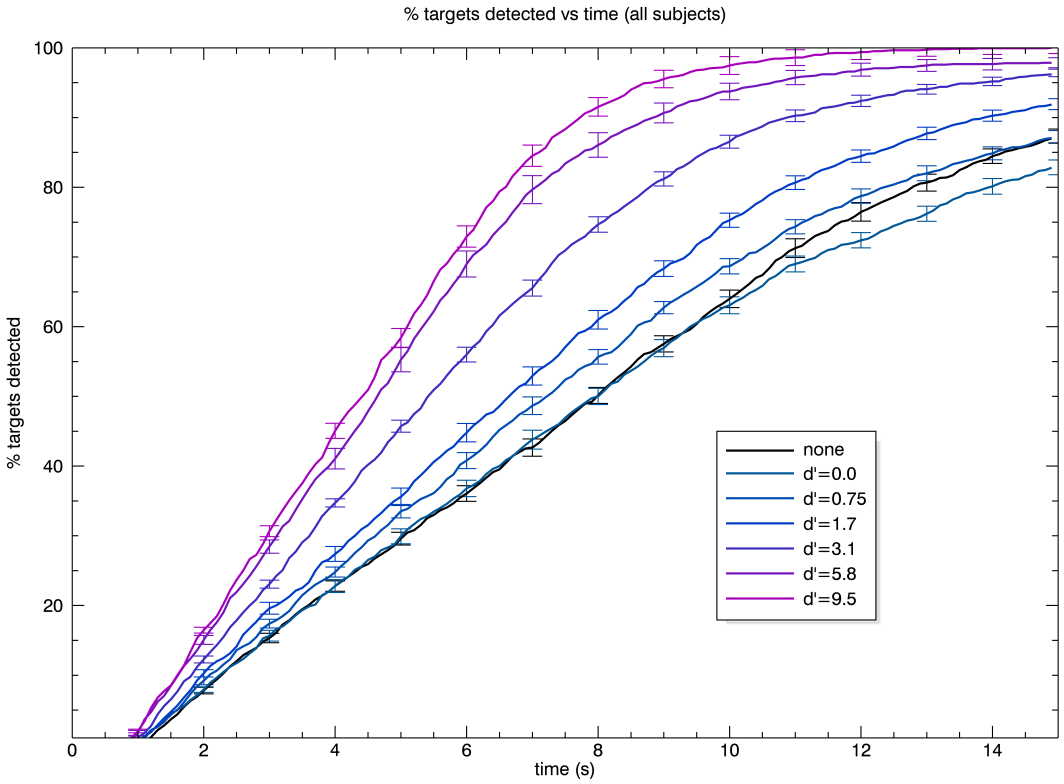


Figure 3.4: Experiment 1. Percentage of the ten targets ("2" digits) detected as a function of time and condition. Each target was removed as it was clicked. The mean over all subjects after between subject variability correction is shown (± SE of the mean). The data are corrected for between-subject variability according to the procedure in [138].

3.2.1 Methods

The stimulus arrays in Experiment 2 were like those in Experiment 1, except that two display sizes were studied: 10×10 and 7×7. In addition to a no-highlighting control condition, we simulated stochastic oracle-based classifiers with $\theta = 1.737$ and $\theta = 3.322$, corresponding to $d' = 1.05$ and $d' = 2.56$. Stimulus arrays were generated as in Experiment 1 except that each display contained exactly one target ("2"). The remaining 99 or 48 elements were randomly selected from set of nontarget digits with no repetition of digit tokens throughout the experiment.

Each subject was presented with 90 trials consisting of 15 blocks each containing exactly one trial in each of six conditions (three highlighting conditions crossed with two grid sizes). Conditions within a block were randomized. There was no indication to the subject that one block of displays had ended and another was beginning. All conditions in each block were presented before the next block. If the subject failed to find the "2" in 45 seconds the experiment ended and the subject was rejected. The subject was also rejected if more than six nontarget digit tokens were clicked.

When subjects identified the target, they were shown their response latency and a 'next' button lit up which would initiate the following trial. Each trial time started with the display onset. For each trial the location of the "Next" button below the latency message implicitly reset the mouse pointer to a common screen position relative to the grid of digits which itself was always drawn in the center of the screen. Additionally, each trial began by presenting a red fixation cross in the center of the screen for one second before presenting the display of digit tokens.

For each trial, each digit token clicked and the time of the click were stored. When the digit token clicked was the target "2" the trial ended. Nontarget digit tokens clicked were recorded but no feedback was given to the subject.

After the 90 trials were complete the recorded per trial information was sent back to the server for analysis and the subject was told that the experiment was finished and that they would be paid. If the experiment was terminated early because the subject clicked too many nontarget digits or failed to locate the target in 45 s on any trial the subject was told that they would not be

paid and their results were rejected. As in Experiment 1, it was felt necessary to add the possibility of rejection to prevent subjects from simply clicking randomly and to keep them focused on the task.

Out of 56 subjects who started the experiment 37 completed all 90 trials successfully while 19 subjects did not and were removed from the analysis. These subjects either quit the experiment on their own without finishing or were rejected. The experiment was unable to differentiate between subjects who quit and who were rejected. As in Experiment 1, subjects were only allowed to participate once, were adults who were not color blind and accepted the consent form required by the University in order to participate in the experiment.

3.2.2 Results

Figure 3.5 shows the mean reaction time for each of the three highlighting conditions (control, $d' = 1.05$, and $d' = 2.56$) and the two display sizes (7×7 and 10×10).

An ANOVA was used to evaluate the results indicating that the three highlight conditions (control, $d' = 1.05$, $d' = 2.56$) yield reliably different response latencies ($F(2, 72) = 14.4, p < 0.001$). The subject's median latency was used as the dependent variable. The two display sizes (7x7, 10x10) yield reliably different response latencies ($F(1, 36) = 127.6, p < 0.001$) as well.

Additionally, there is a reliable highlight condition by size interaction ($F(2, 72) = 5.18, p = 0.008$). The search slope in the control condition appears steeper than in the $d' = 1.05$ condition, which in turn is steeper than in the $d' = 2.56$ condition. Thus, highlighting reduces the time to search for each item in the display. This finding indicates that the benefit of highlighting is, as one would expect, to guide attention to relevant locations in a display.

The slope of the best fit line to each condition is shown in Figure 3.5 along with the results of paired t-tests and two-sided Wilcoxon signed-rank tests [218] showing that the two highlight conditions lead to a significant reduction in the slope when compared to the control condition. This indicates that highlighting, in particular soft highlighting, is enabling subjects to locate targets more quickly as the display size changes. The nonparametric test was included as an additional

indication of the strength of any differences deemed significant.

The best fit lines for Figure 3.5 are,

$$\mathrm{RT(control)} = 29.6\ x + 1377.6 \tag{3.4}$$

$$\mathrm{RT(d' = 1.05)} = 20.6\ x + 1525.9 \tag{3.5}$$

$$\mathrm{RT(d' = 2.56)} = 16.2\ x + 1513.9 \tag{3.6}$$

$$\tag{3.7}$$

where RT is the estimated reaction time for display size x. As indicated in Figure 3.5 there is a statistically significant difference in slopes between the highlighted and control conditions. This also applies to the intercepts.

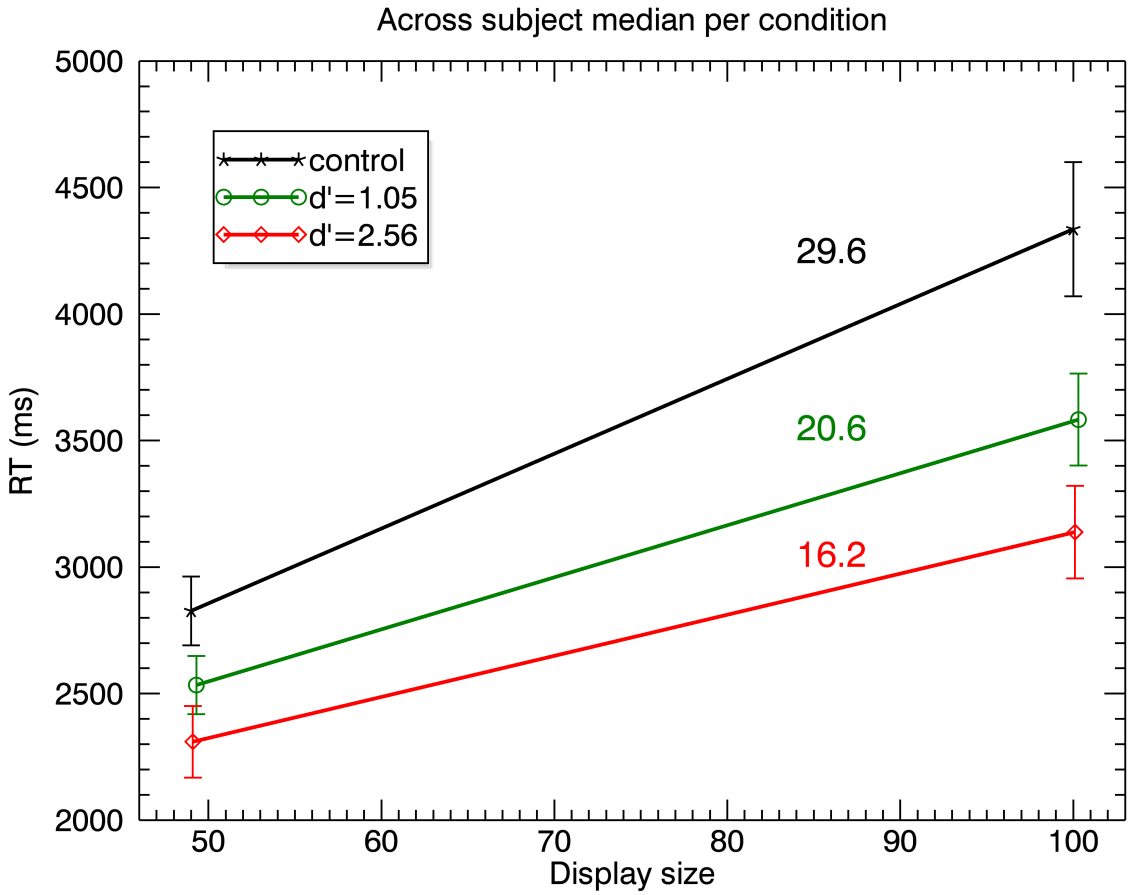
The intercepts above are usually thought of as a fixed time taken to do low-level perceptual processing of the display and motor preparation for making a response. The statistically significant difference in the intercepts between the control and two highlighted conditions indicates that the highlighted displays require a bit more overhead to parse because of their non-uniformity, specifically, that the faint features take longer to extract because the intensity differences create textures and contours that are distracting insofar as the search task is concerned.

3.3 Experiment 3: Comparing Soft Versus Hard Highlighting

Experiments 1 and 2 provide evidence that soft highlighting is superior to no highlighting as long as the highlights are based on a classifier that has some discriminative ability. In Experiment 3, we turn to the more subtle comparison of the effects of soft versus hard highlighting on search efficiency.

3.3.1 Methods

Experiment 3 was set up similar to Experiment 1, with 10×10 arrays and 10 targets. We compared three classifier qualities, corresponding to $\theta = \{1.514, 1.737, 1.907\}$ (or $d' = \{0.75, 1.0, 1.25\}$)




```
control - d'=1.05: 29.579 vs 20.568, t(36)=1.890 (p=0.0668), w(36)=1550.0 (p=0.0790)
d'=1.05 - d'=2.56: 20.568 vs 16.243, t(36)=1.184 (p=0.2441), w(36)=1492.5 (p=0.2563)
control - d'=2.56: 29.579 vs 16.243, t(36)=3.178 (p=0.0030), w(36)=1652.5 (p=0.0042)
```

Figure 3.5: Experiment 2. The $\bar{x} \pm$ SE of the median reaction time across all subjects for each grid size and condition. Values at $x = 49$ are for the 7x7 grid while values at $x = 100$ are for the 10x10 grid. The slope of the lines is marked. Paired t-test and Wilcoxon signed-rank tests show that soft highlighting leads to faster target localization as display size increases. The intercepts for the best fit lines are control: 1377.6, d'=1.05: 1525.9, and d'=2.56: 1513.9 with identical statistical test results to the slopes.

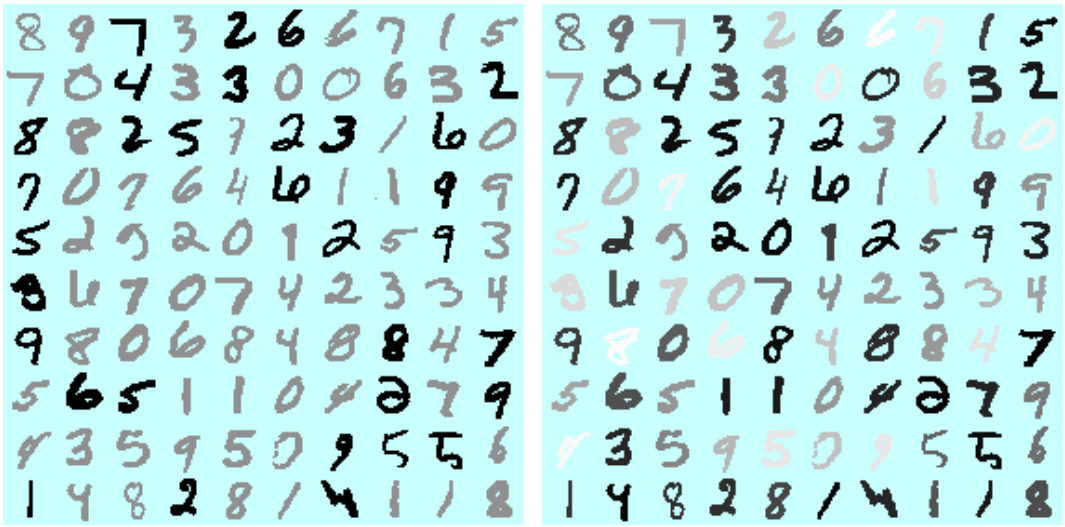


Figure 3.6: Sample digit arrays for Experiment 3. (Left) hard highlighting using binary intensities. (Right) soft highlighting in which intensity varies according to confidence of the stochastic classifier ($\theta = 1.737$, $d' = 1.05$).

with either soft or hard highlighting. As in Experiments 1 and 2, soft highlighting manipulated the intensity of a display element proportional to the simulated classifier output. Hard highlighting thresholded the classifier output, which ranged from 0 to 1, at 0.5. Any output below the threshold was set to a display intensity corresponding to a classifier output of 0.2 for soft highlighting, and any output above the threshold was set to a display intensity corresponding to a classifier output of 1.0. Figure 3.6 shows an example of hard versus soft highlighting for $\theta = 1.737$ ($d' = 1.05$). Experiment 3 included a no-highlighting control condition for a total of 7 conditions.

Subjects were given 15 seconds in which to locate all ten "2" digits randomly distributed on a 10x10 grid. There were exactly ten examples of each digit token (0-9). Digits tokens were used only once per experiment.

Experiment 3 was run in a manner identical to Experiment 1 with the same block size, randomization, and presentation to subjects. See Section 3.1.1 for details.

Out of 54 subjects who started the experiment 41 completed all trials. The remaining 13 either quit the experiment on their own or were rejected. The experiment was unable to differentiate between subjects who quit and who were rejected.

3.3.2 Results

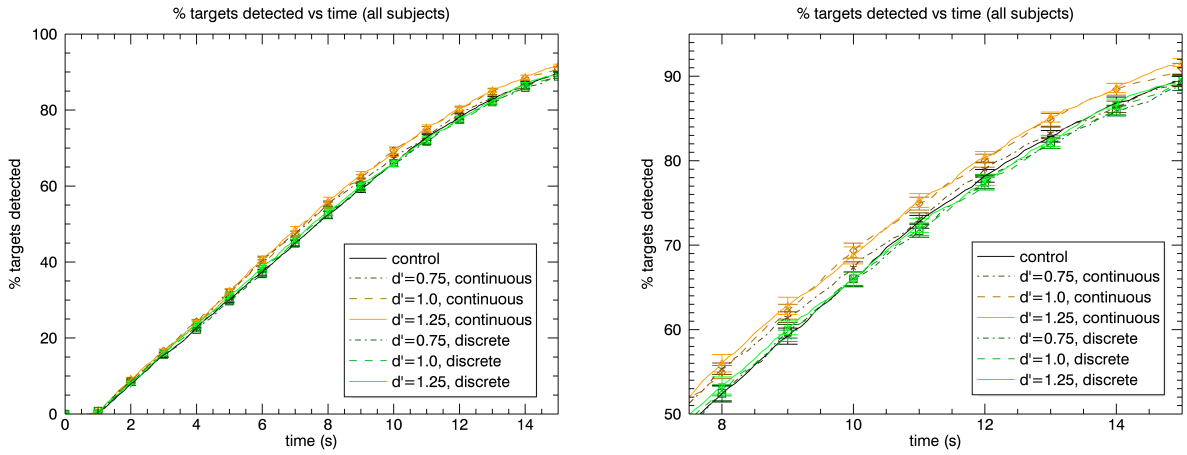
Figure 3.7 shows the percentage of targets detected by time for all seven conditions. The full range plot is on the left while the right has zoomed in to show differences between the conditions. From the figure, especially in the zoomed region, it is clear that there is a small but consistent increase in the fraction of targets found for the soft highlighting conditions compared to hard highlighting.

Figure 3.7, bottom, shows the results of paired t-tests and two-sided Wilcoxon signed-rank tests between the mean number of targets found per trial for the soft and hard highlighting conditions. The $d' = 1.25$ case is statistically significant indicating that even a moderately performant classifier allows subjects to locate more targets per trial when soft highlighting is present.

Figure 3.8 shows the difference between the mean number of targets detected in each of the six experimental conditions and the number detected in the control condition, as a function of time within a trial. In this plot, values above zero mean that more targets were located by that time than in the control condition while values less than zero mean fewer targets were located by that time than in the control condition. Note that a three second wide smoothing window was applied.

Soft highlighting is better than hard highlighting for all conditions since each soft highlighting curve is above the hard highlighting curves. Also, hard highlighting is slightly better than the control condition but not consistently so as later times drop below zero.

We divided the first 12 seconds of each trial into 4 bins of 3 seconds. We computed the number of targets found within each bin and performed a three way ANOVA with subject as the random factor and highlight quality ($d' = 0.75, 1.00, 1.25$), highlighting condition (soft, hard), and time window (0-3, 3-6, 6-9, 9-12) as three within-subject independent variables. We observe a main effect for highlighting condition ($F(1, 40) = 10.213$, $p = 0.003$), with soft highlighting superior to hard highlighting. We also observe a main effect of time ($F(3, 120) = 933$, $p < 0.001$), which simply reflects the fact that more targets are found around 6 or 9 sec than around 3 or 12 sec. We do not observe a main effect of highlight quality ($F(2, 80) < 1$), which is not terribly surprising given the




```
soft-hard (d'=0.75): 8.943 vs 8.887, t(40)=0.608 (p=0.546), w(40)=1738.5 (p=0.730)
soft-hard (d'=1.00): 9.093 vs 8.947, t(40)=1.619 (p=0.113), w(40)=1804.0 (p=0.340)
soft-hard (d'=1.25): 9.163 vs 8.951, t(40)=2.907 (p=0.006), w(40)=1819.0 (p=0.274)
```

Figure 3.7: Experiment 3. Percent of targets detected by time and condition. Left: full range plot. Right: zoomed to show differences between conditions. From the plot it is clear that there is a small but consistent increase in the fraction of targets found over time for the soft highlighting case compared to hard highlighting. Below, the mean number of targets found per trial by condition with paired t-test and two-sided Wilcoxon signed-rank test results. There is an effect between soft and hard highlighting for the d'=1.25 case where subjects were finding more targets per trial, on average.

small range of d' tested. No interactions involving these three factors are significant at the 0.05 level.

The effect size of soft versus hard highlighting was assessed at 6 and 9 seconds for each of the 3 highlight quality conditions. The effect sizes range from small (Cohen's d = 0.22 for $d' = 0.75$ at 9 sec) to medium (Cohen's d = 0.53 for $d' = 0.75$ at 6 sec). Here are all the Cohen's d values for soft versus hard,

6 sec	$d' = 0.75$	Cohen's $d = 0.528$
6 sec	$d' = 1.00$	Cohen's $d = 0.243$
6 sec	$d' = 1.25$	Cohen's $d = 0.248$
9 sec	$d' = 0.75$	Cohen's $d = 0.219$
9 sec	$d' = 1.00$	Cohen's $d = 0.324$
9 sec	$d' = 1.25$	Cohen's $d = 0.265$

Figure 3.9 shows the difference between the mean number of detected targets in corresponding soft versus hard highlighting conditions, as a function of time. Values of the curve greater than zero indicate that at a particular instant of time, subjects found more targets with soft versus hard highlighting.

The plot shows the difference, soft - hard, at each time point for the number of targets detected by that time. In all cases the difference, including error bars, is above zero meaning that subjects consistently found more targets when using soft highlighting. The error bars were calculated using the between subject variability correction of [138]. Even the weakest classifier ($d' = 0.75$) shows this effect.

For a classifier of a given quality, our results indicate that soft highlighting, which leverages the graded output of the classifier, supports human visual search better than hard highlighting, which thresholds the classifier output.

3.4 Experiment 4: Variable Number of Targets

Experiments 1-3 show that even when a weak classifier is used to provide soft highlights, the classifier can boost human performance on detecting targets that are present in the display. However, because Experiments 1 and 3 each contained exactly 10 targets per display and subjects were

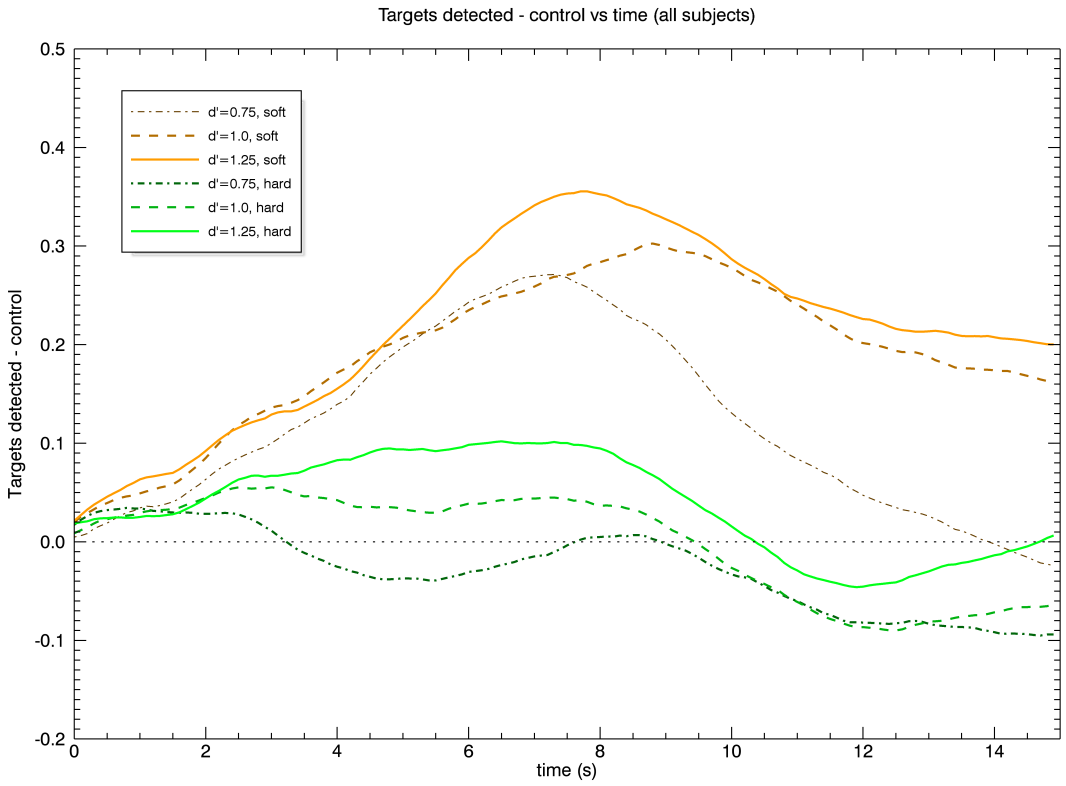


Figure 3.8: Experiment 3. The mean number of targets located across all subjects for each condition minus the mean number of targets located across all subjects for the control condition of no highlighting as a function of time. Values above zero mean that more targets were detected relative to the control condition while values less than zero mean fewer targets were detected. A three second smoothing window was applied.

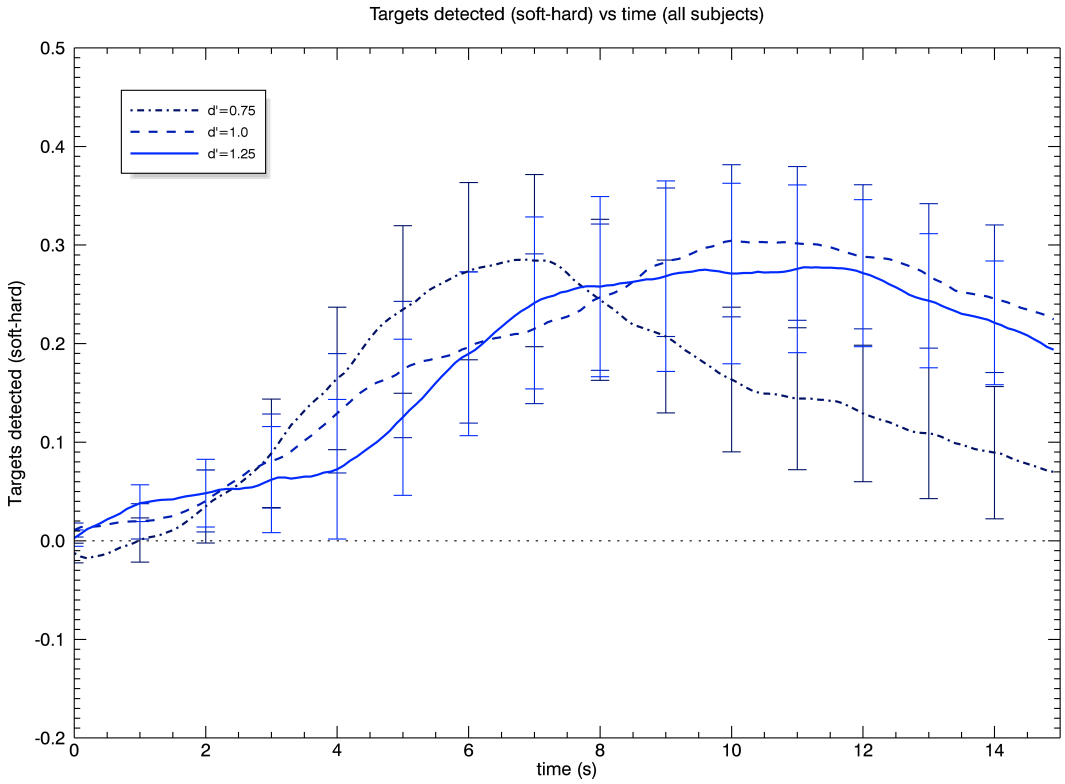


Figure 3.9: Experiment 3. The difference between the mean number of targets detected by that time for the soft and hard highlighting conditions for each d' value. Values above zero mean that by that time subjects had found more targets, on average, when soft highlighting was used than in the hard case. The error bars were calculated using a between subject variability correction [138] and a three second smoothing window was applied.

instructed to continue searching until all targets had been found, we have no evidence concerning the effect of highlighting on an individual's decision to quit searching, and therefore, on the possibility of missed targets. In Experiment 4, we conducted a version of the digit token search task in which the number of targets varied from 0 to 2 and subjects were instructed to continue searching until they were confident that no more targets remained. A button was included in the display to allow subjects to terminate a trial. A trial would also terminate automatically once subjects had found all targets in the display. Trials did not time out.

3.4.1 Methods

Following Experiments 1-3, digit token intensities were drawn from a beta distribution with $\theta = \{1, 2.333, 4, 9\}$ corresponding to $d' = \{0, 1.69, 3.07, 5.77\}$ a control condition of constant 0.5 intensity was also used making 5 conditions total.

Subjects were presented with a grid of 10x10 digits just as in Experiments $1 - 3$ along with a "Done" button that was enabled after 1 second to prevent subjects from clicking through the experiment too quickly. Digits were used only once per experiment.

Subjects completed 80 trials during the experiment with the first five trials regarded as practice and not used in the analysis. The first five trials included one with 0 targets and 2 each with 1 or 2 targets present. The practice trials were presented in random order.

The 75 experimental trials were presented in 5 blocks of 15, with the blocks of 15 composed of one trial each of the 5 highlighting conditions (4 experimental plus 1 control) crossed with 0, 1, or 2 targets present. The sequence of trials was such that across the experiment (5 blocks of 15 trials) there was exactly one trial per highlighting condition C with previous condition P and target number T. This led to 5 levels of the current condition times 5 levels of the previous condition times 3 levels of number of targets to arrive at 75 trials total.

When a target digit token was clicked it was removed from the display and replaced with the background color.

At the end of a trial the subject was shown the grid of digits with the actual targets high-

lighted. If the target was found it was shown in green and if not found it was shown in red. If a target was missed a buzzer was sounded. No sound was made if all targets were located.

Out of 50 subjects who started the experiment 31 completed all trials. The remaining 19 quit the experiment on their own as no subjects were rejected based on performance. Subjects were only allowed to participate once so all subjects were unique. This included those who started but later quit. Subjects accepted the consent form in order to participate in the experiment.

3.4.2 Results

Figure 3.10 shows the fraction of targets found as a function of time for the one target and two target cases. As seen in earlier experiments, stronger classifiers (larger d') enable subjects to locate targets more quickly. For the two target case a random classifier ($d' = 0$) actually hurt performance while even a weak classifier using soft highlighting helped ($d' = 1.69$).

Figure 3.10 shows that high quality classifiers using soft highlighting (larger d') lead to faster detection and also better asymptotic performance. Faster because the slopes for the higher d' curves are steeper than lower d' curves and better because the asymptotes for the two target condition is higher for better quality classifiers. Our previous experiments were unable to examine this issue because subjects had no way to self-terminate a trial without finding a target.

Figure 3.11 shows the mean time to locate a target by condition for the one target and two target present cases. All adjacent pairs were compared and those that are statistically significant are marked with a single star ($p < 0.05$). Paired t-test and Wilcoxon test results are also shown. Strong highlighting ($d' = 5.77$) leads to faster target localization for both one and two targets present.

Figure 3.12 shows three sets of curves corresponding to trials in which 0, 1, or 2 targets are present in the display. Each set of curves indicates, for each point in time within a trial, the proportion of completed trials across the five conditions. A trial is completed either when the subject clicks the done button (for 0-2 targets present) or when all targets have been found (for 1 or 2 targets present). The Figure seems to indicate that trials with stronger classifiers (i.e., larger

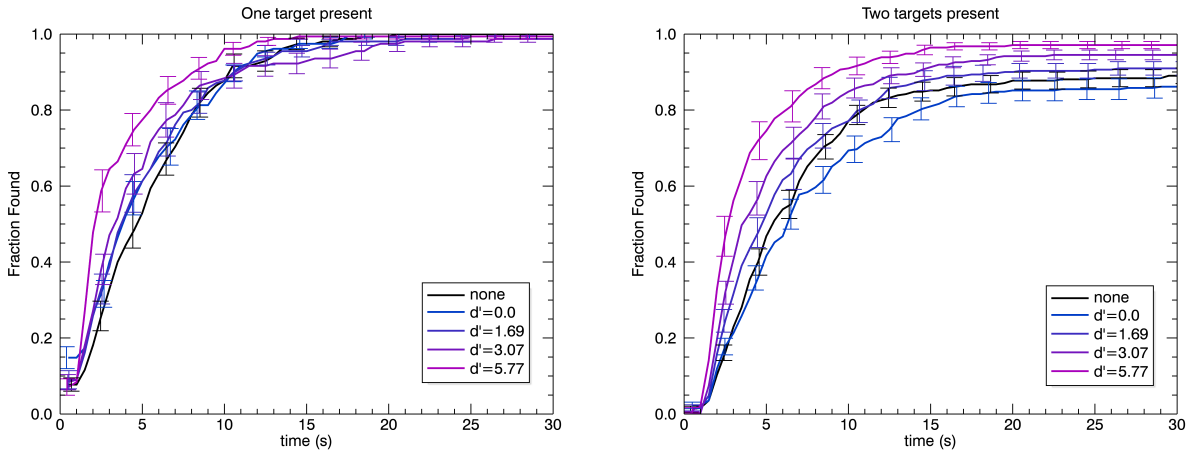
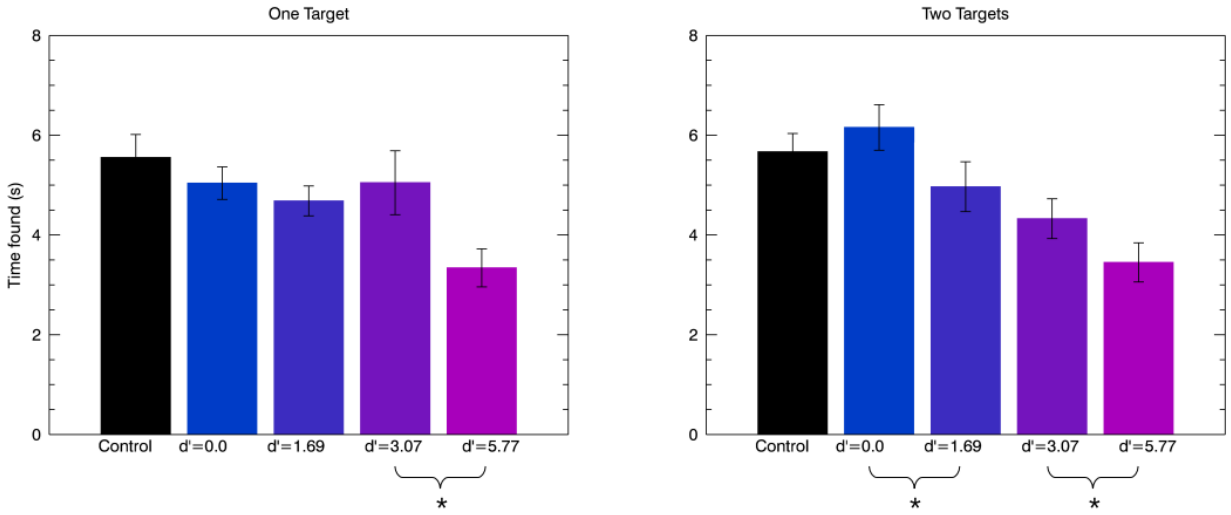


Figure 3.10: Experiment 4. Fraction of targets found as a function of time, number of targets present and condition. Curves are mean across subjects ± SE for each time point.




```
One target:
  d'=3.07 - d'=5.77: 4.3264 vs 3.4504, t(30)= 2.6522 (p= 0.0127)
                                       w(30)= 1165.0 (p= 0.0079)

Two targets:
  d'=0.0  - d'=1.69: 6.1555 vs 4.9680, t(30)= 2.2391 (p= 0.0327)
                                       w(30)= 1129.0 (p= 0.0318)
  d'=3.07 - d'=5.77: 4.3264 vs 3.4504, t(30)= 2.4590 (p= 0.0199)
                                       w(30)= 1154.0 (p= 0.0125)
```

Figure 3.11: Experiment 4. The mean time to locate a target by condition for one target and two targets cases. All pairs were compared. Bars marked with a single star are statistically significant with $p < 0.05$. Paired t-test results are shown along with Wilcoxon signed-rank test results.

d') end sooner, and this effect is amplified as the number of targets increases.

The design of Experiment 4 terminated a trial immediately when any targets present were located. Because of this, we are not able to determine whether subjects spent time after searching after locating the target in the one-target present condition. We suspect that this is not an issue because there is some evidence for quicker self-termination of the trial for larger d' displays when no targets are present (see Figure 3.12, upper left). To the extent that subjects terminate trials faster with larger d', self-termination of 1 and 2 target present displays should show a similar effect.

Figure 3.13 shows the mean time to end a trial by condition along with paired t-tests and two-sided Wilcoxon signed-rank tests. All adjacent pairs of bars were tested. Those that are statistically significant are marked with a single star for $p < 0.05$ and a double star if $p < 0.0001$. The median reaction time of each subject for each condition was calculated and the mean of that time across subjects was used for the plots and the statistical tests. From these results it is clear that stronger highlighting leads to faster localization times.

The analyses above lead to the conclusion that better classifiers (larger d') lead to early termination (faster target localization) in the soft highlighting condition. We found no evidence that a better classifier caused subjects to give up sooner. We found some evidence that highlighting slows subjects relative to the control but only when the classifier was completely uninformative ($d' = 0.0$, single target condition).

We want to know if when subjects failed to find a target whether it was soft highlighting that caused the miss. To examine this, we looked at the number of missed targets for situations where the subject directly terminated the experiment. Specifically, when one target was present and it was missed, when two targets were present and one was missed, and lastly, when two targets were present and both were missed. Plots of these cases are shown in Figure 3.14 where statistically significant differences between the control and a highlight condition are marked with a star ($p < 0.05$) or double star ($p < 0.005$). Below the plot are the results of t-tests and Wilcoxon tests comparing the control condition to each of the highlight conditions for the three cases above. The plot and tests clearly show two things: first, that the uniformative classifier ($d' = 0$) actually hurts performance

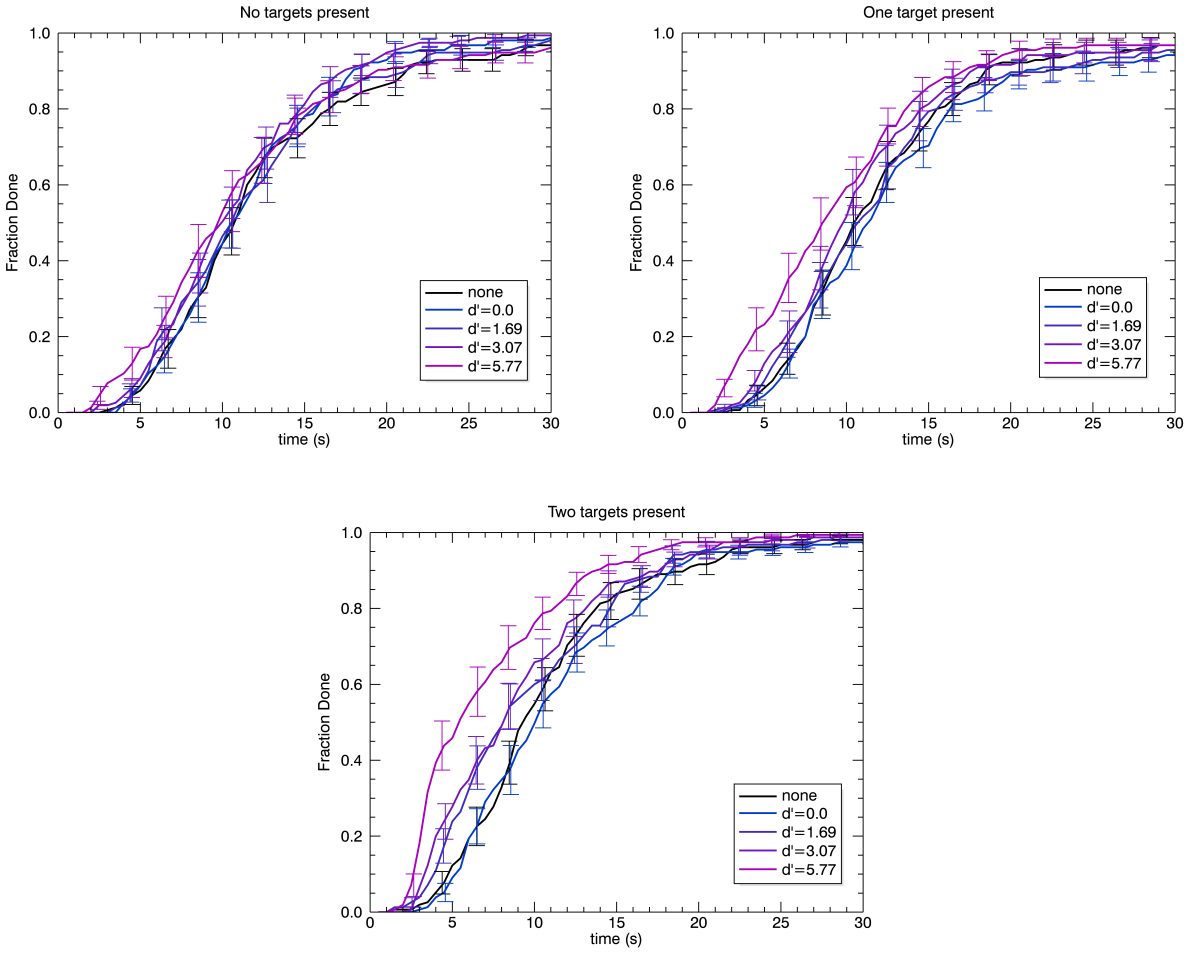
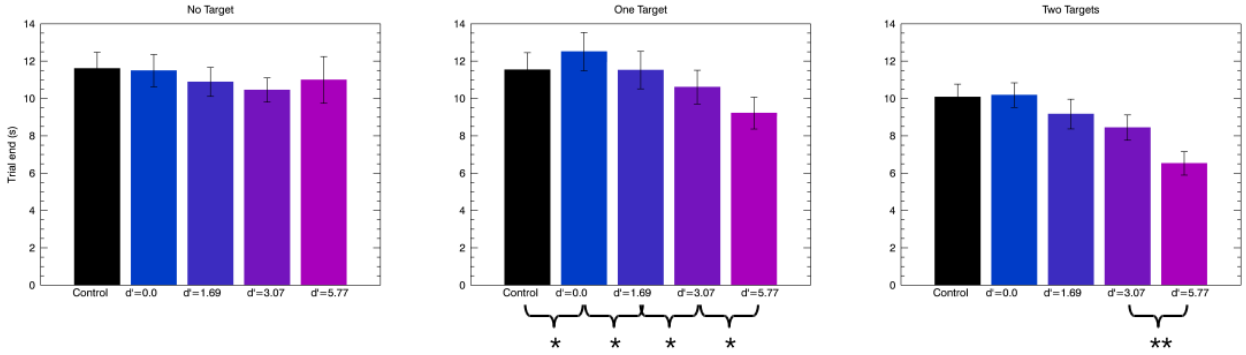


Figure 3.12: Experiment 4. Fraction of trials completed as a function of time, number of targets present and condition. Curves are mean across subjects ± SE for each time point.




```
One target:
  control - d'=0.0 : 11.5364 vs 12.5094, t(30)=  2.1826 (p= 0.0370)
                                         w(30)=   924.0 (p= 0.4598)
  d'=1.69 - d'=3.07: 11.5158 vs 10.6055, t(30)=  2.1230 (p= 0.0421)
                                         w(30)=  1026.0 (p= 0.4859)
  d'=3.07 - d'=5.77: 10.6055 vs  9.2151, t(30)=  2.3116 (p= 0.0278)
                                         w(30)=  1067.0 (p= 0.2026)

Two targets:
  d'=3.07 - d'=5.77:  8.4383 vs  6.5224, t(30)=  6.6789 (p= 0.0000)
                                         w(30)=  1130.0 (p= 0.0307)
```

Figure 3.13: Experiment 4. The mean time to end a trial (in seconds) by condition for no target, one target and two target cases. All pairs were compared. Bars marked with a single star are statistically significant with $p < 0.05$. Bars marked with a double star are significant with $p < 0.0001$. Paired t-test results are shown along with Wilcoxon signed-rank test results.

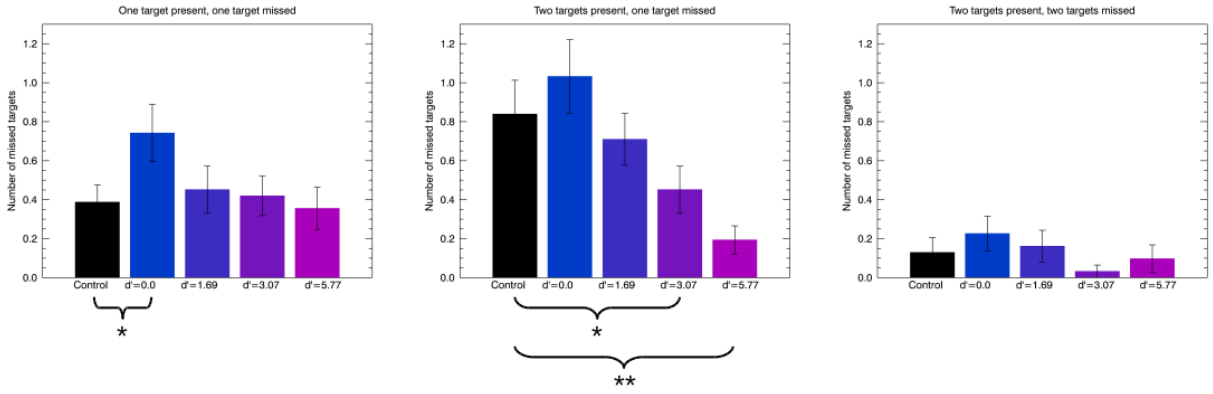
(also seen in Experiment 1) and second, that soft highligting does not hinder target localization but rather results in fewer misses as the quality of the classifier increases (increasing d'). The case of two targets present and both missed was rare and only happened 20 times out of 837 trials where two targets were present.

The analyses of this section demonstrate that soft highlighting leads to improved detection of targets with fewer misses. Specifically, Figure 3.14 shows that subjects missed fewer targets when using soft highlighting in situations where they were able to miss a target.

3.5 Experiment 5: Variable Number of Targets, Subject-Controlled Highlighting

In Experiment 4 the highlights were present throughout the trial. As a result, the darker display elements were more salient and repeatedly attracted attention, whereas the lighter display elements were less salient and subjects may have had difficulty attending to them. To the degree that the highlighting is based on a weak classifier, highlighting has the potential to harm performance by distracting the subject with salient nontargets and masking the targets. Past research has shown that the presence of (hard) highlights can lead to an increase in the number of missed targets [111], [226],[4]. Further, we observed this same effect in Experiment 4 with soft highlighting: lower quality classifiers produced highlights that increased the target miss rate.

Consequently, we hypothesized that subjects may benefit from a scheme in which highlights are present at the onset of a trial but are removed at a later point when the leverage they provide has been exhausted and it becomes easier to search an unbiased display for potentially missed targets. Because we did not know exactly what highlighting schedule would benefit subjects the most, we decided in this experiment to allow subjects to control highlighting themselves. On each trial, highlights were initially present. Subjects were provided with a button that toggled highlights, from on to off and off to on. One goal of this experiment is to determine whether self-directed highlighting improves target detection. But another, looser goal of this experiment is to understand the strategies that subjects use to control highlighting.




```
One target present, one missed:
  control - d'=0.0 : 0.387 vs 0.742, t(30)=-2.160 (p=0.0388), w(30)= 869.0 (p=0.0889)
  control - d'=1.69: 0.387 vs 0.452, t(30)=-0.465 (p=0.6450), w(30)= 974.0 (p=0.9668)
  control - d'=3.07: 0.387 vs 0.419, t(30)=-0.373 (p=0.7120), w(30)= 970.5 (p=0.9206)
  control - d'=5.77: 0.387 vs 0.355, t(30)= 0.215 (p=0.8313), w(30)=1011.0 (p=0.5564)

Two targets present, one missed:
  control - d'=0.0 : 0.839 vs 1.032, t(30)=-1.063 (p=0.2963), w(30)= 926.5 (p=0.4541)
  control - d'=1.69: 0.839 vs 0.710, t(30)= 0.611 (p=0.5458), w(30)= 996.0 (p=0.7669)
  control - d'=3.07: 0.839 vs 0.452, t(30)= 2.555 (p=0.0159), w(30)=1084.0 (p=0.0922)
  control - d'=5.77: 0.839 vs 0.194, t(30)= 3.420 (p=0.0018), w(30)=1168.0 (p=0.0016)

Two targets present, both missed:
  control - d'=0.0 : 0.129 vs 0.226, t(30)=-1.139 (p=0.2636), w(30)= 931.5 (p=0.3002)
  control - d'=1.69: 0.129 vs 0.161, t(30)=-0.328 (p=0.7448), w(30)= 961.5 (p=0.7005)
  control - d'=3.07: 0.129 vs 0.032, t(30)= 1.139 (p=0.2636), w(30)=1008.0 (p=0.2976)
  control - d'=5.77: 0.129 vs 0.097, t(30)= 0.297 (p=0.7685), w(30)= 991.5 (p=0.6546)
```

Figure 3.14: Experiment 4. Mean number of missed targets by condition for the cases where the subject directly terminated the trial. Specifically, cases when there was one target present and it was missed (*left*), when two targets were present and one was missed (*middle*) and lastly when two targets were present and both were missed (*right*). Below, the results of comparisons between the control condition and the four highlight conditions (t-tests and Wilcoxon signed-rank tests). Significant differences are marked in the figure with a single star ($p < 0.05$) or double star ($p < 0.005$). As seen in previous experiments (e.g. Experiment 1) the uninformative classifier ($d' = 0$) actually causes more misses relative to the control condition.

3.5.1 Methods

The conditions and configuration of Experiment 5 exactly mimic those of Experiment 4 with the addition of a button allowing the subjects to freely toggle highlighting on or off. When on, the grid of digits appeared as in Experiment 4 according to the classification strength simulated by the beta distribution. When off, the grid appeared as in the control condition in Experiment 4.

During a trial the grid of digit tokens was displayed in the center of the subject's browser window. Directly below the grid was the done button (labeled "Done") used to end the trial. Directly below the done button was the toggle button (labeled "Toggle highlighting"). All three display elements: the grid of digit tokens, the done button, and the toggle button, were aligned vertically on the display.

3.5.2 Results

Figure 3.15 shows the fraction of trials completed as a function of time by number of targets present in the scene. As in Experiment 4, Figure 3.12, there is a trend of soft highlighting enabling faster times as the number of targets increases. Figure 3.16 shows the mean time to end a trial by condition with significant pairs indicated (computation as in Experiment 4). Here the two target case is significant for the strongest highlighting.

A similar set of plots can be made to examine the fraction of targets found as a function of time as shown in Figure 3.17 where Figure 3.18 shows the mean time to locate a target for both one and two target cases along with the result of a paired t-test across subjects for different conditions. In this case the only significant adjacent pairing is again for the strongest highlighting condition.

The results above closely match those of Experiment 4 and further support the conclusion that soft highlighting is beneficial to subjects.

Unlike Experiment 4, in Experiment 5 subjects were able to toggle highlighting on and off at will. In Figure 3.20 we see the mean fraction of trials in which the subject toggled highlighting at least once by number of targets and condition.

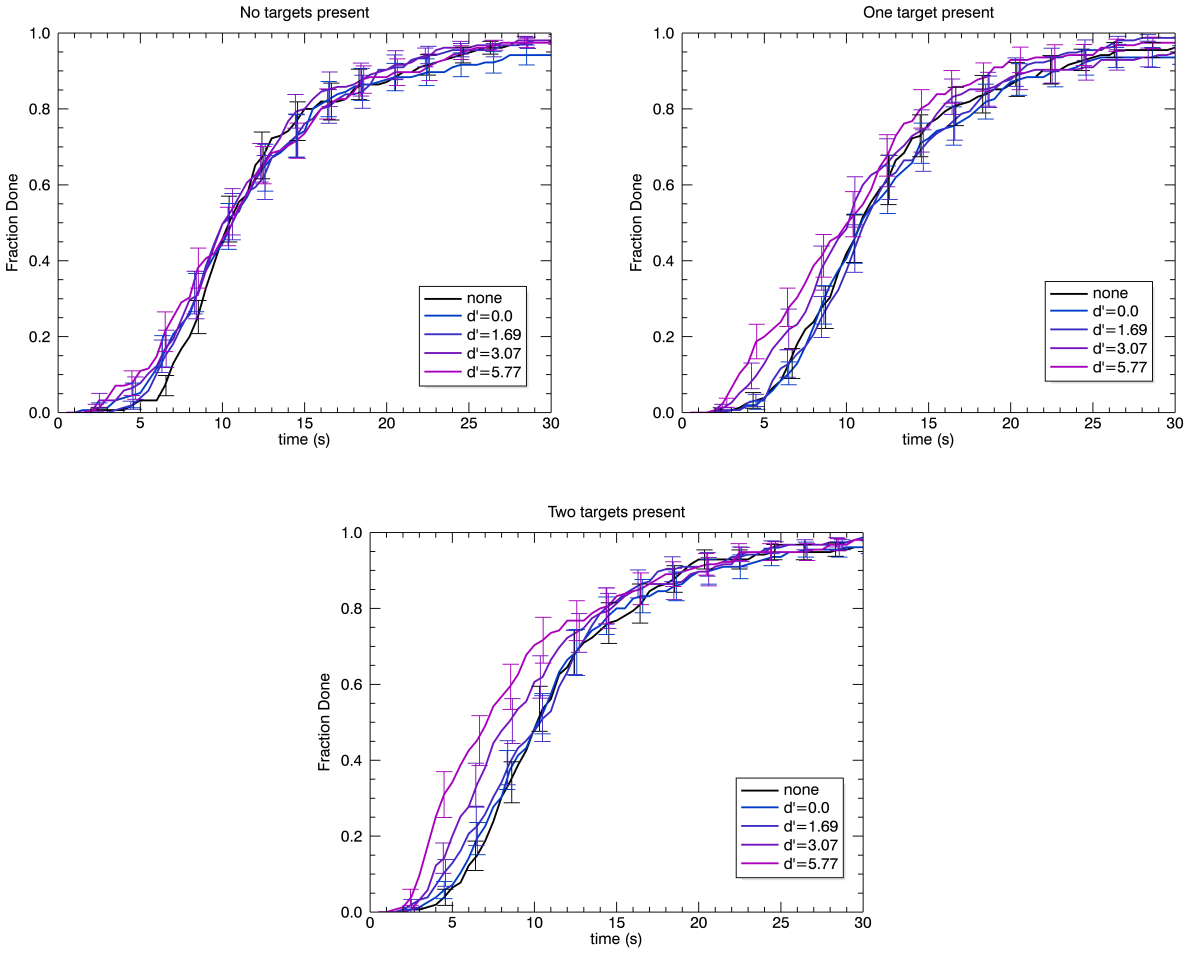
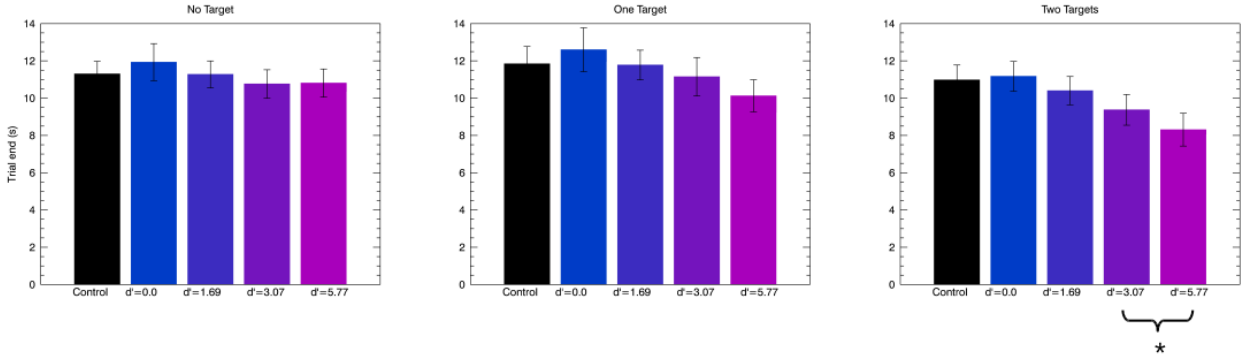


Figure 3.15: Experiment 5. Fraction of trials completed as a function of time, number of targets present and condition. Curves are mean across subjects ± SE for each time point.




```
Two targets:
  d'=3.07 - d'=5.77: 9.3745 vs 8.3064, t(30)= 3.0474 (p= 0.0048)
                                       w(30)= 1066.0 (p= 0.2077)
```

Figure 3.16: Experiment 5. The mean time to end a trial (s) by condition for no target, one target and two target cases. Bars marked with a single star are statistically significant with $p < 0.01$. Paired t-test results are shown along with Wilcoxon signed-rank test results.

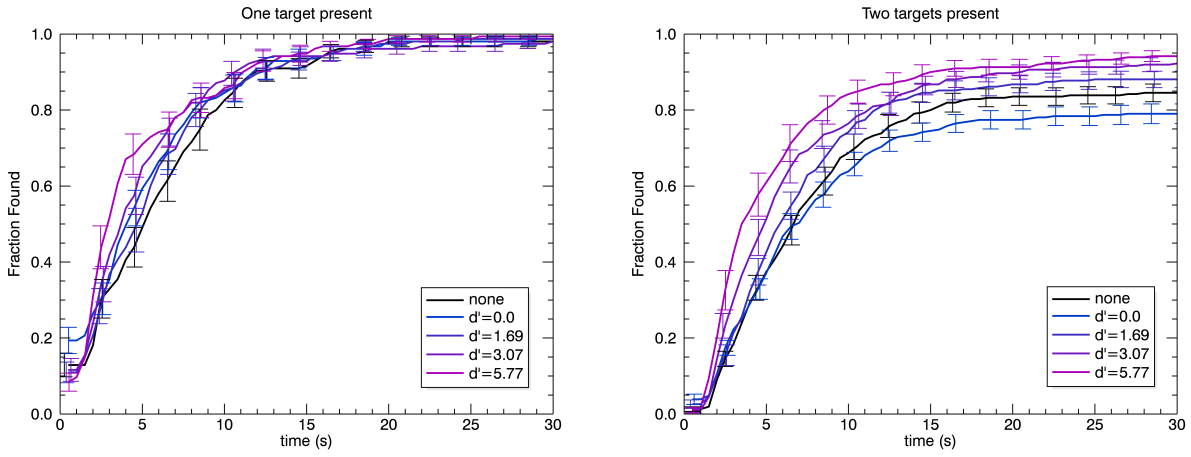
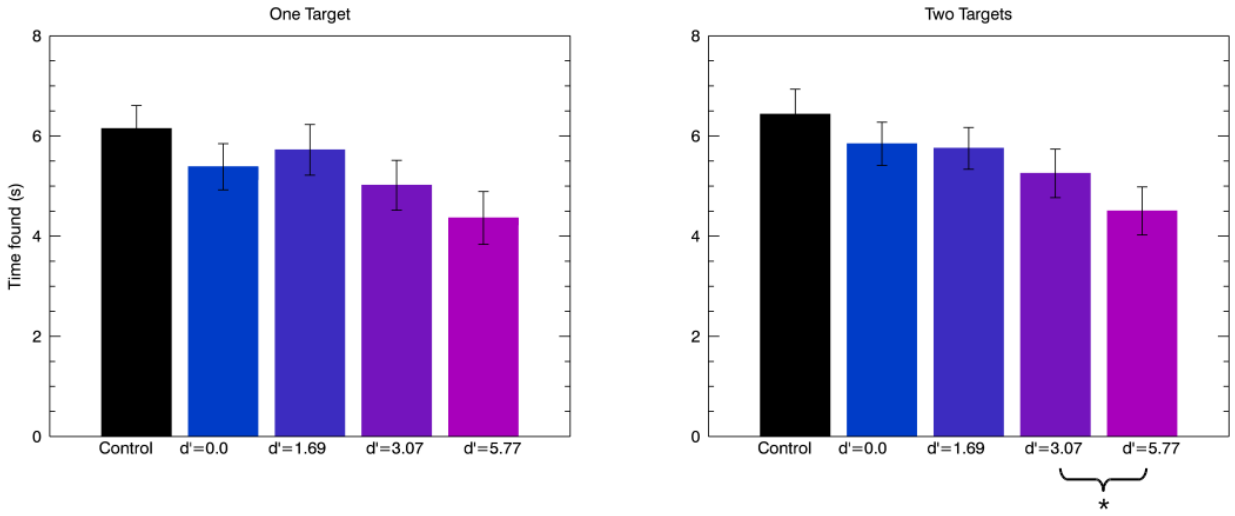


Figure 3.17: Experiment 5. Fraction of targets found as a function of time, number of targets present and condition. Curves are mean across subjects ± SE for each time point.




```
Two targets:
  d'=3.07 - d'=5.77: 5.2528 vs 4.5030, t(30)= 2.8215 (p= 0.0084)
                                       w(30)= 1072.5 (p= 0.1765)
```

Figure 3.18: Experiment 5. The mean time to locate a target by condition for one target and two targets cases. Bars marked with a star are statistically significant with $p < 0.01$). Paired t-test results are shown along with Wilcoxon signed-rank test results for significant pairs. All pairs of bars were tested.

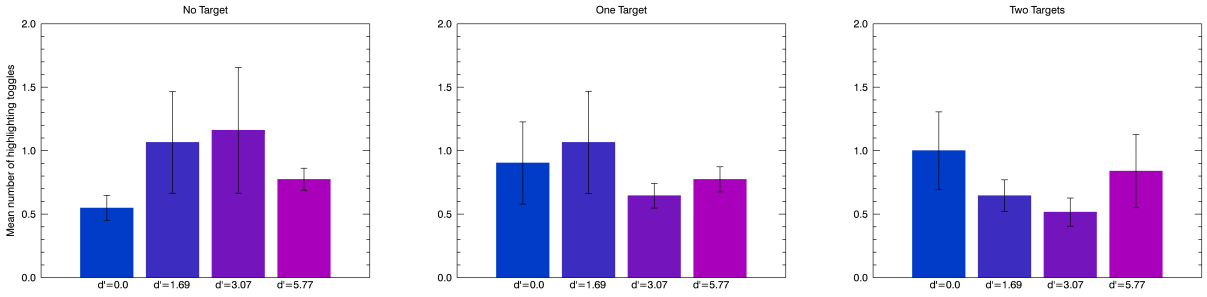
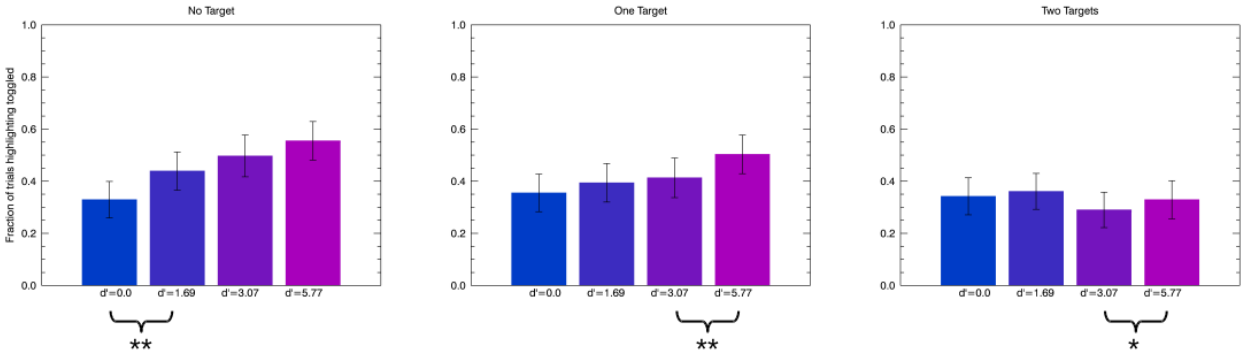


Figure 3.19: Experiment 5. Mean number of highlighting toggles by number of targets and condition. If a subject turns highlighting off and then back on it is counted as two toggles. Statistical tests showed no pair-wise differences as significant. On the whole, subjects chose to either leave highlighting on or turned it off and left it off.

One can also look at the number of times subjects toggled highlighting. Here "toggle" means a change of highlighting state so that turning highlighting off and then back on would be counted as two changes. Figure 3.19 shows the mean number of highlighting toggles by number of targets and condition. Paired t-tests showed all pair-wise differences as significant however, no trend is apparent. What is apparent is that subjects most often left highlights on for the entire trial (see Figure 3.20) or when they did change the highlighting state they turned the highlights off and left them off.

This behavior was unanticipated. It was expected that subjects would toggle highlighting on and off, perhaps several times, in order to add motion to the display in the hopes that it would make targets stand out. Understanding why subjects did not do this is a possible area for future research.

Figure 3.20 suggests a trend for the 0- and 1-target trials: the likelihood that subjects will toggle highlighting increases with the quality of the classifiier providing highlights. The statistics of highlights in the 4 experimental conditions are quite different: with a weak classifier, highlights are continuous in $[0, 1]$; with a strong classifier, highlights are strongly binary, close to 0 or 1. Consequently, if subjects were concerned that they had missed a target in a display with strong highlights, they would need to turn off highlighting to inspect all display elements.




```
No targets:
  d'=0.0  - d'=1.69: 0.3290 vs 0.4387, t(30)= -3.1788 (p= 0.0034)
                                       w(30)=  902.50 (p= 0.2763)

One target:
  d'=3.07 - d'=5.77: 0.4129 vs 0.5032, t(30)= -3.1047 (p= 0.0041)
                                       w(30)=  921.5  (p= 0.4234)

Two targets:
  d'=3.07 - d'=5.77: 0.3613 vs 0.2903, t(30)=  2.4762 (p= 0.0191)
                                       w(30)=  1027.0 (p= 0.4455)
```

Figure 3.20: Experiment 5. Mean fraction of trials in which the subject toggled highlighting at least once. Bars marked with a single star at the same height are statistically significant with $p < 0.05$. Those marked with double stars are statistically significant with $p < 0.005$. Paired t-test results are shown along with Wilcoxon signed-rank test results.

We have no clear explanation for why the same trend was not observed for 2-target trials. However, 2-target trials are special in that once subjects detected a first target, if the trial did not end then subjects knew for certain that a second target was present. In the 0- and 1-target trials, they had no such assurance as they continued their searches.

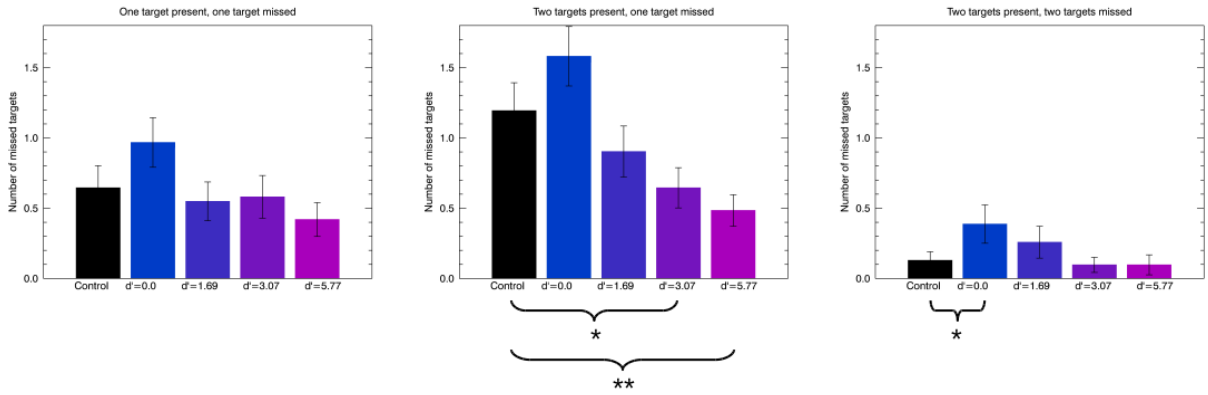
As in Experiment 4, we wanted to know if when subjects failed to find a target whether it was soft highlighting that caused the miss even when the subject had the ability to turn highlighting off. Figure 3.21 shows misses for the cases when the subject ended the trials on his or her own. Below the plot are the results of t-tests and Wilcoxon tests comparing the control condition to each of the highlight conditions. The plot and tests mirror the results for Experiment 4 (see Figure 3.14) showing that even when highlighting was under the subject's control that the uniformative classifier ($d' = 0$) actually hurts performance and that soft highligting does not hinder target localization but rather results in fewer misses as the quality of the classifier increases (increasing d'). The case of two targets present and both missed was rare and only happened 30 times out of 837 trials where two targets were present.

3.6 Discussion

The experiments of this chapter were intended to initiate an investigation into the utility and effectiveness of soft highlighting techniques as compared to no highlighting or the more traditional hard highlighting. The issues raised in [111], [226] and [4] provided the impetus for the experiments.

In Experiment 1 we were interested in whether soft highlighting would lead to improved target detection rates. We measured the time it took subjects to find a fixed number of targets in a grid of digits whose highlight itensity was determined by a stochastic classifier. The results of Experiment 1 indicated that this was indeed the case. The most intriguing finding of Experiment 1 is that the classifier providing highlights does not have to be terribly strong to be helpful. Even a weak classifier with $d' = 0.75$ provided subjects enough signal that detection speed improved over the control condition.

In Experiment 2, we asked whether soft highlighting affects the time to process each display




```
One target present, one missed:
  control - d'=0.0 : 0.645 vs 0.968, t(30)=-1.718 (p=0.0960), w(30)= 884.0 (p=0.1591)
  control - d'=1.69: 0.645 vs 0.548, t(30)= 0.571 (p=0.5722), w(30)=1006.0 (p=0.6403)
  control - d'=3.07: 0.645 vs 0.581, t(30)= 0.349 (p=0.7299), w(30)=1005.0 (p=0.6518)
  control - d'=5.77: 0.645 vs 0.419, t(30)= 1.191 (p=0.2429), w(30)=1048.0 (p=0.2477)

Two targets present, one missed:
  control - d'=0.0 : 1.194 vs 1.581, t(30)=-1.460 (p=0.1546), w(30)= 890.0 (p=0.2076)
  control - d'=1.69: 1.194 vs 0.903, t(30)= 1.393 (p=0.1738), w(30)=1047.0 (p=0.2967)
  control - d'=3.07: 1.194 vs 0.645, t(30)= 2.373 (p=0.0243), w(30)=1108.5 (p=0.0477)
  control - d'=5.77: 1.194 vs 0.484, t(30)= 3.803 (p=0.0007), w(30)=1148.0 (p=0.0095)

Two targets present, both missed:
  control - d'=0.0 : 0.129 vs 0.387, t(30)=-2.794 (p=0.0090), w(30)= 908.5 (p=0.1637)
  control - d'=1.69: 0.129 vs 0.258, t(30)=-1.278 (p=0.2111), w(30)= 943.5 (p=0.4665)
  control - d'=3.07: 0.129 vs 0.097, t(30)= 0.441 (p=0.6621), w(30)= 992.0 (p=0.6906)
  control - d'=5.77: 0.129 vs 0.097, t(30)= 0.571 (p=0.5722), w(30)=1005.5 (p=0.4257)
```

Figure 3.21: Experiment 5. Mean number of missed targets by condition for the cases where the subject directly terminated the trial. Specifically, cases when there was one target present and it was missed (*left*), when two targets were present and one was missed (*middle*) and lastly when two targets were present and both were missed (*right*). Below, the results of comparisons between the control condition and the four highlight conditions (t-tests and Wilcoxon signed-rank tests). Significant differences are marked in the figure with a single star ($p < 0.05$) or double star ($p < 0.005$). As seen in previous experiments (e.g. Experiment 4) the uninformative classifier ($d' = 0$) actually causes more misses relative to the control condition.

element or the time to perform initial segmentation and processing of the display. We found that highlighting affects the rate of search, as determined by decreasing search slopes (response time per display element) as the quality of the classifier increased. This finding is consistent with highlighting modulating attention and allowing subjects to search fewer display elements.

In Experiment 3, we directly compared soft and hard highlighting. We contrasted soft versus hard highlighting with the same quality classifier. The soft classifier provided a graded signal whereas the hard classifier provided a binary signal. However, in matched displays the target discriminability provided by the classifier as identical. Experiment 3 clearly shows a uniform advantage of soft over hard highlighting.

Experiments 1-3 studied search in a situation where each display had a fixed, nonzero number of targets. (There were 10 targets per display in Experiments 1 and 3, one target per display in Experiment 2.) A more naturalistic scenario involving visual search is when the number of targets in the display is unknown from trial to trial. For Experiment 4 we varied the number of targets between zero and two and asked subjects to search until they were confident all targets had been found. Subjects had the ability to terminate a trial at will, leading to the potential of missed targets. One concern about highlighting is that it may facilitate detection of easy targets but cause subjects to give up quicker and therefore obtain more target misses. This concern was not supported by Experiment 4: the stronger highlights yielded faster detection and if anything lower miss rates. (For single-target displays, there was no difference in miss rates depending on the classifier quality. For two-target displays, increased classifier quality led to reliably lower miss rates.) Another concern is that on target-absent trials, subjects may search longer if the display contains highlights. This concern was also alleviated by the experiment: with 0-targets displays, subjects were no slower to end a trial if the display contained meaningful highlights (i.e., highlights produced by a classifier with $d' > 0$) than they were in the control condition.

In Experiments 1-4, highlights were present from the start of the trial until the trial ended. In Experiment 5, we replicated Experiment 4 but offered subjects control over highlights. Each trial began with highlights turned on a toggle button allowed subjects to switch the highlights on

and off. We expected that subjects would turn highlighting off in cases where highlighting was not helpful, the weak highlighting cases. Instead, we observed that subjects turned highlighting off for displays highlighted by strong classifiers but did not toggle highlighting for displays highlighted by weak classifiers. Strong highlighting makes targets more likely to be highly salient but also makes nontargets much less salient (lighter colored against a light background). Because of this, it is possible that subjects were turning highlighting off to avoid missing potentially less salient targets due to a false negative case of the classifier with the masking of potentially unhighlighted targets increasing as the number of strongly highlighted targets decreases which is the strong highlighting case.

Given that we found a robust advantage of soft highlighting for the artificial, segmented displays studied in this series of experiments, the next step of our research program is to explore highlighting in more complex, naturalistic displays, e.g., satellite imagery. The analysis of satellite imagery is a crucial task in the modern world and is presently performed in a very labor-intensive manner by many image analysis. The rate at which satellite imagery is acquired is increasing so any speed up or improvement of the output from imagery analysis would be important.

The challenge of naturalistic displays is that the display elements are not neatly segmented, so more sophisticated methods are needed to determine image highlights. To this end, we abandoned our stochastic classifiers and focused on training a modern, state-of-the-art deep learning classifier for detecting targets in satellite imagery. The development of this classifier is the focus of Chapter 4. With the classifier fully specified and trained, in Chapter 5 we use the classifier to explore highlighting in complex, continuous images.

Chapter 4

Training a Classifier to Locate McDonald's Restaurants in Satellite Imagery

In Chapter 3, we described experiments with artificial, segmented images. In Chapter 5, we will turn to experiments with naturalistic images, in particular, satellite imagery. Simulating a classifier for naturalistic images is problematic for several reasons. First, the image is continuous and a target could potentially be centered at every pixel of the image. Second, the benefit of highlighting will depend not only on overall classifier quality but on the specific sorts of errors that the classifier makes. For these reasons, we constructed deep neural network classifiers that will be used in the experiments presented in Chapter 5.

In this Chapter, we describe the development of these classifiers and argue that they reflect a state-of-the-art approach to machine learning. We present the classifiers in Section 4.1 characterizing their performance in Section 4.2 and finally offer our justification for the classifier selected for the experiments of Chapter 5 in Section 4.3. We follow with three implementation sections on building a training data set (Section 4.4), training classifiers (Section 4.5) and testing classifiers (Section 4.6).

4.1 The McDonald's Classifiers

It is now well-known that state-of-the-art classifiers for object recognition use convolutional neural networks [115] [109]. A convolutional neural network is an extension of a traditional feed-forward neural network that prepends the fully connected layers with one or more convolutional and pooling layers.

The purpose of the convolutional layers is to enable the network to learn a set of filters which are useful for detecting parts of objects in the images of the training set. Pooling takes the output of these convolutional filters and groups them into larger blocks so that higher layers in the network have receptive fields that span increasingly larger patches of the images. For example, see [116].

Visually, a convolutional neural network is arranged as in Figure 4.1. The input image is on the left. The first convolutional layer (`conv1`) convolves a set of small kernels over the input image producing a set of output bands representing the effect of the kernels on the input. Here "band" refers to a 2D array of responses to the convolution kernel and is directly analogous to the red, green or blue bands of an RGB image. Pooling takes these bands and rescales them spatially (`pool1`) by keeping the largest value in each band when convolving a 2x2 region over the input with a stride of 2, ie, keep the largest value in each 2x2 region, per band. This process is repeated (`conv2`, `pool2`) for the next set of convolutional and pooling layers. Finally, the last pooling layer output is passed through two fully connected layers (`fc1`, `fc2`) to an output logistic layer to calculate the final probability of target present in the input image.

We experimented with five different classifier architectures to identify the best performing model. The architecture parameters were selected based on results from a proprietary prototype Bayesian optimization search tool based on [200]. The five architectures explored are variations on the base architecture returned by the search tool and are summarized in Table 4.1. A wider search could have been implemented but was not because it was felt that the search tool had already produced a strongly performing network.

Each classifier followed the basic plan of two convolutional and pooling layers followed by one or more fully connected layers leading to a logistic output layer. In Table 4.1 the convolutional layers are represented as a triplet of integers followed by the letter "c" as in `35-7-1 c`. This is shorthand for a convolutional layer with 35 output filters each consisting of a 7 by 7 pixel kernel which is convolved over the input with a stride of one pixel. Each convolutional layer included a rectified linear nonlinearity followed by local response normalization. See [109] for justification.

After the convolutional layer comes a pooling layer. In Table 4.1 these are denoted as two

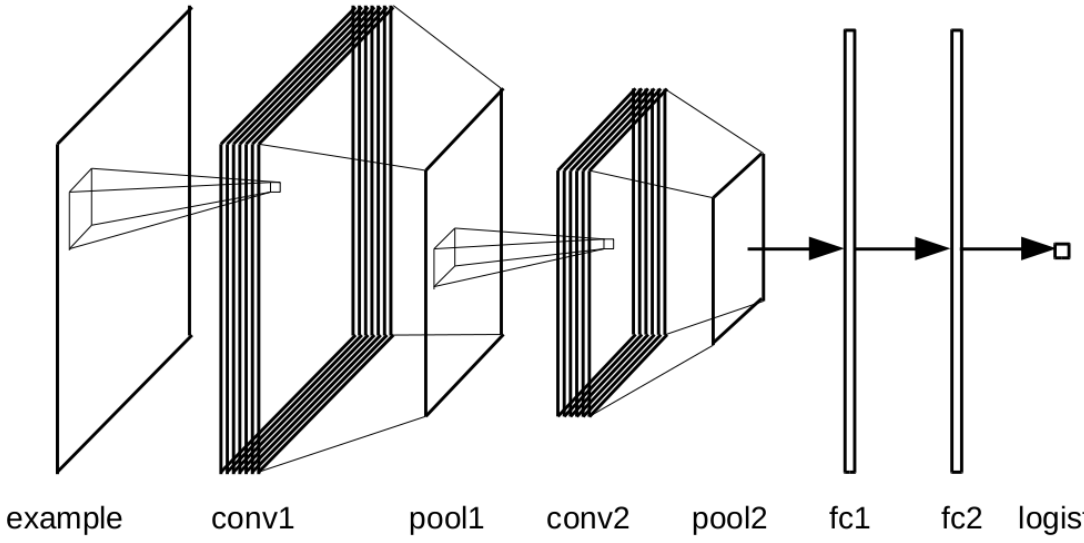


Figure 4.1: A schematic representation of a convolutional neural network. This example matches the architecture of the networks discussed in this chapter. The input image is on the left. The first convolutional layer (`conv1`) convolves a set of small kernels over the input image producing a set of output bands representing the effect of the kernels on the input. Pooling takes these bands and rescales them spatially (`pool1`). This process is repeated (`conv2`, `pool2`) for the next set of convolutional and pooling layers. Finally, the last pooling layer output is passed through two fully connected layers (`fc1`, `fc2`) to an output logistic layer to calculate the final probability of target present in the input image.

mcdonalds1	mcdonalds2	mcdonalds3	mcdonalds4	mcdonalds5
input	input	input	input	input
↓	↓	↓	↓	↓
35-7-1 c	35-7-1 c	35-7-1 c	35-7-1 c	35-7-1 c
↓	↓	↓	↓	↓
ReLU,LRN	ReLU,LRN	ReLU,LRN	ReLU,LRN	ReLU,LRN
↓	↓	↓	↓	↓
3-2 p	2-2 p	2-2 p	2-2 p	2-2 p
↓	↓	↓	↓	↓
128-5-3 c	128-5-3 c	128-5-3 c	128-5-3 c	128-5-3 c
↓	↓	↓	↓	↓
ReLU,LRN	ReLU,LRN	ReLU,LRN	ReLU,LRN	ReLU,LRN
↓	↓	↓	↓	↓
7-3 p	2-2 p	2-2 p	2-2 p	2-2 p
↓	↓	↓	↓	↓
512,ReLU,drop	1000,ReLU,drop	1000,ReLU,drop	1000,ReLU,drop	1000,ReLU,drop
↓	↓	↓	↓	↓
output	1000,ReLU,drop	1000,ReLU,drop	1000,ReLU,drop	1000,ReLU,drop
	↓	↓	↓	↓
	output	output	output	output

model	**image type**	**minibatch**	**mean subtract**	**nontarget:target ratio**
mcdonalds1	grey	128	yes	10:1
mcdonalds2	grey	128	yes	10:1
mcdonalds3	grey	128	no	10:1
mcdonalds4	grey	128	no	1:1
mcdonalds5	pale color	128	no	10:1

Table 4.1: The five architectures trained for McDonald's detection in satellite images. The combination of model architecture (top) and image type, minibatch size, mean image subtraction and training data ratio (bottom) form the different architectures. All learning parameters were the same for each model: train to 1,000,000 minibatches with a base learning rate of 0.01. For each model, the input training data was augmented 10x for targets and put into a 10:1 nontarget:target ratio by selecting 10x as many nontargets as targets. Layer encoding is described in the text.

integers followed by the letter "p". So, `2-2 p` is a maximum pooling layer with a 2x2 pixel kernel with a stride of two pixels. This particular pooling layer reduces the input by a factor of two in each direction. Similarly, a `7-3 p` pooling layer replaces each 7x7 region with the maximum response and then steps over three pixels.

After the second pooling layer are one or more fully connected layers. These are traditional neural network layers which map the output of the layer below to a given number of nodes. In Table 4.1 fully connected layers follow the form `1000,ReLU,drop` for a 1000 node layer which uses a rectified linear nonlinearity and dropout [87] during training with a probability of 0.5.

The bottom part of Table 4.1 details, for each classifier, the source imagery type, minibatch size, whether a mean image was subtracted, and the ratio between nontargets and targets in the training data set. Mean image subtraction involves calculating, per pixel, the mean value across all training data and subtracting the resulting image from each input training example before passing it through the network.

All classifiers used grey scale input imagery of the type shown in Figure 4.4 from the source described in Section 4.4 except for the `mcdonalds5` classifier which used pale RGB color instead. The exact same mix of train and test data was used as for the other classifiers but the original RGB data was converted to pale color by mapping from RGB to HLS color space (hue-lightness-saturation), dividing the saturation by 2, and mapping back to RGB. Figure 4.2 shows example target images in their original pansharpened color, pale color and grey scale versions.

4.2 Comparing the Classifiers

The classifiers defined in Table 4.1 were run against test data, never seen by the classifier during training, to produce a series of test statistics:

Figure 4.2: Example targets as full pansharpened RGB (left), pale color (center), and grey scale (right). The pale color was formed by mapping the RGB image to the HLS color space, dividing saturation by two, and mapping back to RGB.

Model	TP	TN	FP	FN	SENS	SPEC	ACC	AUC	EER
mcdonalds1	2597	29803	117	395	0.8680	0.9961	0.9844	0.9926	0.0309
mcdonalds2	2700	29783	137	292	0.9024	0.9954	0.9870	0.9931	0.0255
mcdonalds3	2651	29768	152	341	0.8860	0.9949	0.9850	0.9920	0.0301
mcdonalds4	2950	29367	553	42	0.9860	0.9815	0.9819	0.9981	0.0164
mcdonalds5	3256	29627	6	23	0.9930	0.9998	0.9991	0.9995	0.0015

Table 4.2: The McDonald's classifier's performance on the test data set. The statistics are true positive count (TP), true negative count (TN), false positive count (FP), false negative count (FN), sensitivity (SENS), and specificity (SPEC), all at a threshold probability of 0.5. Also accuracy (ACC), area under the ROC curve (AUC), and the equal error rate (EER).

TP	true positive count (at a 50% threshold)
TN	true negative count
FP	false positive count
FN	false negative count
SENS	sensitivity (at a 50% threshold)
SPEC	specificity (at a 50% threshold)
ACC	accuracy (at a 50% threshold)
AUC	area under the ROC curve
EER	equal error rate

where the designation "at a 50% threshold" means a probability value of 0.5 was used to decide whether the sample is counted as a TP, TN, FP, or FN. For most neural networks this is a reasonable choice of threshold value. Table 4.2 shows the statistics above for each of the McDonald's classifiers. Recall that the test data seen by each classifier was the same.

Table 4.2 shows clear differences between the classifiers. The most informative columns are AUC and EER for the area under the ROC curve and the equal error rate, respectively. All other columns in Table 4.2 implicitly use a threshold value of 0.5.

The ROC curve is generated by varying the threshold used to tabulate the TP, TN, FP and FN counts from $[0, 1]$ in even increments. Each set of TP, TN, FP, and FN counts for each threshold leads to a specific sensitivity and specificity value. Each of these values are plotted as $(\text{sensitivity}, 1 - \text{specificity})$ to generates a single point on the ROC curve. Changing the threshold from zero up to one traces out the ROC curve parametrically. The curves below were generated using 100 threshold steps. The area under the ROC curve is estimated from the points,

$$\text{AUC} = \frac{1}{2}\sum_{i=1}^{N} |(x_i - x_{i-1})(y_i + y_{i-1})|$$

with $x = 1 - \text{specificity}$ and $y = \text{sensitivity}$. The equal error rate is estimated as the smallest distance from

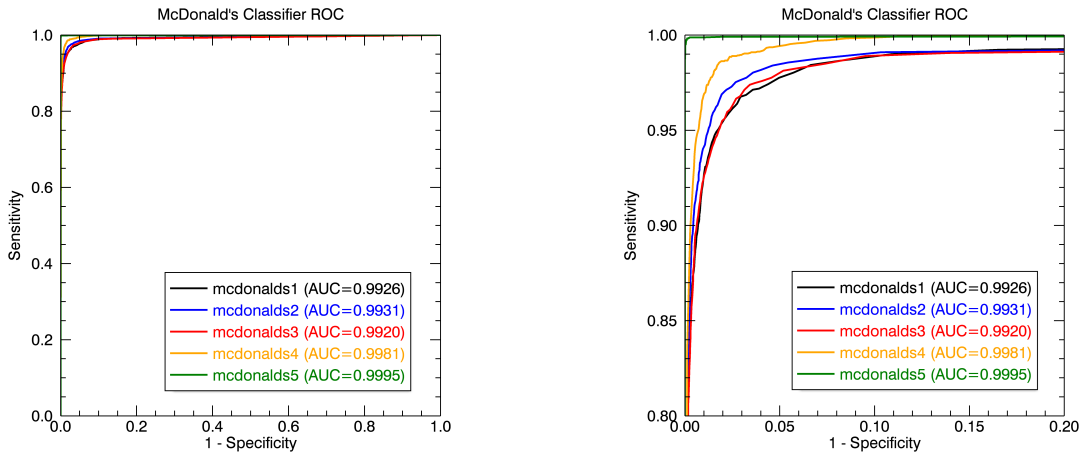


Figure 4.3: ROC curves for the five McDonald's classifiers along with AUC and EER values. Left: the full range plot. Right: zoomed to show the differences between the classifiers. All five classifiers performed quite well on the test data.

the curve points to the upper left corner of the ROC plot divided by the length of the diagonal ($\sqrt{2}$),

$$\text{EER} = \frac{1}{\sqrt{2}} \min_i((x_i - 0)^2 + (y_i - 1)^2))^{\frac{1}{2}}$$

Figure 4.3 shows the ROC curves for each of the five classifiers. The full ROC curve is on the left clearly indicating that the classifiers all do a good job of classifying the test data. The curves are zoomed on the right to show differences. The AUC and EER for each is indicated in the figure.

4.3 Final Classifier Selection and Justification

The experiments in Chapter 5 all used the same McDonald's classifier. The classifier that was selected for the experiments was `mcdonalds3`. Summary statistics from Table 4.2 for this classifier are,

Model	TP	TN	FP	FN	SENS	SPEC	ACC	AUC	ERR
mcdonalds3	2651	29768	152	341	0.8860	0.9949	0.9850	0.9920	0.0301

Clearly, as seen in Figure 4.3, `mcdonalds3` was not the best performing of the classifiers, so why was it selected for the experiments? Two key factors went into selecting this classifier over the

others:

(1) The classifier operates on grey scale imagery.

(2) The classifier is not so good that it will not make mistakes. In particular, the FP and FN counts for the test data are such that targets will be missed by the classifier along with false target detections.

The first reason is important because a practical goal of this work is to use a high-quality classifier on a hard task that is directly applicable to real-world use. In this case, the real-world use is an image analyst evaluating satellite imagery. While consumer-grade systems like Google Earth show nice color images, color imagery is expensive to acquire because it depends on the acquisition of high resolution panchromatic images and simultaneous (or registered) images from a multispectral sensor in order to generate a pansharpened RGB image. This is not practical nor desirable for the image analyst whose job is to locate rare targets in high resolution imagery in a timely manner or to characterize all target structures (or all structures if mapping) as to type and use. For such work panchromatic grey scale imagery is the norm, hence the desire to use grey scale imagery in the classifier experiments.

The second reason is also practical. The McDonald's target was selected because it is fairly distinctive to people who are familiar with the environment of suburban North America. Moreover, the selected training set consciously focused on those McDonald's that were of the "classic" red roof variety. All of this was to make the exercise of locating McDonald's less frustrating for the novices who were the subjects of the experiments. As it turned out, using even pale color as in mcdonalds5 gave the classifier so strong a clue as to the target that the resulting classifier was extremely good at finding them. See the mcdonalds5 results in Table 4.2 or the ROC curve in Figure 4.3.

This level of classifier performance is likely unrealistic for the sorts of targets that an image analyst would be seeking to characterize and as seen above, color imagery is not typically used. This argues against mcdonalds5 as the classifier. Color was tested to see how much harder the pure grey scale task might be. Performance considerations like these also disqualified mcdonalds4.

Of the remaining candidate classifiers `mcdonalds3` fell in the middle in terms of EER (0.0301) and so was the one selected as meeting the criteria of operating on grey scale images while also showing enough errors to be a plausible example of a modern state-of-the-art classifier for a hard problem.

The performance of all the classifiers considered is a testament to the paradigm shift that has happened in artificial neural networks. All of the classifiers are excellent and even spectacular by the standards of classical pattern recognition. The splitting of hairs as to performance is a nice problem to have when selecting a classifier for what in the past would have been a very hard task indeed.

4.4 Implementation: Building a Training Data Set

The satellite imagery used in these experiments is *pansharpened* three band (RGB) imagery from MapQuest available for academic use through their developer program. Pansharpening is a technique which merges high spatial resolution panchromatic imagery with lower resolution multispectral imagery typically using the visible red, green and blue light bands. The resulting RGB image is close to the color that would be seen by the human eye. Pansharpened imagery accounts for most of the imagery seen on websites like Google Maps. For examples and further details see [102] [106].

We convert the original pansharpened imagery to grey scale in order more closely follow the workflow of a typical image analyst who looks almost exclusively at *panchromatic* imagery when locating targets. Panchromatic imagery is presented as a single grey scale value representing the response measured over a wide range of frequencies including visible light. Panchromatic imagery is directly analogous to a black and white photograph.

A publicly available point-of-interest database containing approximate latitude and longitude coordinates for all McDonald's restaurants in the continental United States was used in conjunction with the Mapquest static maps API to extract 2000 by 2000 pixel PNG format images centered on the given lat/lon position at "zoom level" 18. Zoom level is an arbitrary measure used by

MapQuest that is somewhat ambiguous depending on the location on the globe. Level 18 is the highest zoom level available. The coordinates for the McDonald's were approximate, therefore, a human observer verified the actual location of the restaurant if it was visible in the image.

To simplify the task somewhat, we decided to use only McDonald's restaurants that have a "classic" appearance—a rectangular building with a red trimmed roof surrounded by a drive-through and parking—. Not every restaurant follows this design, and the exceptions come in all shapes and sizes. Therefore, the human observer was instructed to select only those that appeared to follow the classical design, up to 3000 examples. Eight targets were rejected because of incomplete image tiles near the restaurant leaving 2992 unique McDonald's in the data set. It was from these that training and testing data were derived.

Using the 2992 target examples, 100 by 100 pixel patches were selected with the offset location chosen by the human observer as the center pixel. The observer was instructed to click on the center of the building itself. The 100 by 100 pixel patches were large enough to contain the entire restaurant along with a good portion of the surrounding drive-through and parking lot. These image patches were used as positive training examples.

Negative training examples were selected randomly from the region immediately around the restaurant, covering about 0.1 to 0.15 miles in radius. The size of the pixels varies slightly from lat/lon position to lat/lon position due to the ambiguous nature of the zoom levels. The negative patch examples were selected so that no portion of the McDonald's patch was included. The selection of negative patches from the vicinity of the restaurant ensures that the training examples given to the classifier are strong lures. For example, selecting negative patch examples from forested areas would yield a classifier that discriminated suburban scenes from forested scenes.

In many machine learning training scenarios, positive and negative training examples are balanced in a 1:1 ratio. However, in cases where the classifier will be applied by convolving a sliding window over a larger image, as is the case for these experiments, the classifier will see a very different ratio of negative to positive examples, easily 10,000:1 or greater. Because of this, it is often helpful to change the training ratio to emphasize the negative examples. Ideally, the

positive-to-negative ratio should reflect the priors in actual usage.

This has the effect of offering the classifier many more instances of things it might encounter in practice that are similar to but different from the positive examples. For these experiments, it means the classifier is more likely to learn important differences between a McDonald's and other buildings that are often found nearby. Therefore, a ratio of 10:1 negative to positive training examples was used. The larger the negative to positive training example ratio is the longer training will take to converge so 10:1 was felt to be a good compromise and is based on previous experience in training deep neural networks to detect objects in satellite imagery.

Modern deep convolutional neural networks are typically trained with tens of thousands to millions of training examples. This data set included 2992 McDonald's. Therefore, data augmentation was used to increase the number of positive training examples. No preprocessing beyond data augmentation was performed. Each positive training patch was augmented nine times so that ten versions of the patch were present in the training data set giving a total of 29,920 positive training examples. Augmentation consisted of a randomly selected horizontal or vertical flip followed by a randomly selected positive or negative 90 degree rotation.

Data augmentation is a common practice in modern machine learning. The set of possible augmentations includes things like translational jitter, arbitrary rotations, etc. We chose not to use translational jitter because our application uses a sliding window which will move over an input image and eventually localize any targets in the center. We did not use arbitrary rotations to avoid introducing any image artifacts from regions that may not have actual data and because flips and 90 degree rotations, along with the naturally random orientation of the McDonald's in satellite imagery, covers, statistically, the majority of the orientations that will be encountered in practice.

In order to have the desired 10:1 negative to positive training example ratio, 100 randomly selected nontarget patches were extracted from the 2000 by 2000 pixel images centered on each McDonald's training patch ensuring that there was no overlap with the positive patch. This generated a negative training set of 299,200 unique patches representing the context in which McDonald's restaurants are typically found.

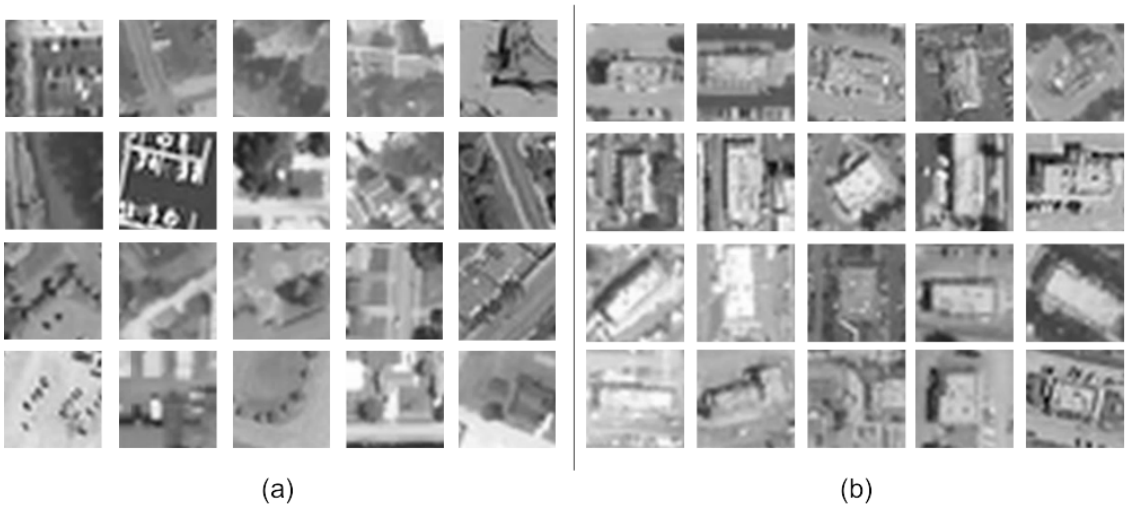


Figure 4.4: Representative (a) negative examples and (b) McDonald's restaurants.

The data set—both positive and negative examples—was partitioned into three disjoint subsets: one for training, one for validation during training, and a final set of test data used to generate ROC curves and other statistics. The 10:1 ratio was maintained for each set. The training set included 266,588 patches, positive and negative. The validation set included 29,620 patches and the test set 32,912 patches. All augmented versions of an example were restricted to the same set—training, validation, testing—that the original source example was placed into so that no mixing of examples was allowed. Additionally, the 10:1 negative to positive example ratio was maintained for each of the three subsets.

Figure 4.4 shows representative examples of the positive and negative training examples used in the experiments. Notice that the imagery is of relatively low quality. This is partly due to the zoom level and partly due to the fact that the imagery is pansharpened.

4.5 Implementation: Training Classifiers

The experiments in Chapter 5 depend critically on having a well-trained and implemented classifier. In this section we detail the development of the classifiers evaluated for the satellite experiments.

Several popular toolkits are available for training and using convolutional networks. These include Theano [21], Torch [44], and Caffe [97]. The classifiers in this chapter used the Caffe toolkit. Caffe uses a text file to specify the architecture of the network to be trained. Additionally, Caffe, among other options, reads training, validation, and testing data from text files which specify the full pathname of the input training image followed by an integer class label. The classifiers used label 0 for non-McDonald's examples and label 1 for McDonald's.

An annotated example of a simple convolutional network specified in Caffe is given in Figure 4.5. This network illustrates loading data into Caffe for training, the selection and definition of convolutional and pooling layers along with a top fully connected layer. The loss function is a softmax followed by a logistic regression which is the de facto standard for deep machine learning with images.

A softmax output takes the top level inner product layer vector (the output of each node) and maps it to a probability so if the output of the top level inner product, which has two nodes, is defined to be the vector x with weight vector θ, then the softmax is,

$$h(x) = \frac{1}{e^{\theta_0^T x} + e^{\theta_1^T x}} \begin{pmatrix} e^{\theta_0^T x} \\ e^{\theta_1^T x} \end{pmatrix}$$

which can be interpreted as an output probability of class membership. Note, the classifier in Figure 4.5 is a binary classifier even though, following Caffe convention, two top level outputs are defined. The softmax above is equivalent to logistic regression with a single parameter so that if the probability of class 0 is ϕ then the probability of class 1 is $1 - \phi$.

Caffe makes use of a solver file to define the environment in which training takes place. It specifies important parameters such as the number of minibatches to train, how often to store intermediate models, and the base learning rate. An annotated solver file is shown in Figure 4.6.

The solver file in Figure 4.6 uses a validation file to show the performance of the model on non-training data as training proceeds. This file is optional and Caffe does not take its output into account. Regardless, the samples in the validation file should not and were not used in any analysis

```
1  layers {                  ← data input layer
2    name: "data"
3    type: IMAGE_DATA
4    top: "data"
5    top: "label"
6    image_data_param {
7      source: "train.txt"   ← source training images
8      batch_size: 128
9    }
10   transform_param {       ← data transforms, here scaling [0, 1]
11     scale: .00390625
12   }
13 }
14 layers {                  ← first convolutional layer
15   name: "conv0"
16   type: CONVOLUTION
17   bottom: "data"
18   top: "conv0"
19   convolution_param {
20     num_output: 35        ← learn 35 output filters
21     kernel_size: 7        ← each 7 by 7 pixels
22     stride: 1             ← and step across the input one pixel at a time
23     weight_filler {
24       type: "xavier"      ← select a weight initialization scheme
25     }
26     bias_filler {
27       type: "constant"    ← but initialize the bias term to zero
28       value: 0
29     }
30   }
31 }
32 layers {                  ← first pooling layer
33   name: "pool0"
34   type: POOLING
35   bottom: "conv0"
36   top: "pool0"
37   pooling_param {
38     pool: MAX             ← use max pooling
39     kernel_size: 2        ← with a 2 by 2 kernel
40     stride: 2             ← stepping across by two pixels
41   }
42 }                         ← a full implementation would have additional convolution and pooling layers
43 layers {                  ← first fully connected layer
44   name: "ip0"
45   type: INNER_PRODUCT
46   bottom: "pool0"
47   top: "ip0"
48   inner_product_param {
49     num_output: 500       ← with 500 nodes
50     weight_filler {
51       type: "xavier"
52     }
53     bias_filler {
54       type: "constant"
55       value: 0
56     }
57   }
58 }
59 layers {
60   name: "relu0"           ← and a rectified linear nonlinearity
61   type: RELU
62   bottom: "ip0"
63   top: "ip0"
64 }
65 layers {
66   name: "drop0"           ← and dropout at 50%
67   type: DROPOUT
68   bottom: "ip0"
69   top: "ip0"
70   dropout_param {
71     dropout_ratio: 0.5
72   }
73 }
74 layers {                  ← final output layer
75   name: "ip1"
76   type: INNER_PRODUCT
77   bottom: "ip0"
78   top: "ip1"
79   inner_product_param {
80     num_output: 2
81     weight_filler {
82       type: "xavier"
83     }
84     bias_filler {
85       type: "constant"
86       value: 0
87     }
88   }
89 }
90 layers {
91   name: "loss"
92   type: SOFTMAX_LOSS      ← using a softmax multinomial logistic loss function
93   bottom: "ip1"
94   bottom: "label"
95 }
```

Figure 4.5: A Caffe convolutional neural network definition file. Line numbers added.

```
1  train_net: "train.prototxt"     ← definition of the network for training
2  test_net: "validate.prototxt"   ← a validation network run during training
3  test_iter: 1                    ← use one pass of validation
4  test_interval: 100              ← test every 100 minibatches
5  base_lr: 0.01                   ← base learning rate η
6  momentum: 0.9                   ← use momentum as well
7  weight_decay: 0.0005            ← and weight decay
8  lr_policy: "inv"                ← learning rate update schedule
9  gamma: 0.029                    ← learning rate update parameter
10 power: 0.75                     ← learning rate update parameter
11 display: 100                    ← show progress every 100 minibatches
12 max_iter: 1000000               ← run this many minibatches
13 snapshot: 10000                 ← store a model every 10,000 minibatches
14 snapshot_prefix: "model"        ← model file name prefix
15 solver_mode: GPU                ← use a GPU to finish before the universe dies
16 device_id: 0                    ← which GPU
```

Figure 4.6: A Caffe solver file. Line numbers added.

of the final model. Caffe simply runs for the specified number of minibatches and then stops. A minibatch is a random selection of training inputs and it is the average error over the minibatch that is used in backpropagation to update the model weights. Typical minibatch sizes are on the order of 100 samples. This is in contrast to classical neural networks which use a pass through all the input training data. It is a concession to runtime efficiency and is the "stochastic" part of "stochastic gradient descent" which is the optimization technique used by Caffe and other deep neural network toolkits. When Caffe completes its training the output will be a final model which stores the weights and biases for the connections implied by the specification in the `train.prototxt` file. This file typically has an extension of `.caffemodel` and it is the combination of the `.prototxt` and `.caffemodel` files that fully specifies a trained network.

4.6 Implementation: Classifying Test Images

A fully trained network can be run against test data in order to characterize its performance. The test samples are stored in a file in the same format as the training samples with a fully qualified pathname followed by a class label 0 or 1. This allows a simple command line to execute the test,

```
$caffe test -model=test.prototxt -weights=model.caffemodel -iterations=1 -gpu=0
```

where `test.prototxt` is a file very similar to the training file listed in Figure 4.5 but referencing the test data instead. The trained weights are in `model.caffemodel`. The `-iterations=1` part of the command makes a single pass through the test data.

Out of the box, Caffe does not include a layer to do anything other than give the overall accuracy of the network on the test data. This was insufficient for our purposes so Caffe was extended with a new layer written in C++ to output, for each input test sample, the actual class label along with the softmax probability for each class defined. The new layer, called `STATS`, is specified as,

```
layers {
  name: "stats"
  type: STATS
```

```
  bottom: "prob"
  bottom: "label"
  top: "stats"
  stats_param {
    output_file: "test.prob"
  }
}
```

where it reads the top level softmax per class probabilities and writes them to disk for each input sample. With the network output and input class label in the `.prob` file it was possible to calculate the ROC curve and other important metrics.

Chapter 5

Experiments with Satellite Imagery

The results of the synthetic image experiments described in Chapter 3 are encouraging: soft highlighting aids subjects in detecting targets compared to no highlighting or hard highlighting. However, the displays in Chapter 3 are synthetic and unlike the contexts in which soft highlighting might be put in practice, e.g., satellite or medical image analysis. In Chapter 5, we extend the results of Chapter 3 to the domain of satellite imagery. We hypothesized that the improvements in target detection observed with soft highlighting of synthetic images would extend to real-world images highlighted with the output of an actual machine learning classifier trained to detect a challenging target in satellite images. Chapter 4 described the classifier and the domain in more detail.

5.1 Highlighting Satellite Images

Each of the experiments in this chapter presents subjects with satellite images in one of three conditions: control, soft highlighting, or hard highlighting. The control condition used no highlighting and consisted of a plain grey scale image. Soft and hard highlighting are described in detail below. Figure 5.1 shows one of the test images in each of these possible states with the control condition on the left, soft highlighting in the middle and hard highlighting on the right.

Section 5.1.1 describes the process by which a trained convolutional neural network was used to generate *heat maps* representing classifier output probability at every location in a satellite image. Sections 5.1.2 and 5.1.3 describe how we use the heat map to perform soft and hard highlighting

Figure 5.1: An example of the three highlighting conditions used in the experiments of Chapter 5. Left: control condition, no highlighting. Center: soft highlighting. Right: hard highlighting.

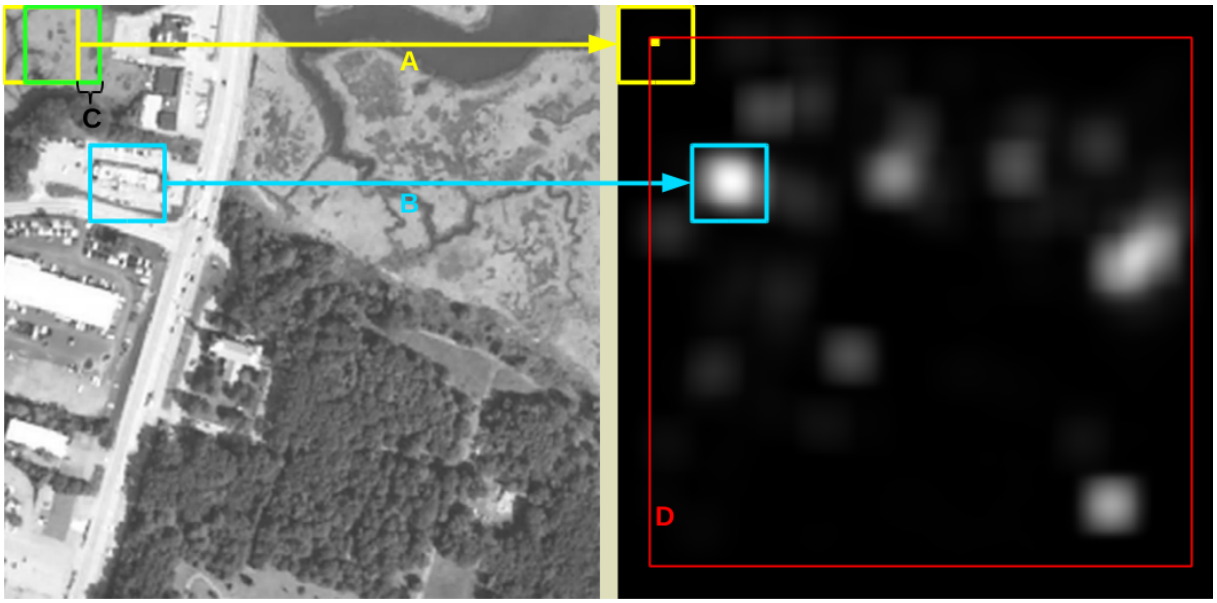
on satellite images, respectively.

5.1.1 Creating a Heat Map

The soft and hard highlights are based on the output of a classifier that analyzes each region of an image. This output is represented as a probability map or *heat map* and is a 2D array of numbers representing the classifier's confidence that the target is present at each pixel in the image. The algorithm used a sliding window of a fixed size and convolved it over the input image.

The heat map generation process is illustrated in Figure 5.2. The input image is on the left. Classification takes a small sliding window (A in the figure) and convolves it over the input image. Each window is presented at the input of the classifier to generate an output probability value. The heat map (right) is built up by filling the center values of the corresponding location in the array to the output probability. The sliding window is then moved by a step size (C) until the sliding window has completely covered the input image.

The half window sized border at the edge of the image (as indicated in Figure 5.2) was excluded to avoid edge effects in the convolution. The raw heat map was smoothed slightly with a Gaussian filter ($\sigma = 50$ pixels). This smoothing is considered part of the heat map generation process and was used for both soft and hard highlighting. Figure 5.3 shows the effect of this smoothing on a heat map.



A Mapping between image patch and heat map output
B A patch with the target present
C Step size between patches
D Actual heat map area (half patch size border)

Figure 5.2: Heat map generation from an input image. The input image, left, is classified by applying a sliding window (A) over the image. The output probability that a target is located in the current sliding window is used to assign the center pixels of the corresponding location in the heat map (A, right). The size of the center region is determined by the sliding window step size (C). In (B, left) the sliding window is over the target which produces a large response in the heat map (B, right) corresponding to a high probability that a target is present. (D, left) shows the region of the heat map with valid data, edges were ignored. N.B. the heat map, as shown on the left, was processed with a subtle Gaussian filter to smooth the response spatially. See Figure 5.3.

Figure 5.3: A heat map, as generated from the process in Figure 5.2, before smoothing (left) and after smoothing with a Gaussian filter of width 50 pixels (right). Smoothing added to the continuous nature of the soft highlighting approach at the expense of possible information such as the partial outline of the target in the upper left of the heat map (bright region).

5.1.2 Soft Highlighting of Satellite Images

The hallmark of soft highlighting is a continuous presentation or variation of the image highlight. We experimented with several different soft highlighting techniques and chose one, to be described next, that was visually salient yet had minimal degradation of image features.

The source satellite image is a single grey scale band. The highlighted image is a new three band image in HLS (hue-lightness-saturation) color space [98]. In HLS, H represents the hue or position around the color wheel, L represents the lightness or intensity, and S corresponds to the fullness of the color represented by the hue value. Our algorithm fixes the hue to a desired color, here red, sets the lightness to the single-band grey value of the image, and sets the saturation to the heat map value. Finally, the HLS image is mapped back to RGB color space for display.

The saturation-adjustment algorithm preserves the contrast inherent in the source image. It simply enhances the selected hue based on the confidence of the classifier at each location. This was deemed to be better than simple alpha-blending [166] which sets the output pixel to a weighted average of the input pixels using a single parameter, α,

$$\texttt{out} = \alpha\ \texttt{red} + (1 - \alpha)\ \texttt{image}$$

where `image` is the satellite image, `red` is a constant red image and α is the heat map value. By its nature, alpha-blending obscures the image as seen in Figure 5.4 right. We were interested in informing the viewer while preserving original image content which is what HLS soft highlighting does (Figure 5.4, left).

Figure 5.5 shows a continual gradation of soft highlighting from left to right corresponding to a highlight intensity (saturation) of zero on the left and one on the right. An intensity of 0.5 is indicated in the middle. The example image contains a target McDonald's (circled) with a highlight intensity of approximately 0.25.

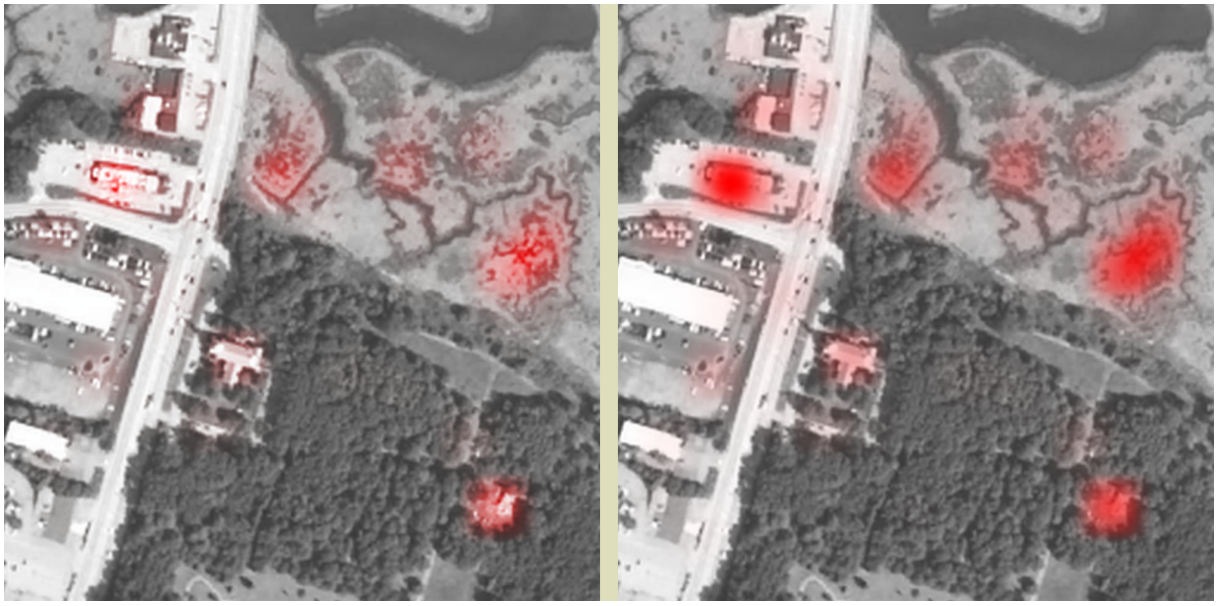


Figure 5.4: A comparison of soft highlighting with saturation adjustment (left) and alpha-blending (right). Saturation adjustment preserves contrast in the image whereas blending obscures potentially important visual features.

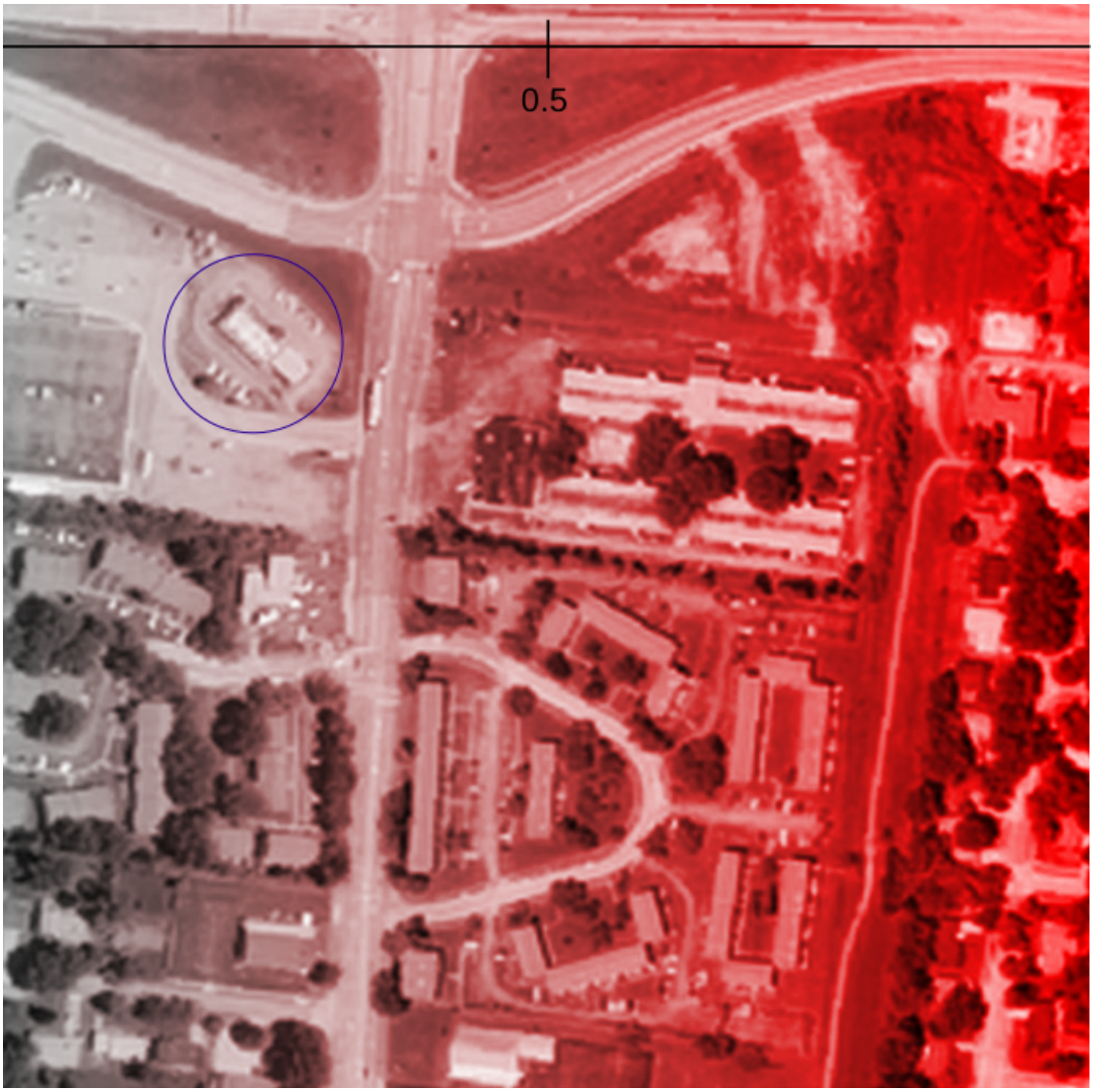


Figure 5.5: An example of soft highlighting as a continuous gradation, left to right, varying the highlight intensity from zero (left) to one (right). A highlight intensity of 0.5 is marked. A target McDonald's is present in this image (circled) and corresponds to a highlight intensity of approximately 0.25.

5.1.3 Hard Highlighting of Satellite Images

Soft highlighting relies on continuous variation whereas hard highlighting makes a decision as to where to place markers in the image based on the classifier output. In the experiments of this chapter the marker is always a red square.

Our marker placement algorithm, while there are still heat map values above the cut off threshold, repeats the following to locate the position of the next marker:

(1) Locate the largest value in the heat map.

(2) Locate the position of this maximum value and if no markers exist within a fixed radius (one half the marker width) create a new marker centered on the maximum position.

(3) Set all the heat map values within the new marker to zero. This prevents selecting the same maximum value location a second time.

(4) Repeat from Step 1 until no new markers can be created. This happens when the maximum heat map value remaining falls below the threshold (0.5) or a maximum number of markers has been created.

For various thresholds, we generated a histogram over the set of images of of the number of markers (hard highlights) that would be placed in an image (Figure 5.6). Based on the heuristic that we wanted to obtain a half dozen markers per image, we chose a threshold of 0.5. The threshold of 0.5, as compared to 0.6, has fewer images with only few highlights. The threshold of 0.5, as compared to 0.4, has fewer images with over 6 highlights. With a threshold of 0.5, every image had at least 2 markers present. The heuristic was chosen to mimic a target recognition system which would strive to limit the number of markers typically shown.

5.2 Experiment 6: Locating a Single Target in a Satellite Image

In Experiment 2 (Section 3.2) subjects used synthetic images and we timed how long they took to locate a single target digit ("2") in a grid of handwritten digits. In Experiment 6 we

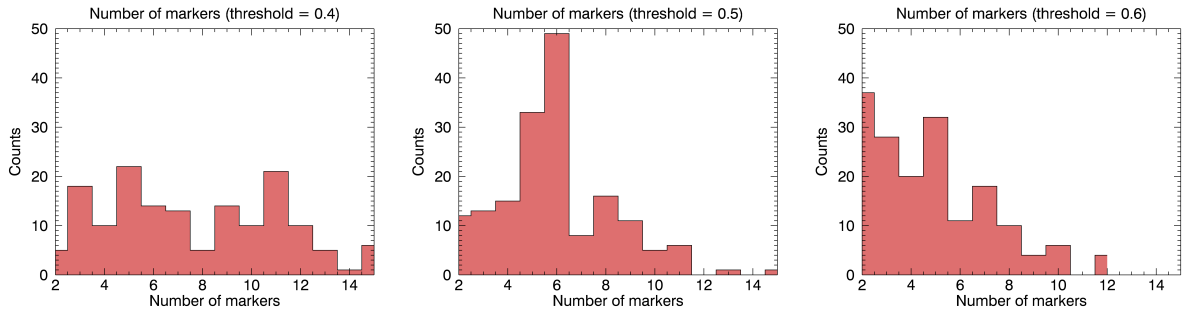


Figure 5.6: Histogram showing the distribution of the number of markers for different threshold cutoffs. The threshold cutoff was the smallest probability value to be considered for a marker. Right: threshold of 0.4. Center: threshold of 0.5. Right: threshold of 0.6.

moved from synthetic images to satellite images and changed our search target from a digit token to a particular type of structure—a McDonald's restaurant. This experiment explores whether the effects seen in Experiments 1-5 using synthetic images persist when images that are more natural are used.

Subjects viewed grey scale satellite images from mapquest.com to locate the single McDonald's restaurant present in every image. Three experimental conditions were studied: soft highlighting, hard highlighting, and a control with no highlighting. The highlights on the images, both soft and hard, were generated from the output of the `mcdonalds3` classifier described in Section 4.1 and applied as outlined in Section 5.1.

5.2.1 Methods

The experiment was run on Amazon's Mechanical Turk and all subjects ($n = 84$) were screened to be from either the United States or Canada. We chose this population because the satellite images were taken from the same region, and although subjects had no prior experience searching satellite imagery for restaurants, matching the subjects to the geographic region of the images at least ensured that they would be familiar with the type of suburban environments in which the restaurants are typically located. As North Americans, the subjects had a good idea that McDonald's are situated on main streets, often at corners, that they have drive throughs

and a parking lot, etc. Our hope was to remove some inter-subject variability via the geographic restriction on our subject pool. All subjects who actually completed the experiment were from the United States as determined by examination of their IP addresses as provided by Amazon. Subjects were told that their IP address would be used for that purpose before agreeing to participate. Future experiments were able to use Amazon's own filtering system so that checking IP addresses became unnecessary, only subjects from the United States were able to select the experiment.

Of 289 subjects who started the experiment, 84 completed it successfully. Subjects were free to leave the experiment voluntarily at any time. And we rejected any subject whose defocused their browser window at the start of a trial; we had this requirement because we were measuring reaction times and Mechanical Turk subjects have the tendency to multitask. Subjects were told prior to the beginning of the experiment that changing window focus would result in the termination of the experiment. Of the 205 subjects who did not complete the experiment, 105 of them were rejected for changing window focus. Subjects were only permitted to perform the experiment once, whether they completed it or not.

Subjects were shown a representative satellite image and 20 small example images of a McDonald's restaurant in order to give them a sense of the variation within the target. Subjects were then given these specific directions:

> Each image will contain exactly one McDonald's restaurant. Simply locate the restaurant and click on it with the mouse to move to the next image. There are 34 images in the experiment and you must locate the McDonald's in each image in order to be paid.
>
> Although there is no limit on time or the number of clicks you can make, avoid unnecessary clicks. If you click on a location other than the McDonald's, you will hear a buzzing sound to remind you to click carefully.
>
> During this experiment, you may not cause the browser window to defocus (e.g,. by clicking another window, tab or taskbar). If you defocus the browser window, the HIT will end immediately and you will not be paid.

The images used in the experiment were selected from a set of 245 images centered on a McDonald's and surrounding region (2000 x 2000 pixels) that were in a held-out test set not used for classifier training. A randomly placed subset of 1000 x 1000 pixels was pulled from the image

to locate the target somewhere within but not centered. The target was completely visible in the image. This image was classified as in Section 5.1.1 and the resulting image and heat map were subset again to remove the half sliding window size border that was unclassified. The resulting 900 x 900 pixel image and heat map were resized to 600 x 600 pixels to fit entirely within the subject's browser window.

The experiment was run inside of the subject's browser within the environment of Amazon's Mechanical Turk using the external question interface. Images were pre-loaded at the start of the experiment; subjects saw an animated count of trial images as they were loaded.

Each subject performed 34 trials separated into two blocks of 17 trials each. On each trial, subjects had to click on the McDonald's target with the computer mouse. The first two trials of each block were treated as practice and were not included in the analysis, leaving 15 trials per block (condition).

At the start of each block, subjects received instructions specific to the block. Each block involved a different condition of the experiment, as detailed below. The specific instruction text at the start of each block was:

- Control:

  > You will now see a series of 17 black and white images which look similar to this example: .... Remember, your task is to click on the McDonald's location as rapidly as possible and with as few errors as possible. You will receive feedback if you click on an incorrect location.

- Soft:

  > You will now see a series of 17 black and white images that have been shaded red where the computer believes a restaurant may be. The computer is not perfect but it can provide assistance in locating the McDonald's. The images look similar this example: .... Remember, your task is to click on the McDonald's location as rapidly as possible and with as few errors as possible. You will receive feedback if you click on an incorrect location.

- Hard:

  > You will now see a series of 17 black and white images that have red boxes where the computer believes a restaurant may be. The actual restaurant may be outside any of the boxes. The computer is not perfect but it can provide assistance in locating the McDonald's. The images look similar this example:

> .... Remember, your task is to click on the McDonald's location as rapidly as possible and with as few errors as possible. You will receive feedback if you click on an incorrect location.

Subjects were counterbalanced in groups of four. For each set of four subjects 34 unique images were selected from the repository of test images. Images were used only once from the repository. The 34 unique images were used to generate the sequence of images seen by each of the four subjects. The order of the images within the sequence was not changed subject to subject. However, the experimental conditions for each subject followed the pattern:

	Block 1	*Block 2*
Subject 0	soft highlighting	control
Subject 1	control	soft highlighting
Subject 2	hard highlighting	control
Subject 3	control	hard highlighting

The experiment was run counterbalanced in groups of four subjects for a total of 56 subjects.

After the first 56 subjects we ran two sets of 14 subjects each, with a new sequence of test images each time, counter-balanced, using these conditions:

	Block 1	*Block 2*
Subject 0	soft highlighting	hard highlighting
Subject 1	hard highlighting	soft highlighting

for a total of 84 subjects in the experiment. Note, the sequence of images for the control-soft and control-hard subjects was different than the sequence for the soft-hard subjects. This is unfortunate. It would have been better if the same sequence of images were used for both sets of subjects. Additionally, the pool of subjects between the control-soft/control-hard runs and the soft-hard runs were different. Given that the subjects were paid adults selected from Mechanical Turk there is no reason to believe there was any significant difference between the sets.

The subject was free to click on the image at will, as many times as desired. If the click was not within 30 pixels of the target McDonald's a buzzing sound was played to offer immediate feedback. If the click was within 30 pixels of the target McDonald's the trial ended and a pleasant ding sound was played. The location and time from the start of the trial was recorded for each mouse click. Once the target was found the next image in the block was presented immediately

after the subject clicked the "Next" button. At the beginning of each block instructions were given as to the type of highlighting that would be present on the images for that block. When both blocks were completed subjects clicked a "Submit" button to send the experiment results to the server and to signal Amazon that they successfully completed the task. Subjects were paid $1.25 for completing the task which typically took between 10 and 20 minutes.

5.2.2 Results

The fraction of targets found by time and condition is shown in Figure 5.7. Each curve represents one condition: control, soft, and hard. Each curve is monotonically increasing because a target is more likely to be found over time. The steepest curve is for soft highlighting, indicating that subjects were fastest to find targets with soft highlighting; next steepest is the hard highlighting condition; and the control condition is the slowest. The curves do not cross over, indicating that the ranking of conditions persists across time. In the Figure, vertical lines are drawn to indicate the mean latency to locate a target. These means indicate that soft highlighting enabled subjects to locate targets more quickly than either hard highlighting or the control condition of no highlighting. This experiment thus finds results similar those obtained in Experiments 1-5, except that the current experiment uses naturalistic stimuli mirroring the effects seen in Experiments 1-5.

Table 5.1 shows both (parametric) t-tests and a (nonparametric) two-sided Wilcoxon signed-rank test comparing the three conditions. The first set of tests excludes the practice trials; the second set of test includes the practice trials. Results of paired comparisons were the same whether or not the practice trials were included. Both the t-tests and Wilcoxon tests show that soft beats hard and control, The t-test indicates that hard beats control, but the weaker Wilcoxon test does not reach significance. Note that no correction of significance levels was done for multiple comparisons.

The experiment allowed subjects to click freely on the image while searching for the target. It is useful to examine the distribution of the number of nontarget clicks as a way to investigate how much searching subjects were doing before finding the target. In Figure 5.8 we show a histogram of the number of nontarget clicks made in a trial by condition. From this histogram it is clear that soft

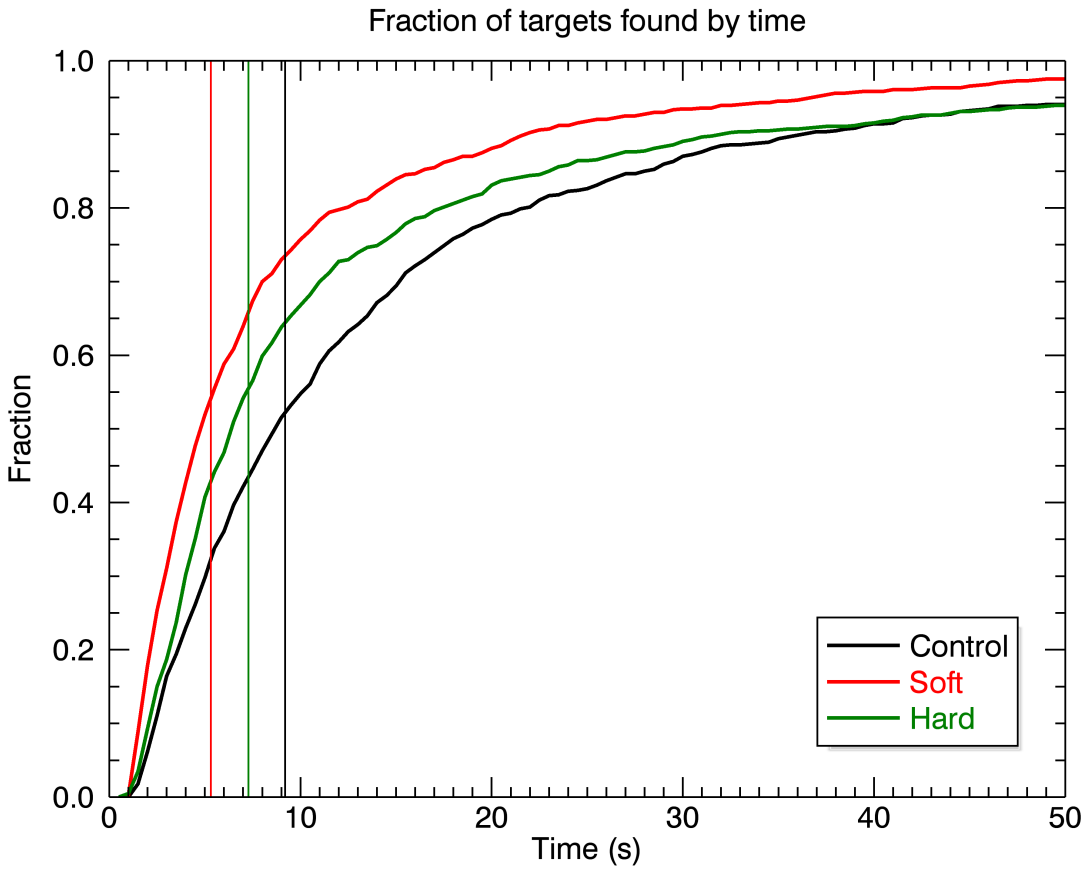


Figure 5.7: Experiment 6. Fraction of targets found by time and condition. The vertical lines are the mean latency to locate the target across each condition.

```
(excluding practice trials)
   control-soft: 9.6504 vs 5.3770, t(83)= 5.3966 (p= 0.0000), w(83)=1069.0 (p= 0.0000)
   control-hard: 8.7385 vs 7.1788, t(83)= 2.2143 (p= 0.0354), w(83)= 891.0 (p= 0.1275)
   soft-hard   : 5.2405 vs 7.3710, t(83)=-2.7299 (p= 0.0110), w(83)= 637.0 (p= 0.0083)

(including practice trials)
   control-soft: 9.7233 vs 5.5467, t(83)= 5.1625 (p= 0.0000), w(83)=1066.0 (p= 0.0000)
   control-hard: 9.1843 vs 7.5802, t(83)= 2.0213 (p= 0.0533), w(83)= 886.0 (p= 0.1493)
   soft-hard   : 5.6603 vs 7.3718, t(83)=-2.1320 (p= 0.0423), w(83)= 651.0 (p= 0.0160)
```

Table 5.1: Experiment 6. Mean latency times along with paired t-test scores and p-values followed by those of the two-sided Wilcoxon signed-rank test. The top scores exclude the first two trials of each block as practice while the bottom scores use all trials. There are no meaningful differences between the two set of scores. Both the t-tests and Wilcoxon tests show that soft beats hard and control.

highlighting leads to fewer overall clicks on nontarget locations relative to hard highlighting and the control condition with most soft targets found within two clicks. Somewhat surprisingly, the distribution of clicks for hard highlighting is somewhat spread out though the majority of targets were found within three clicks. As expected, the distribution of clicks in the control condition was even more variable.

The mean number of nontarget clicks ($\pm$ SE) was 7.6345 ± 0.9356 in the control condition, 6.1143 ± 0.8371 in the hard highlight condition, and 3.6060 ± 0.3078 in the soft highlighting condition. Unpaired t-tests on the log of the number of nontarget clicks shows that soft highlighting leads to significantly fewer clicks than either hard ($t(83) = 2.7475$, $p = 0.0070$, Cohen's $d = 0.5315$) or control ($t(83) = 3.8751$, $p = 0.0002$, Cohen's $d = 0.7730$) conditions. However, there is no significant difference between hard highlighting or control ($t(83) = 1.1886$, $p = 0.2372$, Cohen's $d = 0.2288$).

Figure 5.9 shows the data broken down in terms of pairs of highlighting conditions: control-hard, control-soft, hard-soft. Each subject was tested on one of these three pairs and in a particular order, e.g., control in block 1 and hard in block 2, or hard in block 1 and control in block 2. Each bar represents the time to locate a target (in seconds) for the subset of subjects tested in a given condition on a given block. A single subject's data contributed to two bars in one of the three histograms. We observe the same qualitative pattern of performance regardless of the block. That is, the ordering of performance across conditions does not depend on whether subjects are unpracticed or practiced. An ANOVA was run using two factors: highlighting condition (control, soft, hard) and trial block (1 versus 2). Because of the confound between blocks and highlighting conditions, we could not run an ANOVA to look at the block X condition interaction. However, we were still able to examine the main effects of (pairwise) condition and block. The results of the ANOVA are presented in Table 5.2. These results indicate that there is a statistically significant block effect between control and soft highlighting. A nearly significant effect is seen between control and hard highlighting. Lastly, a marginally significant result is found between soft and hard highlighting. These results are weaker than our overall results because each pairwise test

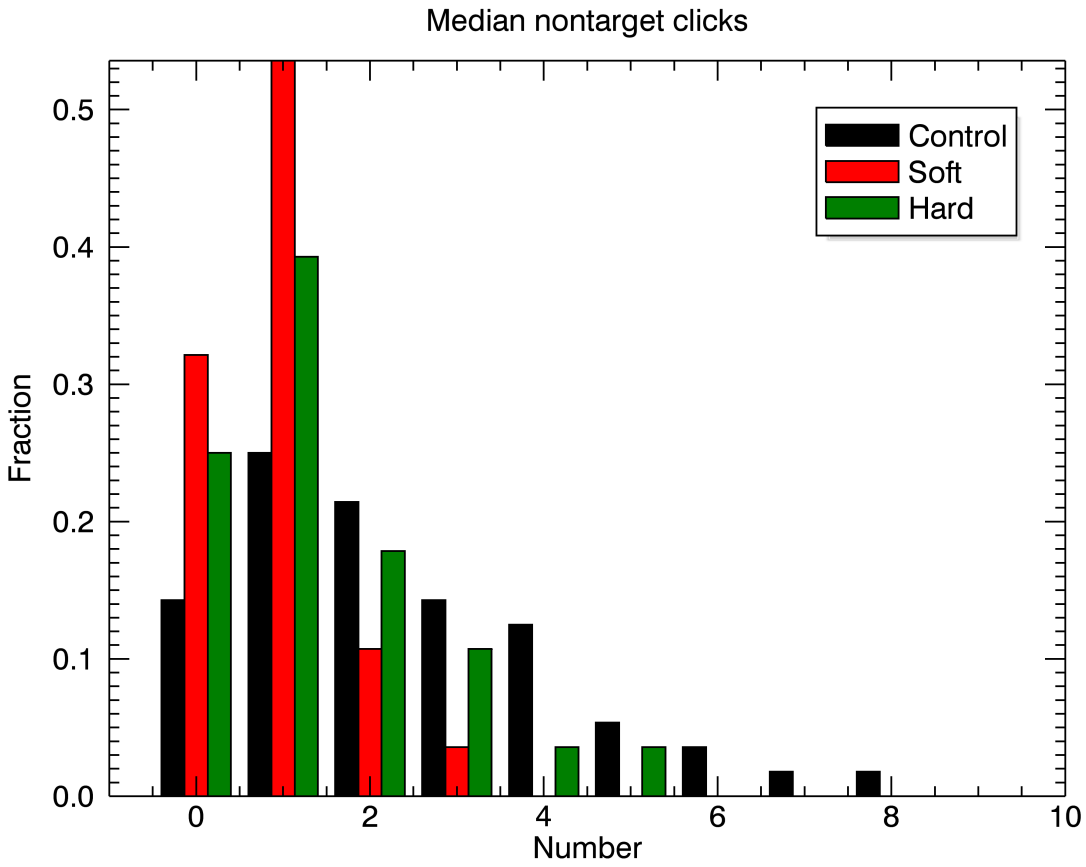


Figure 5.8: Experiment 6. Histogram of the number of nontarget clicks by condition. This can be viewed as a measure of how difficult subjects found the location task with more clicks indicating more attempts before locating the target. Bins beyond 10 clicks are not shown as very few counts were present.

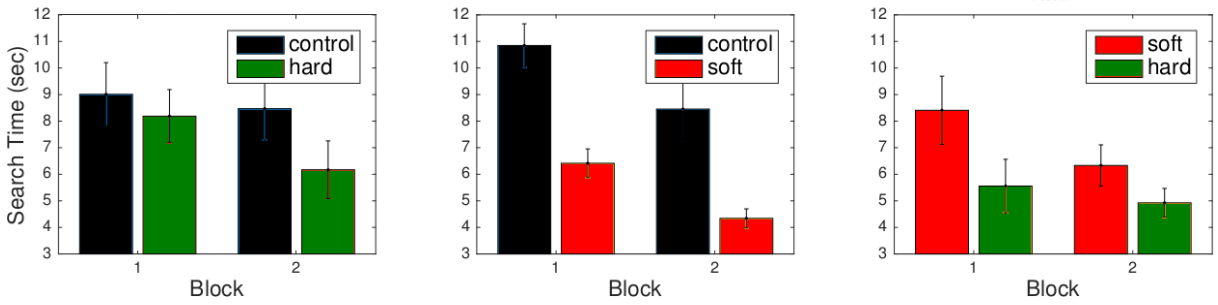


Figure 5.9: Experiment 6. Plots showing the mean time to locate a target by paired conditions: control versus hard, control versus soft, and soft versus hard.

uses only half of the subjects that were included in the overall tests. The advantage of looking at the data in the way they are presented in the Figure is that one can track the within-subject performance across blocks.

We were interested on the relationship between a subject's reaction time (RT) and the heat map probability at the target location. Presumably, if a form of highlighting is effective, then subjects should be faster if the heat map probability is higher. Figure 5.10 shows a scatter plot of individual trial RTs versus the heat map probability at the target location. The points are coded by condition (control is blue, hard is green, and soft is red). (Note that for this plot only the control condition is represented in blue instead of black to make it easier to see through the points in the plot.) The heat map probability represented is the median heat map value in a small region centered on the target McDonald's location in the image.

Regression lines are also shown in the scatter plot. For the soft and hard highlighting conditions, the negative slope of the line indicates a faster reaction time for more highlighted targets (soft condition) based on the classifier output. The negative slope for hard highlighted targets in the case were the classifier output is also higher (not reflected in the hard highlight) might be an indication of the relative ease of locating that target for both humans and the classifier. Because the control condition showed no information to indicate the heat map probabilities, a dependency between RT and heat map probability should be observed only if targets that subjects find easy to identify are also easy for the classifier to identify. Table 5.3 shows the results of paired t-tests

```
Control vs Hard: F(1,26)=   5.4, p = 0.0285
                 block effect for Hard experiment: F(1,26)=3.62, p = 0.0684

Control vs Soft: F(1,26) = 39.8, p = 1.1252e-6
                 block effect for Soft experiment: F(1,26)=10.9, p = 0.0028

Soft vs Hard   : F(1,26)=   8.1, p = 0.0086
                 block effect for Soft vs. Hard  : F(1,26)=3.26, p = 0.0826
```

Table 5.2: Experiment 6. ANOVA results comparing the mean time to locate a target. The block effect for control versus soft highlighting is clearly significant. The block effect for control versus hard highlighting is nearly significant while the effect for soft versus hard highlighting could be considered marginally significant.

between the conditions. The values compared are the mean slope across subject. Each pair was statistically significant indicating that there is an ordering which can be applied across the conditions with soft highlighting leading to the strongest relationship between RT and heat map value, then hard highlighting, and then the control condition. Highlighting helps locate the target in the image with soft highlighting being more effective than hard highlighting. Similar conclusions can be reached by computing the correlation coefficient for each of the three scatter plots: soft, hard, and control highlighting lead to coefficients of -0.43, -0.25, and -0.09, respectively.

The results of Experiment 6 show clearly that soft highlighting can lead to improved target localization (Figure 5.7, Figure 5.10) while reducing the number of false target clicks (Figure 5.9). Beyond the performance benefit of soft highlighting relative to hard highlighting, we obtained evidence that soft highlighting provides better information to subjects about the confidence level of a classifier. Thus, subjects are able to make use of the graded information presented by the soft highlighting scheme.

5.3 Experiment 7: Locating an Unknown Number of Targets (0-1) in a Satellite Image with Real-Time Feedback

As previous research has demonstrated [111, 226, 4], hard highlighting of targets in an image leads to an increase in misses for targets that are not highlighted. Experiment 6 found easier detection of targets with highlighting, but because every image contained a target and each trial

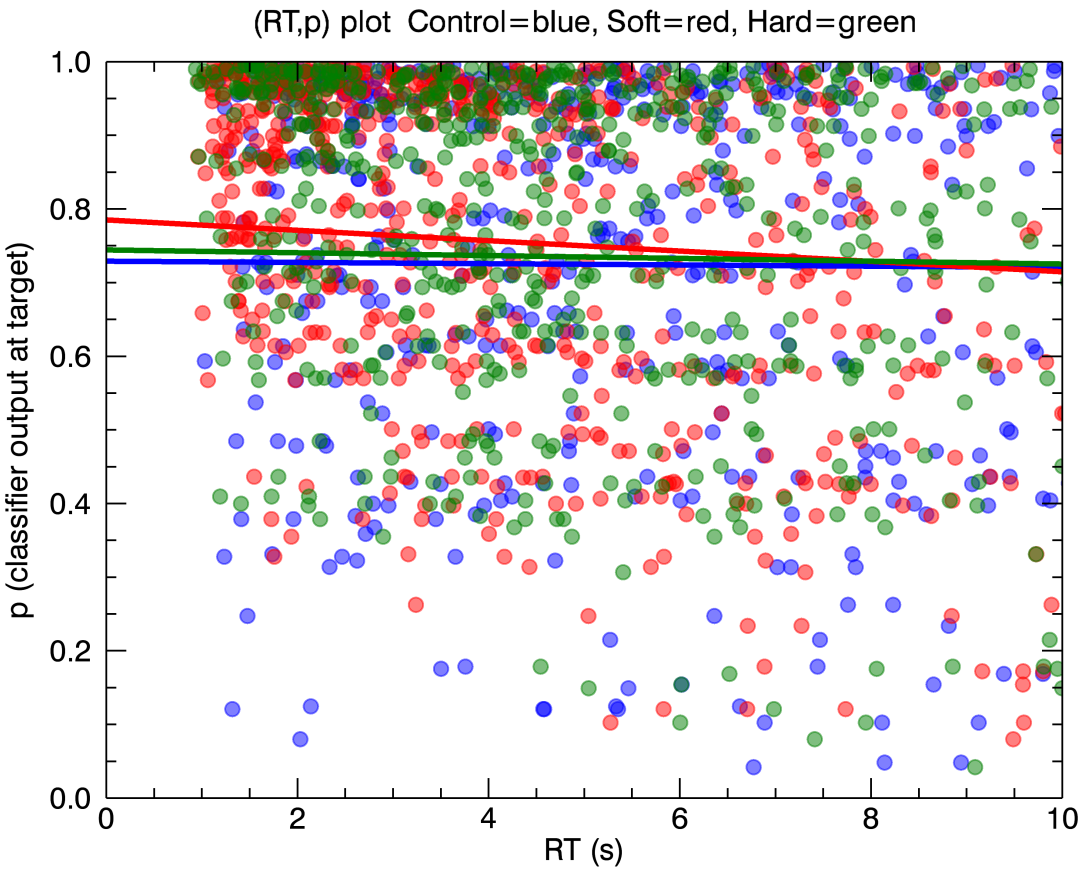


Figure 5.10: Experiment 6. A scatter plot of each trial reaction time (s) and the heat map probability at the target location, by condition. Note, in this plot the control condition is represented in blue instead of black to increase visibility. The probability is the median heat map value in a small region centered on the target location. The lines are best fit lines per condition to the points. The downward slope of the line indicates faster reaction time for strongly highlighted targets in the soft and hard highlighting case. The control condition did not actively indicate the heat map probabilities to the subject hence there is no expected association between the reaction time and the heat map probabilities.

```
control-soft: -0.0042 vs -0.0134 t(28)=  2.999 (p= 0.0041)
control-hard: -0.0018 vs -0.0096 t(28)=  3.292 (p= 0.0018)
soft-hard   : -0.0164 vs -0.0096 t(28)= -2.262 (p= 0.0277)
```

Table 5.3: Experiment 6. The results of paired t-tests on the per subject slopes fit to the (RT,p) data (Figure 5.10). These show that there are statistically significant differences between the slopes again indicating that subjects found soft highlighted targets faster than hard highlighted targets and faster still than unhighlighted images. N.B. the means for the two control conditions are different from each other, the first paired with subjects who also viewed soft highlighted images and the second paired with subjects who also viewed hard highlighted images.

continued until the subject found the target, Experiment 6 was not able to evaluate the effect of highlighting on *misses*. Experiment 7 sought to create a situation where it was possible for subjects to potentially miss a target because some trials contained no target, and subjects were allowed to quit a trial without having found the target. This experiment is analogous to Experiment 4 described in Section 3.4 except that Experiment 7 uses satellite imagery instead of synthetic imagery. We hypothesized that soft highlighting would not only lead to faster detection of targets that are present in the image, but also to fewer missed targets relative to the hard or control conditions.

Like Experiment 6, Experiment 7, using the same three highlighting conditions of control, soft, and hard, asked subjects to view grey scale satellite images to search for a single McDonald's restaurant. The output of the `mcdonalds3` classifier determined the highlights as in Experiment 6.

5.3.1 Methods

Nontarget images, those without a McDonald's, were selected to be nearby a target McDonald's in order to preserve the character of the environment contained in the image. To do this, the latitude and longitude of each McDonald's in the set of test images ($n = 245$) was offset by 0.0045 degrees and a new image was downloaded from the Mapquest server centered on that latitude and longitude. Because of the variation in the actual pixel size for a specified zoom level when downloading the images the offset corresponded to a shift of approximately 0.3 miles away from the McDonald's. This ensured that the target was not present in the nontarget image. Additionally, the set of nontarget images was manually screened to remove images that were, subjectively, deemed too obvious because they consisted mostly or solely of water or vegetation without man-made structures. This screening process left a final tally of 186 nontarget images for use in the experiment.

Experiment 7 was run on Amazon's Mechanical Turk and all subjects ($n = 90$) were required to be from the United States. Of 155 subjects who started the experiment, 90 completed it successfully. Subjects were free to leave the experiment voluntarily at any time. Subjects who changed

their browser window focus were dropped from the experiment. Of the 65 subjects who did not complete the experiment, 43 of them were rejected for changing window focus. Subjects were paid $1.25 for completing the experiment which typically took on the order of ten minutes.

The target images used in the experiment were identical to the set used in Experiment 6. The sequence of images and the conditions applied to them were different as described below but the source test images were the same. The nontarget images were passed to the `mcdonalds3` classifier to generate a heat map for the image. The classifier made mistakes so the resulting heat maps included false positive and false negative outputs as shown in Figure 5.11. Both the target and nontarget images were processed as in Experiment 6 to create the 600 x 600 pixel images presented to the subjects.

The experiment was run inside of the subject's browser within the environment of Amazon's Mechanical Turk.

Each subject completed 36 trials, with 24 target-present trials and 12 target-absent trials. A trial was complete when the McDonald's target was located or the subject clicked a "No McDonald's Present" button below the image to indicate his or her belief that there was no target present. Subjects who did not complete all 36 trials were discarded from the data; these subjects were not paid.

Each block had 9 images. The 9 images were grouped in 3 sets of triples. Within each triple, there was 1 target-absent and 2 target-present images. Each triple was assigned to a different highlighting condition—soft, hard, and control. The order of presentation of the 9 images was randomized within a block. The assignment to highlighting conditions was counterbalanced across subjects, such that for every 3 subjects, each triple appeared in each condition with the following assignments for the 3 triples:

Subject 1	S	S	S	H	H	H	C	C	C
Subject 2	H	H	H	C	C	C	S	S	S
Subject 3	C	C	C	S	S	S	H	H	H

with soft (`S`), hard (`H`), and control (`C`) indicated and representing an image viewed by the subject. This highlighting order was applied to each of the four blocks of images per subject so that each

Figure 5.11: Experiment 7. Example images generated from nontarget heat maps using hard (left) and soft (right) highlighting.

image used each highlighting condition exactly once among the three subjects in the set.

The total experiment consisted of 90 subjects, counterbalanced in sets of three to follow the counterbalancing pattern presented above. The counterbalancing is a Latin square design that ensures that each highlight condition appears equally often in the early, middle, and late portions of the experiment. For each of the 30 sets of 3 subjects, a unique set of target and nontarget images was selected. Once all images and highlights were selected the order of the images within each block, for each subject, was randomly permuted to produce the final sequence of images per subject.

Each subject was presented with the 36 trial images in four blocks of nine with instructions between each block to remind them of the task and to provide a break between blocks. During a trial subjects were free to click anywhere in the image as often as desired. If the click location was not within 30 pixels of a McDonald's center, if one was present in the image, a buzzing sound was played and the trial continued. If a McDonald's was present and clicked a pleasant ding sound was played and the trial ended. The trial also ended if subjects clicked the "No McDonald's Present" button. When clicked, a pleasant ding sound would be played if there was in fact no target present. Otherwise, the buzzing sound was played.

5.3.2 Results

Figure 5.12 presents the fraction of targets detected by condition as a function of time since stimulus onset. The curves in this Figure do not asymptote at 1.0 because subjects may have given up and missed a target. The vertical lines indicate the mean time, across subjects, to correctly locate a target given that a target is present. Highlighting allows subjects to correctly locate a target more quickly than in the control condition (5.3530 ± 0.2546 soft, 5.3039 ± 0.3043 hard, 8.0411 ± 0.8131 control). The mean reaction time is the mean over each subject's median reaction time per condition. There is a highly significant difference for reaction time between control and either highlight condition ($t(89) = 3.1127, p = 0.002$) but no significant difference between the two highlight conditions ($t(89) = 0.1227, p = 0.9025$).

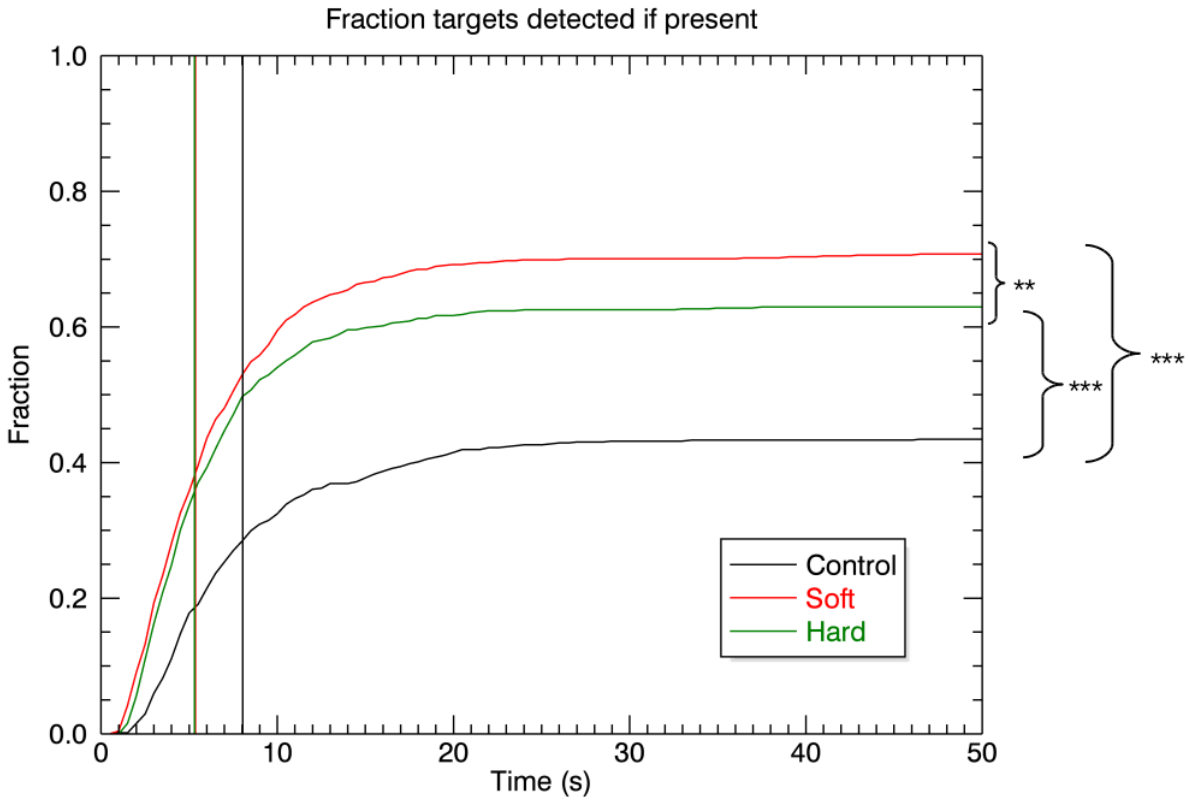
The Figure shows that the soft highlighting curve is steeper and asymptotes at a higher level than the hard highlighting curve, and the hard highlighting curve is steeper and asymptotes at a higher level than the control condition. The qualitative shape of these curves indicates that soft highlighting both allows targets to be found faster and leads to fewer target misses.

At the bottom of the Figure are the results of the (parametric) paired t-test and the (non-parametric) Wilcoxon test supporting the conclusion that soft > hard > control in terms of the asymptotic detection rate. The statistical tests are summarized by the brackets to the right of the Figure indicating significance levels of a paired comparison.

Although soft highlighting leads to significantly more targets detected than hard highlighting, no difference is observed in the mean time to localize a target for soft versus hard. This result contrasts with the result of Experiment 6. Regardless of whether soft highlighting is superior because of its asymptotic performance or because it facilitates detection, all our evidence to this point suggests soft highlighting is typically more effective than hard highlighting, and never less effective.

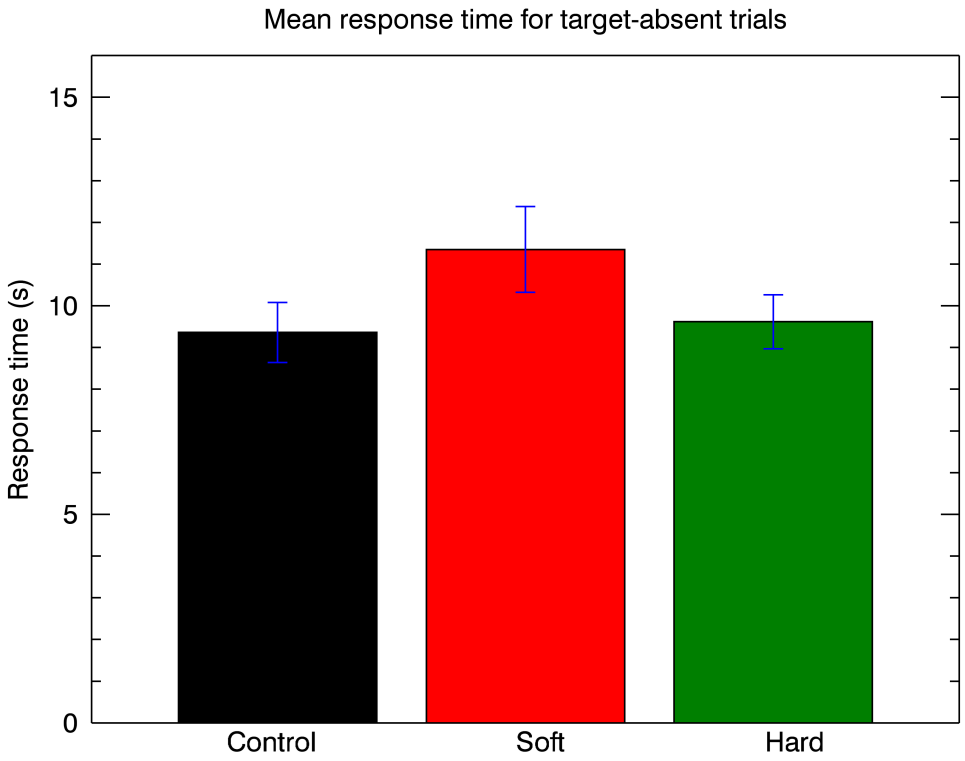
Our analysis of Experiment 7 to this point has focused on target-present trials. We now turn to target-absent trials and examine how much time subjects spend to determine that no target is present. Figure 5.13 shows the mean response time for target-absent trials by condition. The paired t-tests show that for the target-absent trials, subjects were significantly slower to end a trial with soft highlighting than with hard highlighting ($t(89) = 2.9436$, $p = 0.0041$) or the control condition ($t(89) = 3.4744$, $p = 0.0008$). This result is the only evidence we obtained that did not support soft highlighting as the most effective technique.

We examined the number of times subjects clicked on the image by condition to see if highlighting led to more clicks compared to the control condition. The mean number of nontarget clicks ($\pm SE$) in target-absent images for the control condition was 1.8111 ± 0.2129 for hard highlighting 2.7472 ± 0.2945 and for soft highlighting 3.1500 ± 0.4350. The control condition led to significantly fewer clicks than hard ($t(89) = 5.2077$, $p = 0.0000$, Cohen's $d = 0.3840$) or soft ($t(89) = 4.4414$, $p = 0.0000$, Cohen's $d = 0.4121$) highlighting. However, there was no significant difference in the




```
*** control-soft: 0.4389 vs 0.7069, t(89)=-10.9161 (p=0.0000), w(89)= 10618.5 (p=0.0000)
*** control-hard: 0.4389 vs 0.6278, t(89)= -6.5335 (p=0.0000), w(89)=  9982.0 (p=0.0000)
**  soft-hard   : 0.7069 vs 0.6278, t(89)=  3.0974 (p=0.0026), w(89)=  7152.5 (p=0.0038)
```

Figure 5.12: Experiment 7. Fraction of targets detected as a function of time if a target was present in the image by condition. Paired t-test results are below along with two-tailed Wilcoxon signed-rank test results which confirm the t-test results. It is clear that subjects missed fewer targets in the soft highlighting condition than in either the hard or control conditions. The mean time to end the trial when the target was detected is shown with a vertical line.




```
control-soft:  9.3622 vs 11.3497, t(89)=3.4744 (p=0.0008)
control-hard:  9.3622 vs  9.6168, t(89)=0.5762 (p=0.5660)
soft-hard    : 11.3497 vs  9.6168, t(89)=2.9436 (p=0.0041)
```

Figure 5.13: Experiment 7. Mean response time (± SE) for target-absent trials by condition. Based on the unpaired t-test of the log of the results there was a significant difference between the soft condition and the hard and control conditions but no significant difference between control and hard highlighting.

number of clicks between the two highlighting conditions ($t(89) = 1.2915$, $p = 0.1999$, Cohen's $d = 0.1143$). These results indicate that the presence of highlighting, either soft or hard, causes subjects to click on more potential target locations before deciding no target is present.

The structure of Experiment 7 did not allow for a false positive condition, however, the increased number of nontarget clicks in target-absent cases when highlighting was present, especially soft highlighting, implies that were it possible to select a false positive location in the image subjects would have done so more frequently when highlighting was present.

We broke down responses by whether the target was contained within a marker (hard highlighting) or when the assigned probability was below 0.5 or above (soft highlighting). A target presence probability of 0.5 was minimum used when deciding whether or not to put a marker on the image in the case of hard highlighting. The results are shown in Figure 5.14. The Figure shows an increase in localization of targets that are highlighted and above a classifier threshold of 0.5. The highlighting has made target localization easier and the design of the experiment has made it possible to click freely without additional effort.

For hard highlighting, subjects are significantly more likely to locate targets that the classifier has also highlighted with a marker locating the target 40.28% of the time if the target is inside a marker compared to only 30.07% of the time when the target is outside a marker. For soft highlighting, subjects are more likely to locate a target when highlighted above 0.5 than not (47.34% versus 20.54%), even more so than in the case of hard highlighting when the target is inside a marker though the difference between the two is not significant ($t(89) = 1.3789$, $p = 0.1697$). However, when the target is highlighted below 0.5, subjects detected the target only 20.54% of the time compared to 30.07% of the time for hard highlighting ($t(89) = 2.4751$, $p = 0.0143$). This implies that subjects are paying more attention to the soft highlighting present in the image. When highlighted above 0.5 subjects were more likely to locate soft highlighted targets than hard, though the difference was not statistically significant for the number of subjects in the experiment. If this trend is real, it offers more support for the idea that soft highlighting is helping subjects more so than hard highlighting.

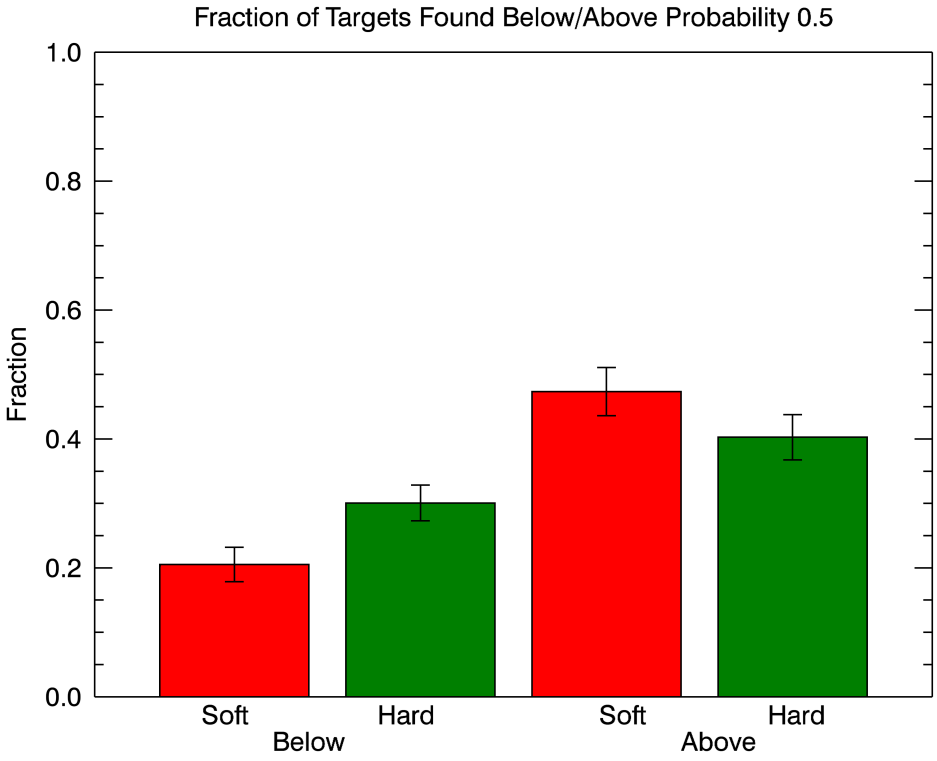


Figure 5.14: Experiment 7. Fraction of targets found (mean ± SE) when the classifier assigned probability was below 0.5 and above 0.5 for both soft and hard highlighting conditions. Subjects were more likely to correctly identify the target when highlighting was present though there was no significant difference between soft and hard highlighting (soft/hard above 0.5, ($t(89) = 1.3789$, $p = 0.1697$). When highlighting was not present over the target subjects were nearly twice as likely to locate it when hard highlighting was present in the image than soft highlighting (soft/hard below 0.5, ($t(89) = 2.4751$, $p = 0.0143$).

We ran an ANOVA with highlight type (soft, hard) and target prediction probability (above 0.5, below 0.5) as within-subject factors. We get a main effect of prediction probability ($F(1, 89) = 34.86$, $p < 0.0001$, Cohen's $d = 0.61$), no effect of highlighting type($F(1, 89) < 1$), and a significant interaction between prediction probability and highlighting type ($F(1, 89) = 5.32$, $p = 0.023$). The number of target present images with the target prediction probability greater than 0.5 was 186 out of 245 (75.9%).

The ANOVA results lead to the belief that subjects are choosing to attend to soft highlights more closely, suggesting that they find the highlights useful, especially that a weak soft highlight is one subjects feel comfortable ignoring just as a strong soft highlight is one subjects feel uncomfortable ignoring.

Next, we investigate how subjects' performance might have changed over the course of the experiment; learning might lead to an improvement in performance, whereas fatigue might lead to a decrement in performance. It is possible to look for a learning effect by comparing, across subjects and conditions, the mean number of correct responses from a subject, either locating the target or correctly stating that no target is present, by block. The mean number of correct responses by block is shown in Figure 5.15 where it is clear that there is no change for any highlighting condition indicating no learning effect is present.

The results of Experiment 7 show that soft highlighting leads to faster target detection and asymptotically better detection (Figure 5.12). Experiment 6 had a similar finding; however, a target was present on every trial in Experiment 6. Experiment 7 thus extends the results of Experiment 6 to the more realistic scenario when targets are not always present. Experiment 7 also noted a slight problem for highlighting: when no target is present, soft and hard highlighting caused subjects to spend more time relative to the control condition to terminate the trial (Figure 5.13). Target-absent trials with highlighting cause subjects to spend more time rejecting highlighted locations than if they had not been highlighted.

Attention is drawn to highlights - hard and soft. This slows subjects a bit and causes them to guess (click more often). This bottom-up deployment of attention may be due to the fact that

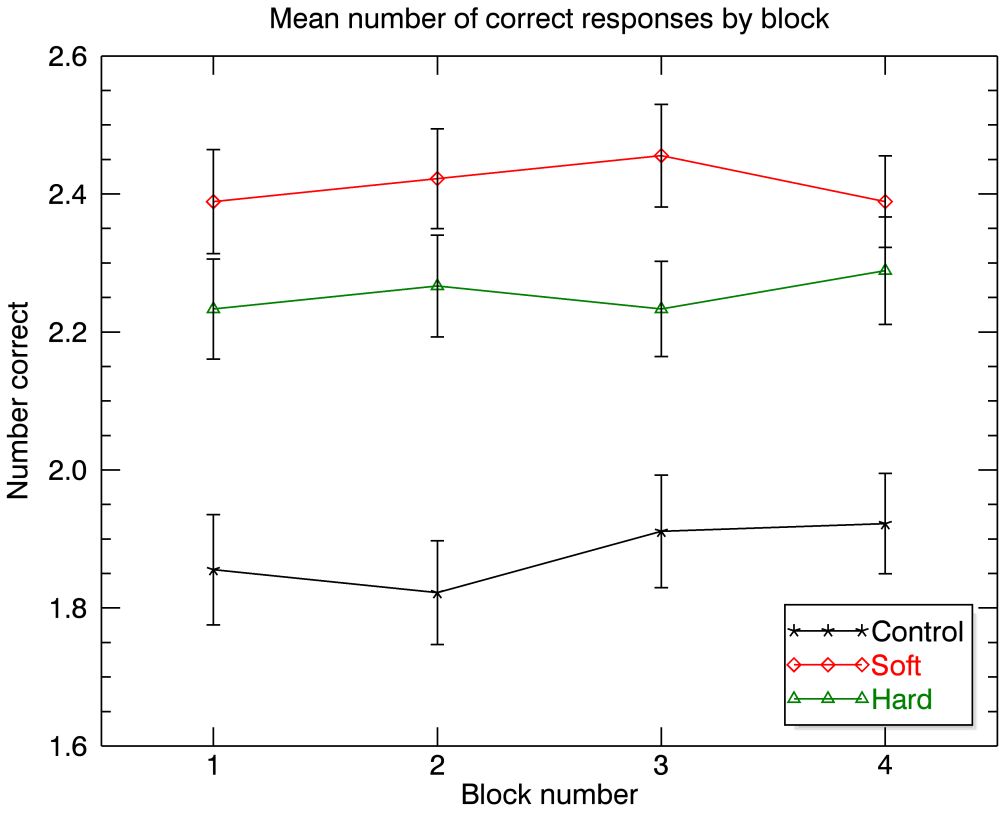


Figure 5.15: Experiment 7. Mean number of correct responses by block and condition. No significant difference across blocks is evident.

subjects are naive to the task and experts might not be distracted in the same way.

5.4 Experiment 8: Locating an Unknown Number of Targets (0-1) in a Satellite Image with No Real-Time Feedback

Experiment 7 offered a more naturalistic stimulus environment than Experiment 6 in that not every trial contained a target. As a result, subjects in Experiment 7 might miss a target by terminating a target-present trial before the target is found. Such a response is often termed a false negative. Experiment 7 did not permit false positives: marking a location as a target which was not actually a target. Experiment 8 attempted to bring the experimental task closer to a real-world scenario by allowing for false positive responses. In real-world scenarios involving image analysis, experts are not given feedback telling them that their hypothesized target location is correct or incorrect. Experiment 8 was designed in this way: on each trial, subjects selected a location or indicated no target was present, and only once they committed was feedback provided. We felt it necessary to provide feedback after each trial because without expertise and without feedback we were concerned that subjects would not perform the task carefully. We hypothesized, as in Experiment 7, that soft highlighting would lead to fewer missed targets when compared to hard highlighting and control conditions. Experiment 8 followed exactly the same procedure as Experiment 7 in selecting images and trial sequences. See Experiment 7 methods section (5.3.1) for details.

5.4.1 Methods

Subjects were presented with 36 images in four blocks of nine trials. Instructions between blocks reminded subjects of the task and provided a break. An image was presented on each trial, after which subjects were free to click anywhere in the image as often as desired. Clicking on the image placed a plus symbol (*cross-hair*) at that location. If the subject clicked in another location the cross-hair moved from the previous location to the new location such that only one cross-hair appeared in the image at a time. No feedback was given when the subject was clicking on the

image to position the cross-hair.

Two buttons appeared below the image on each trial, side by side. The first, as in Experiment 7, was labeled "No McDonald's Present". The second was labeled "Submit Response". If subjects clicked on the "No McDonald's Present" button the trial ended and recorded the subject's belief that no target was in the image. If subjects clicked on "Submit Response" the trial ended and the current cross-hair location, if any, was recorded as the position where the subject believed a target was located.

After each trial, subjects received feedback. A message was displayed indicating whether or not their response—the selected location or their indication that no target was present—was correct. A reduced size version of the trial image was also shown with a cross-hair on the location of the actual target, if present. In addition, if the subject's response was correct a pleasant ding sound was played, otherwise a buzzing sound was played. The feedback remained on the screen until subjects clicked a button to advance to the next trial.

Subjects were paid $1.25 for completing the experiment which typically took on the order of ten minutes.

Of 130 subjects who started Experiment 8, 90 completed it successfully. As in Experiments 6 and 7, subjects were free to leave at any time and were rejected if they clicked outside the browser window during a trial. There were 40 subjects who did not complete the experiment of which 16 left voluntarily and 24 were rejected for clicking outside their browser window during a trial. No subjects who had interacted in any way with Experiments 6 and 7 were allowed to participate in Experiment 8.

5.4.2 Results

Because of the similarity between Experiment 7 and 8, many of the analyses we performed for Experiment 8 were identical to those performed for Experiment 7. However, we conducted additional analyses specific to Experiment 8 that examined false positive responses.

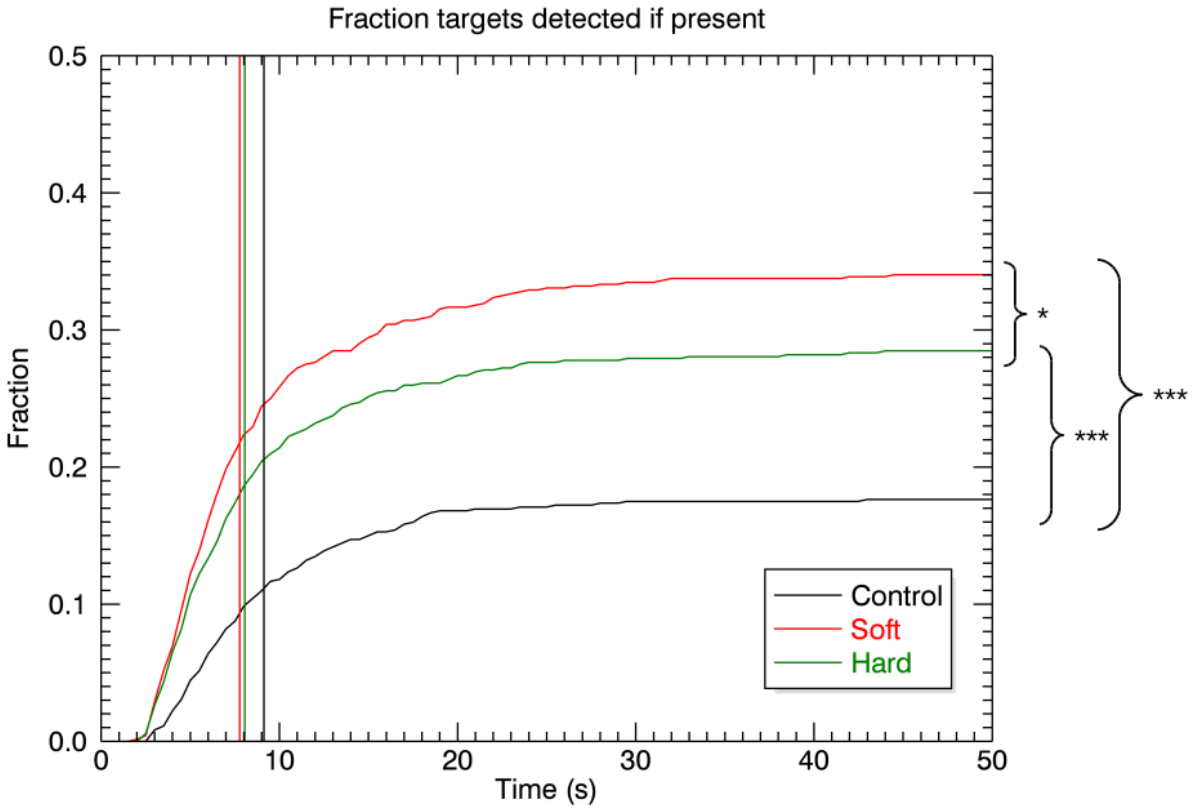
Figure 5.16 shows the fraction of targets detected as a function of time for target present

trials, with one curve for each highlighting condition. Consistent with Experiment 7, subjects are faster to locate the target and have asymptotically better performance with soft highlighting than with hard highlighting or the no-highlighting control. The curves do not asymptote at 1.0 because subjects might miss targets.

Below the Figure are the results of paired t-tests and nonparametric tests showing that the bracketed asymptotic fraction of targets located are reliably different. The vertical lines in the Figure indicate the mean time (mean $\pm$ SE), across subjects, to terminate a target-present trial by correctly locating the target (7.7628 ± 0.4671 soft, 8.0616 ± 0.6419 hard, 9.1305 ± 0.5726 control). The mean reaction time is the mean over each subject's median reaction time per condition. Just as in Experiment 7 (Figure 5.12), highlighting led to faster reaction times to correctly locate targets. However, in this case there is no statistically significant difference between any of the reaction times.

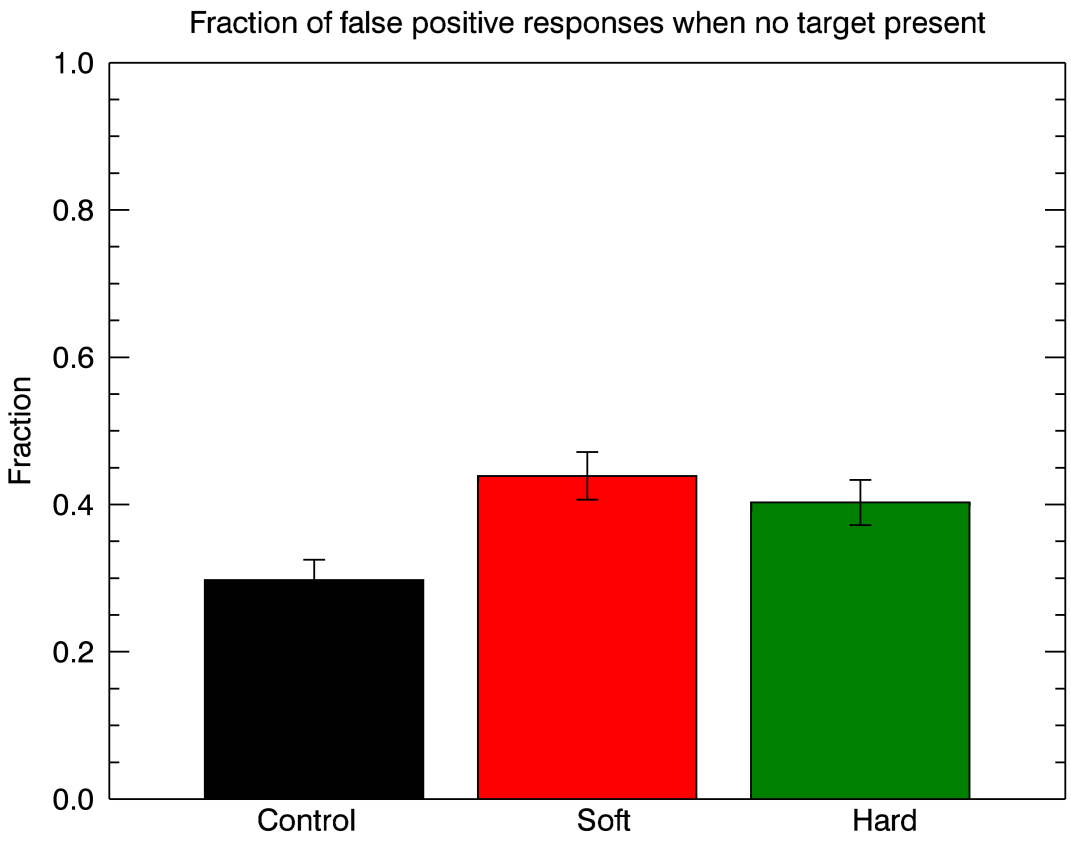
The primary purpose of Experiment 8 over Experiment 7 was to examine false positive responses, i.e., cases where a target is identified by a subject which is not actually a target. Figure 5.17 shows a comparison across conditions of the probability that a location is reported to contain a target when that location is not a target. This comparison includes only target-absent trials. Soft or hard highlighting leads to an increase in the false positive rate compared to the control condition (Figure 5.17). We conjecture that this is due to the fact that subjects are not domain experts and they are somehow trusting the presence of the soft highlighting and markers to be an indication that a target really is present.

Further support for this conjecture comes from an analysis examining the mean response latency for false positive trials in the three highlighting conditions. Figure 5.18 separates the two possible response types showing that for the false positive case, subjects took significantly longer to decide, incorrectly, that a target was present in the image in the control condition relative to either highlighting condition. This seems to indicate that subjects were willing to commit to a false target location when soft or hard highlighting was present but took longer to decide when in the control condition. Subjects appear to be trusting the highlights which leads to faster response




```
*** control-soft: 0.1764 vs 0.3403, t(89)= -6.4665 (p=0.0000), w(89)= 10072.0 (p=0.0000)
*** control-hard: 0.1764 vs 0.2875, t(89)= -5.6410 (p=0.0000), w(89)=  9627.5 (p=0.0000)
*   soft-hard   : 0.3403 vs 0.2875, t(89)=  2.0995 (p=0.0386), w(89)=  7486.5 (p=0.0545)
```

Figure 5.16: Experiment 8. Fraction of targets detected as a function of time if a target was present in the image by condition. Paired t-test results are below along with two-tailed Wilcoxon signed-rank test results which confirm the t-test results. It is clear that subjects missed fewer targets in the soft highlighting condition than in either the hard or control conditions. The mean time to end the trial is shown with a vertical line.




```
control-soft: 0.2972 vs 0.4389, t(89)=-3.6784, p=0.0004, w(89)=7119.5, p=0.0024
control-hard: 0.2972 vs 0.4028, t(89)=-2.8651, p=0.0052, w(89)=7338.5, p=0.0169
soft-hard   : 0.4389 vs 0.4028, t(89)= 0.9501, p=0.3446, w(89)=8372.5, p=0.5022
```

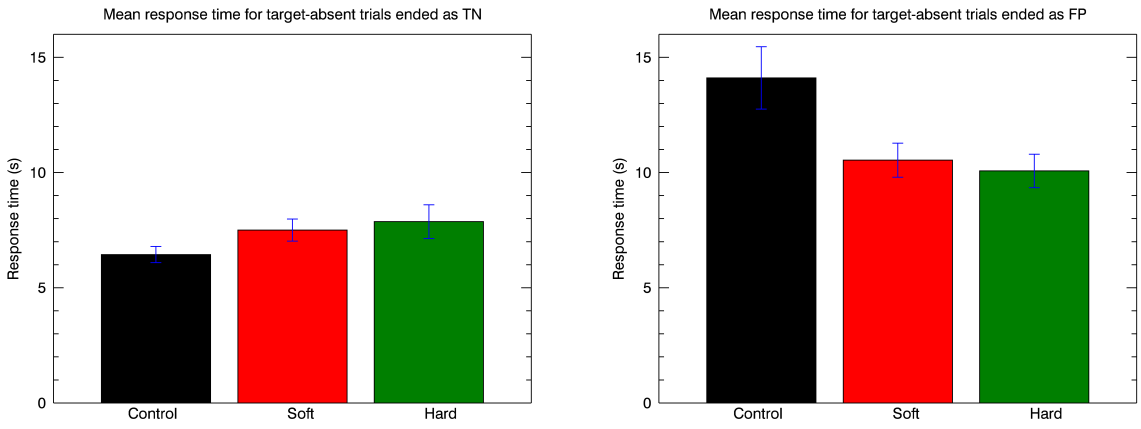
Figure 5.17: Experiment 8. Fraction of false positive no target present trials, by condition. These are trials where the subject marked a location on the image as the target when no target was present. The presence of highlighting, either soft or hard, lead to an increase in the number of false positives over the control condition.

times.

In the graph on the left, for target-absent trials in which subjects correctly decided no target was present, we see only a small difference among conditions. This result is different than what we observed in Experiment 7 (Figure 5.13), where soft highlighting led to an increase in the time to complete the trial relative to hard highlighting and no highlighting. In both Experiment 7 and Experiment 8, soft highlighting slows subjects down in target absent displays relative to target present displays, however, the magnitude of the difference is small (Cohen's $d = 0.26$).

A likely explanation is that Experiment 7 provided continuous immediate feedback, whereas Experiment 8 provided feedback only at the trial's end. Consequently, Experiment 8 provided no incentive to continue clicking on the image. Thus, the absence of immediate feedback in Experiment 8, which made the experiment more naturalistic, appears to have pointed us to an artifact introduced by the design of Experiment 7, and allows us to explain away the one weakness we observed with soft highlighting.

We broke down responses by whether the target was contained within a marker (hard highlighting) or when the assigned probability was below 0.5 or above (soft highlighting). A target presence probability of 0.5 was minimum used when deciding whether or not to put a marker on the image in the case of hard highlighting. The results are shown in Figure 5.19. For hard highlighting, subjects are significantly more likely to locate targets that the classifier has also highlighted with a marker, as was the case in Experiment 7, locating the target 33.67% of the time if the target is inside a marker compared to only 16.06% of the time when the target is outside a marker. For soft highlighting, subjects are more likely to locate a target when highlighted above 0.5 than not (37.77%), even more so than in the case of hard highlighting when the target is inside a marker though the difference between the two is not significant ($t(89) = 1.3444$, $p = 0.1822$). However, when the target is below 0.5, subjects detected the target only 8.11% of the time compared to 16.06% of the time for hard highlighting ($t(89) = 2.8606$, $p = 0.0053$). This implies that subjects are paying more attention to the soft highlighting present in the image. As above for Figure 5.17, we believe this is due to the fact that subjects are not domain experts and that they are somehow




```
TN:
  control-soft: 6.4409 vs 7.5023,   t(89)=2.5049 (p=0.0126)
  control-hard: 6.4409 vs 7.8710,   t(89)=2.2597 (p=0.0243)
  soft-hard   : 7.5023 vs 7.8710,   t(89)=0.2366 (p=0.8131)

FP:
  control-soft: 14.1079 vs 10.5368, t(89)=2.5492 (p=0.0114)
  control-hard: 14.1079 vs 10.0712, t(89)=2.7252 (p=0.0069)
  soft-hard   : 10.5368 vs 10.0712, t(89)=0.1702 (p=0.8649)
```

Figure 5.18: Experiment 8. Mean response time (± SE) for target-absent trials by condition and whether the response was a true negative (left) or false positive (right). Based on the unpaired t-test of the log of the results there was a significant difference between the control condition and the two highlight conditions but no significant difference between soft and hard highlighting.

trusting the presence of soft highlighting to be an indication that a target really is present, even more so than in the case of hard highlighting.

We ran an ANOVA with highlight type (soft, hard) and target prediction probability (above 0.5, below 0.5) as within-subject factors just as in Experiment 7. We get a main effect of prediction probability ($F(1, 89) = 5.03$, $p < 0.001$, Cohen's $d = 1.10$), no effect of highlighting type ($F(1, 89) < 1$), and a reliable interaction between prediction probability and highlighting type ($F(1, 89) = 9.53$, $p = 0.003$). The number of target present images with the target prediction probability greater than 0.5 was 186 out of 245 (75.9%).

The ANOVA results again lead to the belief that subjects are choosing to attend to soft highlights more closely, suggesting that they find the highlights useful, and that a weak soft highlight is one subjects feel comfortable ignoring just as a strong soft highlight is one subjects feel uncomfortable ignoring. Consider again Figure 5.5 showing a target that is in the weak (below 0.5 probability) soft highlight condition.

The subject responses fell, for each trial, into one of four possible outcomes. If the target was present and correctly identified the trial was a true positive (TP). If the target was not present and correctly identified as not present the trial was a true negative (TN). A false positive (FP) is when the subject marked a location as the target when no target was present or the target was in another location. Lastly, if a target was present and the subject indicated that no target was present the trial was a false negative (FN). With these in mind, we can define the false positive rate (FPR) and false negative rate (FNR) as,

$$FPR = \frac{FP}{FP+TN}$$
$$FNR = \frac{FN}{TP+FN} = 1 - \frac{TP}{TP+FN}$$

It is not clear *a priori* which highlight condition is best and each leads to a specific false positive and false negative rate for each condition. The mean FPR for each condition (mean ± SE) was: 0.6169 ± 0.0197 control, 0.7031 ± 0.0196 soft, and 0.4027 ± 0.0306 hard. The mean FNR for each condition was: 0.6560 ± 0.0329 control, 0.3195 ± 0.0320 soft, and 0.7125 ± 0.0184 hard. These rates are linked, so to compare the highlighting conditions we require a measure that integrates them,

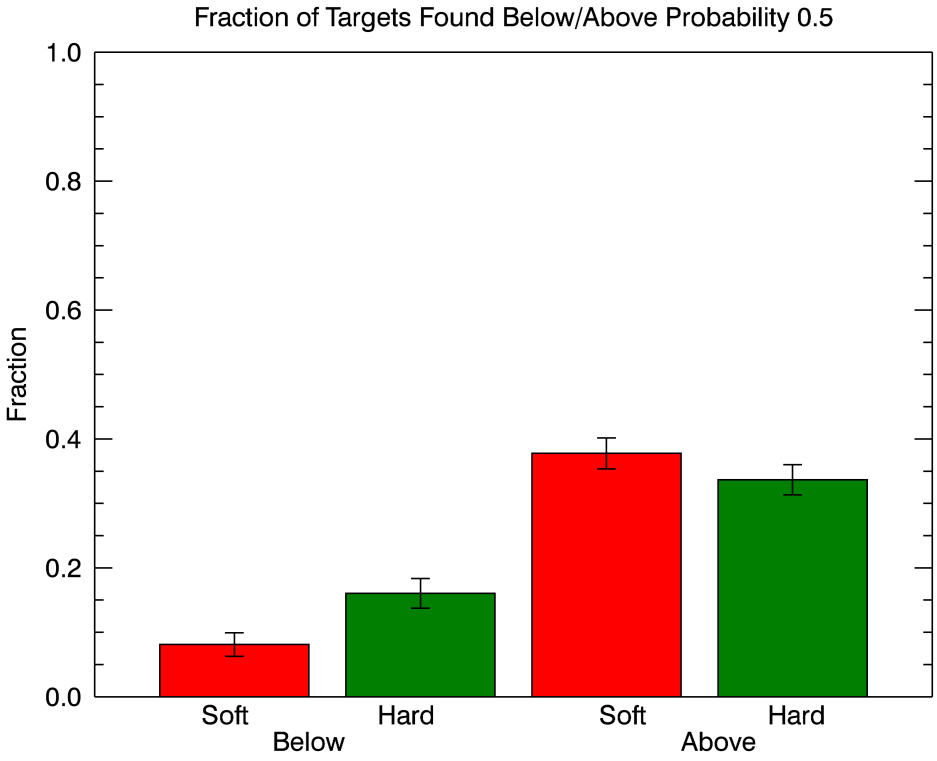


Figure 5.19: Experiment 8. Fraction of targets found (mean ± SE) when the classifier assigned probability was below 0.5 and above 0.5 for both soft and hard highlighting conditions. Subjects were more likely to correctly identify the target when highlighting was present though there was no significant difference between soft and hard highlighting (soft/hard above 0.5, ($t(89) = 1.3444$, $p = 0.1822$). When highlighting was not present over the target subjects were nearly twice as likely to locate it when hard highlighting was present in the image than soft highlighting (soft/hard below 0.5, ($t(89) = 2.8606$, $p = 0.0053$).

i.e., d'. There are more false positives, but also more true positives, when highlighting is present. Highlighting biases subjects to clicking, which leads to more false positives and more true negatives, so neither rate alone is meaningful, hence our calculation of d'.

We calculated d' for each subject, across trials, by condition (control, soft, hard). We adjusted the false positive rate to account for the fact that targets could appear at any location. If we consider the image to be made up of N non-overlapping tiles a false positive may happen in any of these tiles. Therefore, we adjust the FPR prior to calculating d' scaling it by $1/N$. We set $N = 100$ but consider other values as well.

The mean d' across subjects in the control condition was 1.9194 ± 0.1421. For soft highlighting it was 3.1864 ± 0.1411 and for hard highlighting 2.1361 ± 0.0825. The medians closely matched the means. Each subject completed 36 trials. There were 90 subjects total.

Paired t-tests show highly significant differences between soft versus control ($t(89) = 6.2454$, $p = 0.0000$, Cohen's $d = 0.9432$) and soft versus hard ($t(89) = 6.6131$, $p = 0.0000$, Cohen's $d = 0.9581$) but not between hard versus control ($t(89) = 1.3944$, $p = 0.1667$, Cohen's $d = 0.1966$). Varying the number of possible non-overlapping locations from 100 to 50, 25, and 10, which changes the scale factor applied to the false positive rate, led to Cohen's d values above 0.9 for all soft versus control, above 0.8 for all soft versus hard, and above 0.197 for all hard versus control.

If a false positive rate is zero or one it was adjusted by adding or subtracting a small value, ϵ, so that d' could be calculated. The choice of ϵ might affect the d' calculation significantly so we investigated this effect for different ϵ values and how might influence the t-test results. These are shown in Table 5.4 where it is clear that the choice of ϵ does not alter the results of the t-tests. The d' values below use $\epsilon = 0.01$.

If the classifier alone outperforms humans using highlights, there is little point in asking humans to perform the task of locating targets. However, if humans using highlights outperform the classifier alone then highlighting is worthwhile. Therefore, we calculate d' of the classifier itself as if it were a subject. To do this it is necessary to select a probability value to act as a threshold so that a classifier output probability above the threshold is considered the same as a

ϵ	d' control	d' soft	d' hard	control vs soft	control vs hard	soft vs hard
0.1	2.11	2.95	2.12	t(89)=7.06, p=0.0000	t(89)=0.08, p=0.9367	t(89)=8.74, p=0.0000
0.05	2.04	3.03	2.12	t(89)=6.72, p=0.0000	t(89)=0.68, p=0.4996	t(89)=7.89, p=0.0000
0.03	2.00	3.08	2.12	t(89)=6.53, p=0.0000	t(89)=0.97, p=0.3330	t(89)=7.39, p=0.0000
0.01	1.92	3.19	2.14	t(89)=6.25, p=0.0000	t(89)=1.39, p=0.1667	t(89)=6.61, p=0.0000
0.005	1.87	3.24	2.15	t(89)=6.12, p=0.0000	t(89)=1.57, p=0.1189	t(89)=6.25, p=0.0000
0.001	1.78	3.36	2.17	t(89)=5.91, p=0.0000	t(89)=1.86, p=0.0658	t(89)=5.64, p=0.0000

Table 5.4: Experiment 8. Test results comparing d' for each condition as a function of ϵ, the small adjustment used when the false positive rate was either zero or one. The results of the t-tests are not sensitive to ϵ.

subject selecting a particular location while a probability value below the threshold is considered not selecting a particular location. We found the probability value for the classifier that matched the true negative rate of the subjects for images in the control condition. We then calculated the corresponding true positive rate and d'. The true positive rate for humans in the control condition and the classifier are directly comparable because of the matched true negative rate. In the control condition the human's true positive rate is 0.3459 compared to 0.4127 for the classifier.

We also found the threshold value that matched the classifier and human true positives rates in the control condition. From this we calculated the corresponding true negative rates and d'. The true negative rate for humans in the control condition and the classifier are directly comparable in this case. In the control condition the human's true negative rate is 0.9938 compared to 0.3972 for the classifier. These results indicate that the classifier performance matched in this way is significantly less than the human subjects at correctly deciding that no target was present. The fact that the true negative rates for the human subjects is so high indicates that the subjects were able to ignore extraneous highlighting from the classifier.

The d' values for human subjects and the classifier matched for both true negative rate and true positive rate are shown in Table 5.5. From the Table it is clear that the classifier was able to outperform the human subjects in the control condition ($d' = 2.2778$ versus $d' = 1.9194$) but when the human subjects used the classifier output in the soft highlighting case they were able to increase their performance over that of the classifier alone (d'=3.1864) giving strong evidence of a synergy between the subjects and the machine learning classifier. This effect was not seen in the

	d'
humans(control)	1.9194
humans(soft)	3.1864
humans(hard)	2.1361
classifier (matched TN rate)	2.2778
classifier (matched TP rate)	2.1272

Table 5.5: Experiment 8. Comparing d' for human subjects and the `mcdonalds3` classifier. The d' values represent 100 non-overlapping locations as discussed for the human subjects. Two d' values are given, one when matching the classifier true negative rate to the human true negative rate in the control condition and the other when matching the true positive rate to the human true positive rate in the control condition. The classifier performs better than humans in the control condition but when humans combine the classifier output using soft highlighting the two together perform better than either alone. This effect was not seen in the hard highlighting case.

hard highlighting case.

The effect of the number of locations in the calculated classifier d' was investigated giving, for a matched true negative rate, d' values of 2.2778 for 100 locations, 2.0237 for 50 locations and 1.7351 for 25 locations.

Just as in Experiment 7, the trials for this experiment are organized into four blocks of nine images each. It is possible to look for a learning effect by comparing, across subjects and conditions, the mean number of correct responses from a subject, either locating the target or correctly stating that no target is present, by block. The mean number of correct responses by block is shown in Figure 5.20 where a trend may be present in both the hard highlighting and control condition.

The results of Experiment 8 mirror those of Experiment 7 showing that soft highlighting leads to improved target detection (Figure 5.16). However, highlighting does lead to an increase in the number of false positives (Figure 5.18 and Figure 5.17), perhaps due to a bias that highlighting introduces that causes subjects to want to select *some* location as a target. Still, soft highlighting leads to a significant increase in the detected signal with a Cohen's d value above 0.9 compared to hard highlighting and the control condition. It also leads to a synergy with the classifier in the soft highlighting condition so that subjects are able to improve their performance beyond that of the unhighlighted control case and the classifier alone.

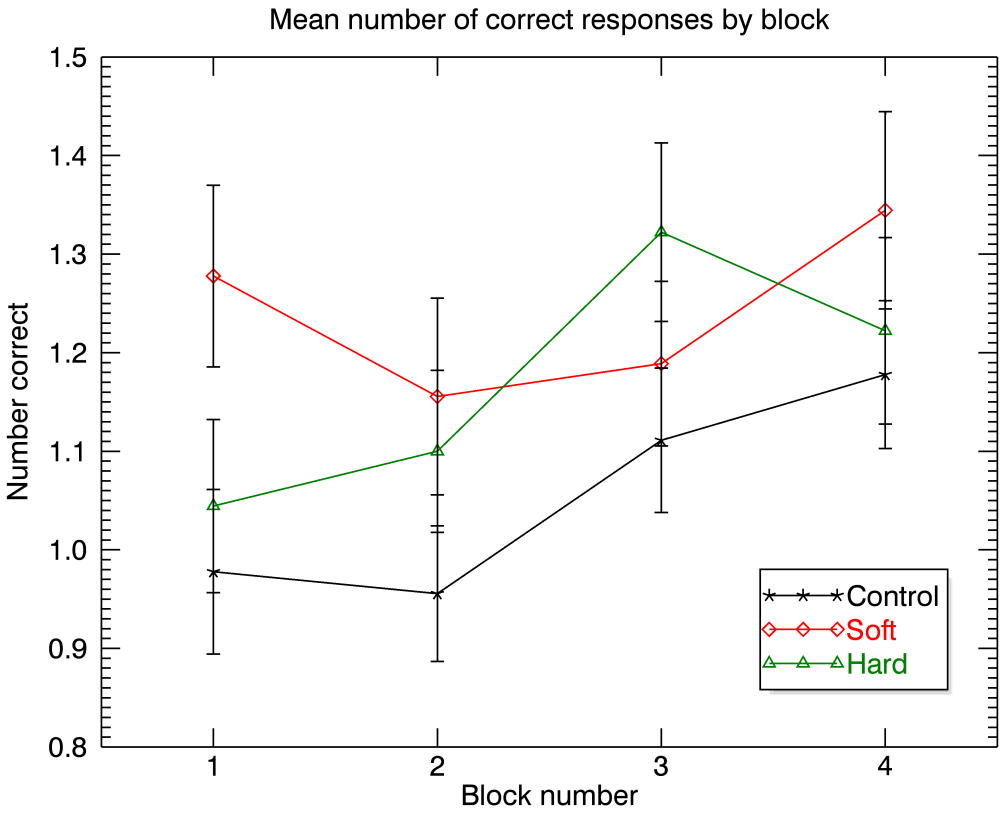


Figure 5.20: Experiment 8. Mean number of correct responses by block and condition. A trend may be present for learning across blocks in the hard and control conditions.

5.5 Discussion

The experiments of this chapter sought to extend the results of Chapter 3 from synthetic images to satellite images.

In Experiment 6 we asked subjects to locate a target object when the target was always present and the subject was free to search, with feedback, until the object was found. The results showed that subjects were quicker to locate the target in the soft highlighting case and with fewer clicks. The counterbalanced structure of the experiment enabled us to show that soft highlighting again leads to a significant reduction in the time to locate targets compared to the control.

Experiment 7 made the search scenario more realistic in that not all images contained a target. Subjects were still free to click at will, with feedback, until the target was located but they were also able to give up if they felt no target was present. This created the possibility of missed targets (false negatives). Soft highlighting led to an increase in the number of targets found when compared to both hard highlighting or the control condition. However, subjects took longer to terminate target-absent trials when highlighting was used, especially soft highlighting. We hypothesized that the increased time to termination was due to the fact that subjects felt compelled to continue clicking until all highlighted locations had been exhausted. We suspected that eliminating feedback would eliminate this compulsion, and indeed, in Experiment 8 when feedback was removed, no increase in search time was observed between target-absent trials with highlighting.

Lastly, in an attempt to make the task more naturalistic, Experiment 8 eliminated real-time feedback available in Experiment 7 and required subjects to commit to a response before offering feedback. In doing so, Experiment 8 allowed for false positive responses, i.e., the selection of a location as target when that location did not contain a target.

We found that soft highlighting lead to more false positives as well as more true positives. Neither rate alone is meaningful, they must be considered together. Our d' analysis integrates the two rates into one measure. We found that in the soft highlighting condition d' was greater

than in the hard or control condition with Cohen's d > 0.9 indicating that soft highlighting helps subjects. We also calculated d' for the classifier itself to see if there was a synergy between subjects on their own and the classifier. The soft highlighting d' value was greater than either subjects without highlighting and the classifier on its own. This indicates that subjects are incorporating the classifier results through soft highlighting and using it successfully to achieve performance exceeding either subjects or the classifier alone. This effect was not seen with hard highlighting.

The results of Experiment 8 provide strong evidence that subjects are able to integrate the classifier results through soft highlighting in order to increase their performance.

Chapter 6

Discussion

The experiments of this dissertation were motivated in part by comments heard from image analysts with whom the author interacts professionally. Image analysts spend the majority of their time viewing, in great detail, satellite imagery in order to locate and characterize objects. Sometimes these objects are targets of high importance that must be located in a time-critical manner. At other times the goal is mapping and every object is important. These analysts have, over the years, used systems that apply hard highlighting to images to show the output of a classifier in order to fully or partially automate their tasks. Historically, analysts have strongly disliked these systems because they are too distracting in appearance and the errors the systems make take too much of the analysts' time to recover from. When asked if there would be interest in a system that showed the output of the classifier while still showing the image so that the analyst would be able to easily ignore classifier errors the answer was a resounding "yes".

The medical imaging community makes use of hard highlighting as well, particularly in CAD (computer-aided detection) systems which are in wide use clinically, especially for mammography. Therefore, how these systems affect human radiologists and others interpreting images has been a topic for research for some time. In Chapter 2 we evaluated key research in this area (see [111, 226, 4]) showing the potential for harm from hard highlighting in terms of missed targets. A missed lesion in mammography could very easily prove fatal.

The anecdotal experience with satellite image analysts combined with the results of research in the medical domain motivated the synthetic image experiments of Chapter 3 and the satellite

image experiments of Chapter 5 as we sought to demonstrate that soft highlighting was, in fact, a viable alternative approach which would allow a human and a computer to cooperate and thereby obtain results that are better than either along could achieve.

6.1 General Discussion of Experimental Results

In Experiment 1 we asked subjects to search for a fixed number of targets in a synthetic image. This experiment showed that subjects were able to use soft highlighting to locate targets more quickly, an effect which only increased as the quality of the stochastic classifier improved (d' value increased). The surprising result of this experiment was that even a weak classifier ($d' = 0.75$) produced highlights that supported the subject's visual search even if the same classifier often highlighted nontargets as well.

In Experiment 2 we asked subjects to search for a single target while we altered the quality of the stochastic classifiers as well as the display size (number of display elements). This allowed us to examine response latencies for a particular condition across display sizes. The results of Experiment 1, the faster localization times, might show themselves in Experiment 2 by changing the search slopes or intercepts as display size changes. If the slopes change then highlights makes it easier for subjects to reject display elements that are not helpful indicating that highlighting guides attention. The intercept reflects fixed preprocessing or motor preparation time.

An ANOVA analysis showed that highlighting reduced the time to search for the target indicating that soft highlighting guided attention to relevant locations in the display. The search slopes demonstrated that as the classifier quality improves subjects were locating targets more quickly, especially in the soft highlighting condition. Lastly, the search intercept showed that the more complex displays with highlights did cause an increase in parsing time but that subjects still found targets more quickly when soft highlighting was present.

Experiments 1 and 2 provided evidence that soft highlighting was superior to no highlighting as long as the highlights are based on a classifier with discriminative ability. In Experiment 3 we turned to a direct comparison of soft and hard highlights and their effects on search efficiency and

found that soft highlighting supports human visual search better than hard highlighting.

For Experiment 4 we asked subjects to search for 0 to 2 targets in the synthetic images and to search until they were confident that no targets remained. This experiment showed that subjects using soft highlighting and a high quality classifier were faster at initial detection and also detected more targets per time point than weaker classifiers. This result is a win-win from a speed-accuracy trade off. Better classifiers (larger d') lead to early termination (faster localization) for soft highlighting but we found no evidence that a better classifier caused subjects to give up sooner. A key finding of this experiment is that soft highlighting leads to fewer missed targets.

Our final experiment with synthetic images and stochastic classifiers, Experiment 5, followed the exact design of Experiment 4 but allowed subjects the freedom to toggle highlighting on or off at will. As in Experiment 4 we found that soft highlighting helped subjects to miss fewer targets.

When we examined how subjects made use of the ability to toggle highlights on and off we were somewhat surprised to see that, for the most part, subjects either left the highlights on or turned them off and left them off. Also, subjects were more likely to turn off highlighting when the classifier was strong (higher d'). We believe this is due to the fact that strong highlighting makes nontargets less salient so subjects may have been turning highlighting off to avoid missing targets.

Experiments 1-5 clearly showed a strong advantage to using soft highlighting. We wanted to know if these results would translate to a real-world example so for the remaining experiments we explored highlighting of satellite images using both soft and hard highlights derived from a modern machine learning classifier.

Experiment 6 looked at how highlighting style: soft, hard or none, affected the reaction time of subjects looking for a single target object in a satellite image. We showed that soft highlighting enabled subjects to locate targets more quickly than either hard highlighting or the control condition of no highlighting which is exactly what we saw in the results of Experiments 1-5.

Since Experiment 6 offered subjects immediate feedback on each click and did not limit the number of clicks they could make they might be locating objects quickly by simply clicking many times on the image. The results of Experiment 6, however, demonstrated that soft highlighting

again led to faster target localization with fewer clicks than either the hard or control condition.

We examined the relationship between reaction time (time to locate the target) and the heat map probability (reflecting the classifier's confidence of a target). We showed that soft highlighting led to a steeper, more negative, slope of the RT vs heat map probability plot. This means that soft highlighting caused subjects to find the targets more quickly than hard highlighting or the control condition which showed no effect between reaction time and heat map probability. An effect between subject reaction time and heat map probability for the control condition would have indicated that targets which were easy for the subject to locate were also easy for the classifier to identify. Experiment 6 provided evidence that not only does soft highlighting lead to improved target localization but that soft highlighting provides better information about the confidence level of the classifier and that subjects were able to use this information.

In Experiment 6 a target was always present and subject were required to locate it. Experiment 7 added the possibility of missing a target (a false negative condition). The results of Experiment 7 also show that soft highlighting leads to faster target detection but because a target was not always present it extends the result of Experiment 6 to the more realistic scenario. Experiment 7 also demonstrated that when no target is present, soft and hard highlighting caused subjects to take longer to terminate the trial compared to the control condition. This additional time was spent clicking on more potential target areas in the image before deciding no target was present. This increased time to terminate a target-absent trial for soft highlighting especially, was concerning, but the results of Experiment 8 offer an explanation.

For Experiment 8 we added the possibility of a false positive. Subjects were no longer receiving real-time feedback to mouse clicks but instead were placing a cross-hair and only receiving feedback after submitting their selection. This allowed for a false positive of selecting a location when no target was present at that location. The results of this experiment showed again that subjects missed fewer targets in the soft highlighting condition than in the other two conditions. It also showed that soft or hard highlighting led to an increase in the number of false positives when no target was present in the image. It seems plausible that this is due to the fact that the subjects

are not domain experts and were trusting the presence of soft highlighting and hard highlighting markers to be an indication that a target was really present. When only true negative trials were examined, i.e., cases where no target was present and subjects correctly indicated that fact there was no difference in the time to end a trial, unlike what was seen in Experiment 7. A likely explanation is that lack of immediate feedback in Experiment 8 removed the incentive to continue clicking on the image. Thus, is seems likely that the increase in search time for soft highlighting cases in Experiment 7 was an artifact of the design and not a true weakness of soft highlighting as a technique insofar as the design of Experiment 7 encouraged subjects to continue clicking by offering immediate feedback. This immediate feedback, combined with the presence of highlighting as an inducement to continue searching for a target is a likely explanation of the results of Experiment 7 and Experiment 8.

The d' analysis of Experiment 8 demonstrates clearly that soft highlighting enables human subjects to improve their performance beyond that of the classifier itself. Human performance when no highlighting is present is slightly worse than the classifier on its own. However, when soft highlighting is present human subjects perform substantially better than the classifier. This effect was not seen in the hard highlighting condition and is strong support for the central hypothesis of this dissertation that soft highlighting is a useful technique for improving human-classifier interaction.

The experiments above demonstrate that soft highlighting helps subjects relative to hard highlighting. A plausible explanation for this is that soft highlighting allows subjects to combine the bottom-up cues in the image with location cues from the highlighting. This lets subjects weigh the relative strength of evidence from each in order to draw a conclusion as to the location of the target. For hard highlighting this balancing between bottom-up cues and location cues may be more difficult as indicated by Krupinski [111] causing subjects to be less efficient at locating the target.

The experiments of this dissertation focused on situations where a rare target was search for in an image. Once the target was localized human subjects had little trouble discriminating it from the background. Even a fuzzy McDonald's restaurant looks quite different than the general

background of a gray scale satellite image. The experiments did not directly address a situation where localization and discrimination would be difficult for humans. It is unknown whether soft highlighting would offer any advantage in that situation. For example, target discrimination in mammography is quite difficult, would soft highlighting be an aid or an impediment in that situation (see Section 6.2)?

Machine learning systems need to balance their performance based on the cost associated with errors, either false positives or false negatives. For example, in a military setting, a false negative (missed enemy target) might have a very real cost in terms of human life. Similarly, a missed cancer detection may prove fatal. However, if the cost of a false positive is high, perhaps in terms of effort that must be spent in determining that it is indeed a false positive, it might be desirable to minimize the false positives even at the expense of an increase in false negatives. The soft highlighting techinque explored in this dissertation can by modified to address these concerns. For example, the version of soft highlighting used in Experiments 6-8 applied the probability map generated by the classifier directly to the contrast in the image through the saturation channel as $S = H$ where S is the HLS color space saturation and H is the heat map output from the classifier. This could easily become $S = F(H)$ where $F()$ is a transformation function that can emphasize or de-emphasize the heat map values in order to modulate what might be a false positive or false negative. Searching for an optimal $F()$ for a particular class of images might be worthwhile. As might be enabling users of a system employing soft highlighting to adjust a parameters affecting $F()$ in real-time while viewing the image with highlights.

As a whole, the experiments of this dissertation clearly show an advantage to using soft highlighting techniques in the presentation of classifier output to human viewers. With soft highlighting, viewers were able to locate targets more quickly without giving up prematurely. They were also able to miss fewer targets when using soft highlighting. This could be particularly useful in medical imaging where a missed target can carry a very high cost.

6.2 Directions for Future Work

The results presented here point the way towards potential future research. In Experiments 7 and 8 we examined cases where the target was inside or outside of a hard highlight marker. One could imagine a set of experiments where the same image is presented to different subjects, once with soft highlighting and again with hard highlighting. Are missed targets not inside of a hard highlight marker also missed when the same image is presented with soft highlighting? It is known that the presence of hard highlighting has a dramatic impact on the search pattern subjects use when looking for targets [55]. Is this effect still present or modified when soft highlighting is used? Next, one could imagine experiments where the "classifier" used to highlight the images was either a machine learning algorithm (e.g., a deep neural network) or the result of some crowdsourced classification of the image where a heat map was built from the density of target locations clicked by many different people classifiying the same image. This would allow for an experiment to see if classifier errors are more easily ignored when soft highlighting is used compared to hard highlighting.

Experiments 6-8 used an outline square to mark possible target locations in the hard highlighting condition. This square could be replaced by a filled region using the soft highlight algorithm to create a sort of hybrid soft-hard highlighting condition. In this case, would there be any change to the way subjects perform?

All of the subjects used for Experiments 6-8 were novices, not experts in the interpretation of satellite imagery. The experiments demonstrated that highlighting was likely being used by subjects as in indication that a target is probably present. As novices, this sort of information would be tempting to use, even unconsciously, as a crutch to compensate for lack of experience. It is natural to ask whether similar results would be seen by experts. In that case, will experience be able to integrate the location information of the highlights, in particular soft highlights, in such a way as to enable an overall improvement in performance?

Lastly, as noted in the section above, it is unclear whether soft highlighting would offer any advantage to tasks involving difficult localization *and* discrimination of targets in an image.

This could be investigated, perhaps with digit localization experiments where the digit is barely discernable from the background in the image. The results of classifiers trained with even pale RGB color information, barely visible to humans, indicates that a machine learning classifier might do well in such situations and be able to offer additional information that would improve human performance. A similar study could be done with medical images that are difficult to interpret even when a potential target is localized (e.g., mammography, fracture detection, etc.)

6.3 Final Thoughts

The medical community has known for some time, since at least 1993 when Krupinski published her early work detailing the negative effects of hard highlighting on target detection [111], that simply placing a marker or drawing a box around a region to draw attention to it is potentially catastrophic in terms of missed high-cost targets. Yet, this approach still seems to be the norm. This effect seems general and one could imagine it showing up in other critical tasks like baggage handling or automatic target recognition systems for the military where a missed target could also be a matter of life or death. It is hoped that this dissertation is able to contribute in some small way towards addressing these concerns by initiating an area of research into new ways to present classifier output to humans so that the strengths of each are maximized while their deficiencies are minimized.

Bibliography

[1] Jean-Francois Abramatic and Oliver D Faugeras. Sequential convolution techniques for image filtering. Acoustics, Speech and Signal Processing, IEEE Transactions on, 30(1):1–10, 1982.

[2] Ahmed S Abutaleb. Automatic thresholding of gray-level pictures using two-dimensional entropy. Computer vision, graphics, and image processing, 47(1):22–32, 1989.

[3] Holger G Adelmann. Butterworth equations for homomorphic filtering of images. Computers in Biology and Medicine, 28(2):169–181, 1998.

[4] Eugenio Alberdi, Andrey Povyakalo, Lorenzo Strigini, and Peter Ayton. Effects of incorrect computer-aided detection (cad) output on human decision-making in mammography. Academic radiology, 11(8):909–918, 2004.

[5] Eugenio Alberdi, Andrey A Povyakalo, Lorenzo Strigini, Peter Ayton, and Rosalind Given-Wilson. Cad in mammography: lesion-level versus case-level analysis of the effects of prompts on human decisions. International Journal of Computer Assisted Radiology and Surgery, 3(1-2):115–122, 2008.

[6] Douglas G Altman. Practical statistics for medical research. CRC press, 1990.

[7] M Analoui. Radiographic image enhancement. part i: spatial domain techniques. Dentomaxillofacial Radiology, 30(1):1–9, 2001.

[8] James R Anderson. Land-use classification schemes. Photogrammetric Engineering, 1971.

[9] Marcelo E Andia, Johannes Plett, Cristian Tejos, Marcelo W Guarini, María E Navarro, Dravna Razmilic, Luis Meneses, Manuel J Villalon, and Pablo Irarrazaval. Enhancement of visual perception with use of dynamic cues 1. Radiology, 250(2):551–557, 2009.

[10] Harry C Andrews. Monochrome digital image enhancement. Applied optics, 15(2):495–503, 1976.

[11] Ronald L Arenson, SB Seshadri, HL Kundel, D DeSimone, F Van der Voorde, WB Gefter, DM Epstein, WT Miller, JM Aronchick, and MB Simson. Clinical evaluation of a medical image management system for chest images. American Journal of Roentgenology, 150(1):55–59, 1988.

[12] Edward Argyle and A Rosenfeld. Techniques for edge detection. Proceedings of the IEEE, 59(2):285–287, 1971.

[13] Samuel G Armato, Maryellen L Giger, Catherine J Moran, James T Blackburn, Kunio Doi, and Heber MacMahon. Computerized detection of pulmonary nodules on ct scans 1. Radiographics, 19(5):1303–1311, 1999.

[14] Kazuo Awai, Kohei Murao, Akio Ozawa, Masanori Komi, Haruo Hayakawa, Shinichi Hori, and Yasumasa Nishimura. Pulmonary nodules at chest ct: Effect of computer-aided diagnosis on radiologists detection performance 1. Radiology, 230(2):347–352, 2004.

[15] MA Badamchizadeh and A Aghagolzadeh. Comparative study of unsharp masking methods for image enhancement. In Multi-Agent Security and Survivability, 2004 IEEE First Symposium on, pages 27–30. IEEE, 2004.

[16] Reynold Bailey, Ann McNamara, Nisha Sudarsanam, and Cindy Grimm. Subtle gaze direction. ACM Transactions on Graphics (TOG), 28(4):100, 2009.

[17] Jay A Baker, Eric L Rosen, Joseph Y Lo, Edgardo I Gimenez, Ruth Walsh, and Mary Scott Soo. Computer-aided detection (cad) in screening mammography: sensitivity of commercial cad systems for detecting architectural distortion. American Journal of Roentgenology, 181(4):1083–1088, 2003.

[18] Corinne Balleyguier, Karen Kinkel, Jacques Fermanian, Sebastien Malan, Germaine Djen, Patrice Taourel, and Olivier Helenon. Computer-aided detection (cad) in mammography: Does it help the junior or the senior radiologist? European journal of radiology, 54(1):90–96, 2005.

[19] Isaac Bankman. Handbook of medical image processing and analysis. 2008.

[20] Stephen Barrass and Gregory Kramer. Using sonification. Multimedia systems, 7(1):23–31, 1999.

[21] Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[22] Lucy Bastin, Peter F Fisher, and Jo Wood. Visualizing uncertainty in multi-spectral remotely sensed imagery. Computers & Geosciences, 28(3):337–350, 2002.

[23] Nicholas Edward Bearman. Using sound to represent uncertainty in spatial data. 2013.

[24] Nick Bearman and Peter F Fisher. Using sound to represent spatial data in arcgis. Computers & Geosciences, 46:157–163, 2012.

[25] Nick Bearman and Andrew Lovett. Using sound to represent positional accuracy of address locations. The Cartographic Journal, 47(4):308–314, 2010.

[26] Thomas Blaschke. Object based image analysis for remote sensing. ISPRS journal of photogrammetry and remote sensing, 65(1):2–16, 2010.

[27] Isabelle Bloch. Fuzzy spatial relationships for image processing and interpretation: a review. Image and Vision Computing, 23(2):89–110, 2005.

[28] Ivona Brajevic and Milan Tuba. Cuckoo search and firefly algorithm applied to multilevel image thresholding. Cuckoo Search and Firefly Algorithm, pages 115–139, 2014.

[29] Glenn Brauen and DR Fraser Taylor. Linked audio representation in cybercartography: Guidance from animated and interactive cartography for using sound. Revista Brasileira de Cartografia, (60/3), 2009.

[30] John Canny. A computational approach to edge detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on, (6):679–698, 1986.

[31] Morton J Canty. Image analysis, classification and change detection in remote sensing: With algorithms for envi/idl and python. 2014.

[32] Sébastien Caquard, Glenn Brauen, Benjamin Wright, and Paul Jasen. Designing sound in cybercartography: from structured cinematic narratives to unpredictable sound/image interactions. International Journal of Geographical Information Science, 22(11-12):1219–1245, 2008.

[33] Francine Catté, Pierre-Louis Lions, Jean-Michel Morel, and Tomeu Coll. Image selective smoothing and edge detection by nonlinear diffusion. SIAM Journal on Numerical analysis, 29(1):182–193, 1992.

[34] M Emre Celebi and Quan Wen. Variance-cut: A fast color quantization method based on hierarchical clustering. In Electronics, Computer and Computation (ICECCO), 2013 International Conference on, pages 103–106. IEEE, 2013.

[35] Heang-Ping Chan, Berkman Sahiner, Mark A Helvie, Nicholas Petrick, Marilyn A Roubidoux, Todd E Wilson, Dorit D Adler, Chintana Paramagul, Joel S Newman, and Sethumadavan Sanjay-Gopal. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: An roc study 1. Radiology, 212(3):817–827, 1999.

[36] Chip-Hong Chang, Pengfei Xu, Rui Xiao, and Thambipillai Srikanthan. New adaptive color quantization method based on self-organizing maps. Neural Networks, IEEE Transactions on, 16(1):237–249, 2005.

[37] Tao Chen, Kai-Kuang Ma, and Li-Hui Chen. Tri-state median filter for image denoising. Image Processing, IEEE Transactions on, 8(12):1834–1838, 1999.

[38] Heng-Da Cheng, Xiaopeng Cai, Xiaowei Chen, Liming Hu, and Xueling Lou. Computer-aided detection and classification of microcalcifications in mammograms: a survey. Pattern recognition, 36(12):2967–2991, 2003.

[39] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu. Global contrast based salient region detection. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 409–416. IEEE, 2011.

[40] Shao-Yi Chien and Liang-Gee Chen. Reconfigurable morphological image processing accelerator for video object segmentation. Journal of Signal Processing Systems, 62(1):77–96, 2011.

[41] Roland T Chin and Chia-Lung Yeh. Quantitative evaluation of some edge-preserving noise-smoothing techniques. Computer Vision, Graphics, and Image Processing, 23(1):67–91, 1983.

[42] B Chitprasert and KR Rao. Discrete cosine transform filtering. In Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on, pages 1281–1284. IEEE, 1990.

[43] Klaus Christoffersen, David D Woods, and George T Blike. Discovering the events expert practitioners extract from dynamic data streams: the modified unit marking technique. Cognition, Technology & Work, 9(2):81–98, 2007.

[44] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. Torch7: A matlab-like environment for machine learning. In BigLearn, NIPS Workshop, number EPFL-CONF-192376, 2011.

[45] Mary L Comer and Edward J Delp. Morphological operations for color image processing. Journal of electronic imaging, 8(3):279–289, 1999.

[46] Russell G Congalton. A review of assessing the accuracy of classifications of remotely sensed data. Remote sensing of environment, 37(1):35–46, 1991.

[47] Larry S Davis. A survey of edge detection techniques. Computer graphics and image processing, 4(3):248–270, 1975.

[48] Stephanie Deitrick and Robert Edsall. The influence of uncertainty visualization on decision making: An empirical evaluation. Progress in spatial data handling, pages 719–738, 2006.

[49] Stephanie A Deitrick. Uncertainty visualization and decision making: Does visualizing uncertain information change decisions. In Proceedings of the XXIII International Cartographic Conference, pages 4–10, 2007.

[50] Jia Deng, Jonathan Krause, and Li Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pages 580–587. IEEE, 2013.

[51] Yining Deng, Charles Kenney, Michael S Moore, and BS Manjunath. Peer group filtering and perceptual color image quantization. In Circuits and Systems, 1999. ISCAS'99. Proceedings of the 1999 IEEE International Symposium on, volume 4, pages 21–24. IEEE, 1999.

[52] Yining Deng and BS Manjunath. Unsupervised segmentation of color-texture regions in images and video. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23(8):800–810, 2001.

[53] Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized medical imaging and graphics, 31(4):198–211, 2007.

[54] Piotr Dollár and C Zitnick. Fast edge detection using structured forests. 2014.

[55] Trafton Drew, Corbin Cunningham, and Jeremy M Wolfe. When and why might a computer-aided detection (cad) system interfere with visual search? an eye-tracking study. Academic radiology, 19(10):1260–1267, 2012.

[56] RB DAugustino and ES Pearson. Testing for departures from normality. Biometrika, 60:613–622, 1973.

[57] Michael Egmont-Petersen, Dick de Ridder, and Heinz Handels. Image processing with neural networksa review. Pattern recognition, 35(10):2279–2301, 2002.

[58] Chun-Nian Fan and Fu-Yan Zhang. Homomorphic filtering based illumination normalization method for face recognition. Pattern Recognition Letters, 32(10):1468–1479, 2011.

[59] Jiu-Lun Fan and Bo Lei. A modified valley-emphasis method for automatic thresholding. Pattern Recognition Letters, 33(6):703–708, 2012.

[60] Peter F Fisher. Hearing the reliability in classified remotely sensed images. Cartography and Geographic Information Systems, 21(1):31–36, 1994.

[61] John H Flowers, Dion C Buhman, and Kimberly D Turnage. Data sonification from the desktop: Should sound be part of standard data analysis software? ACM Transactions on Applied Perception (TAP), 2(4):467–472, 2005.

[62] Keith M Franklin and Jonathan C Roberts. A path based model for sonification. In Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on, pages 865–870. IEEE, 2004.

[63] Matthew T Freedman, Shih-Chung B Lo, Teresa Osicka, Fleming YM Lure, Xin-Wei Xu, Jesse Lin, Hui Zhao, and Ron Zhang. Computer-aided detection of lung cancer on chest radiographs: effect of machine cad false-positive locations on radiologists' behavior. Medical Imaging 2002, pages 1311–1319, 2002.

[64] Matthew T Freedman and Teresa Osicka. Heat maps: an aid for data analysis and understanding of roc cad experiments. Academic radiology, 15(2):249–259, 2008.

[65] Timothy W Freer and Michael J Ulissey. Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center 1. Radiology, 220(3):781–786, 2001.

[66] Werner Frei. Image enhancement by histogram hyperbolization. Computer Graphics and Image Processing, 6(3):286–294, 1977.

[67] Fujifilm. Fujifilm digital mammography cad. Promotional brochure, 2014.

[68] Yong Ge, Sanping Li, V Chris Lakhan, and Arko Lucieer. Exploring uncertainty in remotely sensed data with parallel coordinate plots. International Journal of Applied Earth Observation and Geoinformation, 11(6):413–422, 2009.

[69] Michael Gervautz and Werner Purgathofer. A simple method for color quantization: Octree quantization. New trends in computer graphics, pages 219–231, 1988.

[70] Theo Gevers and Arnold WM Smeulders. Color-based object recognition. Pattern recognition, 32(3):453–464, 1999.

[71] Ashish Ghosh, Badri Narayan Subudhi, and Susmita Ghosh. Object detection from videos captured by moving camera by fuzzy edge incorporated markov random field and local histogram matching. Circuits and Systems for Video Technology, IEEE Transactions on, 22(8):1127–1135, 2012.

[72] Maryellen L Giger, Kunio Doi, H MacMahon, RM Nishikawa, KR Hoffmann, CJ Vyborny, RA Schmidt, H Jia, K Abe, and X Chen. An" intelligent" workstation for computer-aided diagnosis. Radiographics, 13(3):647–656, 1993.

[73] Fiona J Gilbert, Susan M Astley, CR Boggis, Magnus A McGee, Pamela M Griffiths, Stephen W Duffy, Olorunsola F Agbaje, Maureen GC Gillan, Mary Wilson, Anil K Jain, et al. Variable size computer-aided detection prompts and mammography film reader decisions. Breast Cancer Research, 10(4):R72, 2008.

[74] Myrna CB Godoy, Tae Jung Kim, Charles S White, Luca Bogoni, Patricia de Groot, Charles Florin, Nancy Obuchowski, James S Babb, Marcos Salganicoff, David P Naidich, et al. Benefit of computer-aided detection analysis for the detection of subsolid and solid lung nodules on thin-and thick-section ct. American Journal of Roentgenology, 200(1):74–83, 2013.

[75] RC Gonzales and BA Fittes. Gray-level transformations for interactive image enhancement. mechanism and machine theory, 12(1):111–122, 1977.

[76] Apurba Gorai and Ashish Ghosh. Gray-level image enhancement by particle swarm optimization. In Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on, pages 72–77. IEEE, 2009.

[77] Goesta H Granlund. In search of a general picture processing operator. Computer Graphics and Image Processing, 8(2):155–173, 1978.

[78] DM Green and JA Swets. Signal detection theory and psychophysics. 1966. New York.

[79] Thomas Gruber. What is an ontology? `http://www-ksl.stanford.edu/kst/what-is-an-ontology.html`, 2009. [Online; accessed 18-March-2015].

[80] David Gur, Jules H Sumkin, Howard E Rockette, Marie Ganott, Christiane Hakim, Lara Hardesty, William R Poller, Ratan Shah, and Luisa Wallace. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. Journal of the National Cancer Institute, 96(3):185–190, 2004.

[81] Ernest L Hall. Almost uniform distributions for computer image enhancement. Computers, IEEE Transactions on, 100(2):207–208, 1974.

[82] Paul Heckbert. Color image quantization for frame buffer display. ACM Siggraph Computer Graphics, 16(3):297–307, 1982.

[83] Mark A Helvie, Lubomir Hadjiiski, Erini Makariou, Heang-Ping Chan, Nicholas Petrick, Berkman Sahiner, Shih-Chung B Lo, Matthew Freedman, Dorit Adler, Janet Bailey, et al. Sensitivity of noncommercial computer-aided detection system for mammographic breast cancer detection: Pilot clinical trial 1. Radiology, 231(1):208–214, 2004.

[84] Tomislav Hengl and Norair Toomanian. Maps are not what they seem: representing uncertainty in soil-property maps. In Proc. Accuracy, pages 805–813, 2006.

[85] Thomas Hermann. Sonification for exploratory data analysis. 2002.

[86] Thomas Hermann, T Nattkemper, Walter Schubert, and Helge Ritter. Sonification of multichannel image data. Proc. of the Mathematical and Engineering Techniques in Medical and Biological Sciences (METMBS 2000), pages 745–750, 2000.

[87] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580, 2012.

[88] Clemens Holzhüter, H Schulz, and Heidrun Schumann. Enriched heatmaps for visualizing uncertainty in microarray data. In Poster at the Eurographics Workshop on Visual Computing for Biomedicine (VCBM10), 2010.

[89] G Houle and E Dubois. Quantization of color images for display on graphics terminals. In Proc. IEEE Global Telecomun. Conf., GLOBE-COM86, pages 1138–1142, 1986.

[90] Yu-Zhe Hsiao and Soo-Chang Pei. Edge detection, color quantization, segmentation, texture removal, and noise reduction of color image using quaternion iterative filtering. Journal of Electronic Imaging, 23(4):043001–043001, 2014.

[91] Sheen Hsieh and Kuo-Chin Fan. An adaptive clustering algorithm for color quantization. Pattern Recognition Letters, 21(4):337–346, 2000.

[92] Robert Hummel. Image enhancement by histogram transformation. Computer graphics and image processing, 6(2):184–195, 1977.

[93] Julie A Jackson and Patrick Brady. Radar target classification using morphological image processing. SPIE Defense, Security, and Sensing, pages 805114–805114, 2011.

[94] Nathaniel Jacobson and Walter Bender. Strategies for selecting a fixed palette of colors. OE/LASE'89, 15-20 Jan., Los Angeles. CA, pages 333–341, 1989.

[95] Halszka Jarodzka, Katharina Scheiter, Peter Gerjets, and Tamara Van Gog. In the eyes of the beholder: How experts and novices interpret dynamic stimuli. Learning and Instruction, 20(2):146–154, 2010.

[96] Halszka Jarodzka, Tamara van Gog, Michael Dorr, Katharina Scheiter, and Peter Gerjets. Learning to see: Guiding students' attention via a model's eye movements fosters learning. Learning and Instruction, 25:62–70, 2013.

[97] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093, 2014.

[98] George H Joblove and Donald Greenberg. Color spaces for computer graphics. In ACM siggraph computer graphics, volume 12, pages 20–25. ACM, 1978.

[99] Willie D Jones. Sight for sore ears [visual aids for the blind]. Spectrum, IEEE, 41(2):13–14, 2004.

[100] Keechul Jung, Kwang In Kim, and Anil K Jain. Text information extraction in images and video: a survey. Pattern recognition, 37(5):977–997, 2004.

[101] Shaun K Kane, Jeffrey P Bigham, and Jacob O Wobbrock. Slide rule: making mobile touch screens accessible to blind people using multi-touch interaction techniques. In Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility, pages 73–80. ACM, 2008.

[102] V Karathanassi, P Kolokousis, and S Ioannidou. A comparison study on fusion methods using evaluation indicators. International Journal of Remote Sensing, 28(10):2309–2341, 2007.

[103] Manpreet Kaur, Jasdeep Kaur, and Jappreet Kaur. Survey of contrast enhancement techniques based on histogram equalization. IJACSA) International Journal of Advanced Computer Science and Applications, 2(7), 2011.

[104] Tom Kimpe and Tom Tuytschaever. Increasing the number of gray shades in medical display systemshow much is enough? Journal of digital imaging, 20(4):422–432, 2007.

[105] Josef Kittler, John Illingworth, and J Föglein. Threshold selection based on a simple image statistic. Computer vision, graphics, and image processing, 30(2):125–147, 1985.

[106] Ronald T Kneusel and Peter N Kneusel. Novel pet/ct image fusion via gram-schmidt spectral sharpening. In SPIE Medical Imaging, pages 86692Y–86692Y. International Society for Optics and Photonics, 2013.

[107] Takeshi Kobayashi, Xin-Wei Xu, Heber MacMahon, Charles E Metz, and Kunio Doi. Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. Radiology, 199(3):843–848, 1996.

[108] Ivan Kopecek and Radek Oslejsek. Hybrid approach to sonification of color images. In Convergence and Hybrid Information Technology, 2008. ICCIT'08. Third International Conference on, volume 2, pages 722–727. IEEE, 2008.

[109] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[110] Elizabeth A Krupinski. Visual scanning patterns of radiologists searching mammograms. Academic radiology, 3(2):137–144, 1996.

[111] Elizabeth A Krupinski, Calvin F Nodine, and Harold L Kundel. Perceptual enhancement of tumor targets in chest x-ray images. Perception & psychophysics, 53(5):519–526, 1993.

[112] Elizabeth A Krupinski, Calvin F Nodine, and Harold L Kundel. Enhancing recognition of lesions in radiographic images using perceptual feedback. Optical Engineering, 37(3):813–818, 1998.

[113] Harold L Kundel, Calvin F Nodine, and ELIZABETH A KRUPINSKI. Computer-displayed eye position as a visual aid to pulmonary nodule interpretation. Investigative radiology, 25(8):890–896, 1990.

[114] Phaedon C Kyriakidis. Towards a systems approach to the visualization of spatial uncertainty. In UCGIS Workshop: Geospatial Visualization and Knowledge Discovery Workshop, 2003.

[115] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[116] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 609–616. ACM, 2009.

[117] Jong-Sen Lee. Digital image enhancement and noise filtering by use of local statistics. Pattern Analysis and Machine Intelligence, IEEE Transactions on, (2):165–168, 1980.

[118] San-Kan Lee, Chien-Shun Lo, Chuin-Mu Wang, Pau-Choo Chung, Chein-I Chang, Ching-Wen Yang, and Pi-Chang Hsu. A computer-aided design mammography screening system for detection and classification of microcalcifications. International journal of medical informatics, 60(1):29–57, 2000.

[119] JM Lesniak, R Hupse, R Blanc, N Karssemeijer, and G Székely. Comparative evaluation of support vector machine classification for computer aided detection of breast masses in mammography. Physics in medicine and biology, 57(16):5295, 2012.

[120] Jacob Levman, Tony Leung, Petrina Causer, Don Plewes, and Anne L Martel. Classification of dynamic contrast-enhanced magnetic resonance breast lesions by support vector machines. Medical Imaging, IEEE Transactions on, 27(5):688–696, 2008.

[121] Xia Li, Menno-Jan Kraak, and Zhiming Ma. Towards visual representations to express uncertainty in temporal geodata. Analysis, 6(9):3, 2007.

[122] Jae S Lim. Two-dimensional signal and image processing. Englewood Cliffs, NJ, Prentice Hall, 1990, 710 p., 1, 1990.

[123] Damien Litchfield, Linden J Ball, Tim Donovan, David J Manning, and Trevor Crawford. Viewing another person's eye movements improves identification of pulmonary nodules in chest x-ray inspection. Journal of Experimental Psychology: Applied, 16(3):251, 2010.

[124] Leh-Nien D Loo, Kunio Doi, and Charles E Metz. Investigation of basic imaging properties in digital radiography. 4. effect of unsharp masking on the detectability of simple patterns. Medical physics, 12(2):209–214, 1985.

[125] L Luccheseyz and SK Mitray. Color image segmentation: A state-of-the-art survey. Proceedings of the Indian National Science Academy (INSA-A), 67(2):207–221, 2001.

[126] Arko Lucieer. Uncertainties in segmentation and their visualisation. 2004.

[127] Arko Lucieer and Menno-Jan Kraak. Interactive and visual fuzzy classification of remotely sensed imagery for exploration of uncertainty. International Journal of Geographical Information Science, 18(5):491–512, 2004.

[128] Alan M MacEachren, Anthony Robinson, Susan Hopper, Steven Gardner, Robert Murray, Mark Gahegan, and Elisabeth Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. Cartography and Geographic Information Science, 32(3):139–160, 2005.

[129] Ryan MacVeigh and Daniel Jacobson. Increasing the dimensionality of a geographic information system (gis) using auditory display. 2007.

[130] Mark T Madsen and Chan H Park. Enhancement of spect images by fourier filtering the projection image set. Journal of nuclear medicine: official publication, Society of Nuclear Medicine, 26(4):395–402, 1985.

[131] Raman Maini and Himanshu Aggarwal. Study and comparison of various image edge detection techniques. International journal of image processing (IJIP), 3(1):1–11, 2009.

[132] Petros Maragos. Tutorial on advances in morphological image processing and analysis. Optical engineering, 26(7):267623–267623, 1987.

[133] Petros Maragos. A representation theory for morphological image and signal processing. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 11(6):586–599, 1989.

[134] Kanti V. Mardia and TJ Hainsworth. A spatial thresholding method for image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 10(6):919–927, 1988.

[135] David Marr and Ellen Hildreth. Theory of edge detection. Proceedings of the Royal Society of London. Series B. Biological Sciences, 207(1167):187–217, 1980.

[136] Antonio Cesar Germano Martins, Rangaraj Mandayam Rangayyan, Luis Antonio Portela, E Amaro Jr, and Ruggero Andrea Ruschioni. Auditory display and sonification of textured images. In Proceedings of the Third International Conference on Auditory Display ICAD, volume 96, 1996.

[137] Christiane Marx, Ansgar Malich, Mirjam Facius, Uta Grebenstein, Dieter Sauner, Stefan OR Pfleiderer, and Werner A Kaiser. Are unnecessary follow-up procedures induced by computer-aided diagnosis (cad) in mammography? comparison of mammographic diagnosis with and without use of cad. European journal of radiology, 51(1):66–72, 2004.

[138] Michael EJ Masson and Geoffrey R Loftus. Using confidence intervals for graphically based data interpretation. Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 57(3):203, 2003.

[139] Patricia Melin, Claudia I Gonzalez, Juan R Castro, Olivia Mendoza, and Oscar Castillo. Edge-detection method for image processing based on generalized type-2 fuzzy logic. Fuzzy Systems, IEEE Transactions on, 22(6):1515–1525, 2014.

[140] Claudia Mello-Thoms. How does the perception of a lesion influence visual search strategy in mammogram reading? Academic radiology, 13(3):275–288, 2006.

[141] Arpita Mittal and Sanjay Kumar Dubey. Analysis of mri images of rheumatoid arthritis through morphological image processing techniques. IJCS, 10(2-3):118–122, 2013.

[142] MA Mohamed Ali, RJ Toomey, JT Ryan, FC Cuffe, and PC Brennan. A novel teaching tool using dynamic cues improves visualisation of chest lesions by naive observers. In Proc. of SPIE Vol, volume 7263, pages 726304–1, 2009.

[143] Giorgos Mountrakis, Jungho Im, and Caesar Ogole. Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing, 66(3):247–259, 2011.

[144] Serafeim Moustakidis, Giorgos Mallinis, Nikos Koutsias, John B Theocharis, and Vassilios Petridis. Svm-based fuzzy decision trees for classification of high spatial resolution remote sensing images. Geoscience and Remote Sensing, IEEE Transactions on, 50(1):149–169, 2012.

[145] Janne J Näppi. Cade prompts and observer performance: A game of confidence. Academic radiology, 17(8):945–947, 2010.

[146] Patrenahalli M Narendra. A separable median filter for image noise smoothing. Pattern Analysis and Machine Intelligence, IEEE Transactions on, (1):20–29, 1981.

[147] Nasser M Nasrabadi and Robert A King. Image coding using vector quantization: A review. Communications, IEEE Transactions on, 36(8):957–971, 1988.

[148] Hui-Fuang Ng, Davaajargal Jargalsaikhan, Hao-Chuan Tsai, and Chih-Yang Lin. An improved method for image thresholding based on the valley-emphasis method. In Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific, pages 1–4. IEEE, 2013.

[149] Robert M Nishikawa and Andriy Bandos. Predicting the benefit of using cade in screening mammography. Breast Imaging, pages 44–49, 2014.

[150] Robert M Nishikawa and Maria Kallergi. 6.3. computer-aided detection, in its present form, is not an effective aid for screening mammography. Colin G. Orton and William R. Hendee, page 253, 2008.

[151] Eisaku Oho, Norio Baba, Masaru Katoh, Takashi Nagatani, Masako Osumi, Kazunobu Amako, and Koichi Kanaya. Application of the laplacian filter to high-resolution enhancement of sem images. Journal of Electron Microscopy Technique, 1(4):331–340, 1984.

[152] Alan V Oppenheim and Ronald W Schafer. From frequency to quefrency: A history of the cepstrum. Signal Processing Magazine, IEEE, 21(5):95–106, 2004.

[153] Michael T Orchard and Charles A Bouman. Color quantization of images. Signal Processing, IEEE Transactions on, 39(12):2677–2690, 1991.

[154] Stanley Osher and Leonid I Rudin. Feature-oriented image enhancement using shock filters. SIAM Journal on Numerical Analysis, 27(4):919–940, 1990.

[155] Nobuyuki Otsu. A threshold selection method from gray-level histograms. Automatica, 11(285-296):23–27, 1979.

[156] Nikhil R Pal and Sankar K Pal. A review on image segmentation techniques. Pattern recognition, 26(9):1277–1294, 1993.

[157] Sankar K Pal and R King. Image enhancement using smoothing with fuzzy sets. IEEE TRANS. SYS., MAN, AND CYBER., 11(7):494–500, 1981.

[158] Athanasios Papadopoulos, Dimitrios I. Fotiadis, and Aristidis Likas. An automatic microcalcification detection system based on a hybrid neural network classifier. Artificial intelligence in Medicine, 25(2):149–167, 2002.

[159] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: a technical overview. Signal Processing Magazine, IEEE, 20(3):21–36, 2003.

[160] Emerson Carlos Pedrino, José Hiroki Saito, and Valentin Obac Roda. A genetic programming approach to reconfigure a morphological image processing architecture. International Journal of Reconfigurable Computing, 2011:5, 2011.

[161] Eli Peli. Contrast in complex images. JOSA A, 7(10):2032–2040, 1990.

[162] Liane E Philpotts. Can computer-aided detection be detrimental to mammographic interpretation? 1. Radiology, 253(1):17–22, 2009.

[163] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing, 39(3):355–368, 1987.

[164] Johannes Plett, Marcello Guarini, and Pablo Irarrazaval. Enhancement of visual perception through dynamic cues: An application to mammograms. In Image Processing, 2007. ICIP 2007. IEEE International Conference on, volume 2, pages II–441. IEEE, 2007.

[165] Andrea Polesel, Giovanni Ramponi, and V John Mathews. Image enhancement via adaptive unsharp masking. IEEE transactions on image processing, 9(3):505–510, 2000.

[166] Thomas Porter and Tom Duff. Compositing digital images. In ACM Siggraph Computer Graphics, volume 18, pages 253–259. ACM, 1984.

[167] Judith MS Prewitt. Object enhancement and extraction. Picture processing and Psychopictorics, 10(1):15–19, 1970.

[168] Xiaoting Pu, Zhenhong Jia, Liejun Wang, Yingjie Hu, and Jie Yang. The remote sensing image enhancement based on nonsubsampled contourlet transform and unsharp masking. Concurrency and Computation: Practice and Experience, 26(3):742–747, 2014.

[169] Emma Pun, W F Eddie Lau, Robin Cassumbhoy, Anthony J Taranto, and Alexander G Pitman. Clinical experience of the first digital mammographic unit in australia in its first year of use. Med J Aust, 187(10):576–579, 2007.

[170] Giovanni Ramponi. A cubic unsharp masking technique for contrast enhancement. Signal Processing, 67(2):211–222, 1998.

[171] Giovanni Ramponi, Norbert K Strobel, Sanjit K Mitra, and Tian-Hu Yu. Nonlinear unsharp masking methods for image contrast enhancement. Journal of Electronic Imaging, 5(3):353–366, 1996.

[172] Rangaraj M Rangayyan, Antonio CG Martins, and Ruggero A Ruschioni. Aural analysis of image texture via cepstral filtering and sonification. Electronic Imaging: Science & Technology, pages 283–294, 1996.

[173] Alessandra Retico. Computer-aided detection for pulmonary nodule identification: improving the radiologist's performance? Imaging in Medicine, 5(3):249–263, 2013.

[174] John A Richards and JA Richards. Remote sensing digital image analysis. 1999.

[175] Gerhard X. Ritter, Joseph N. Wilson, and Jennifer L Davidson. Image algebra: An overview. Computer Vision, Graphics, and Image Processing, 49(3):297–331, 1990.

[176] Lawrence Gilman Roberts. Machine perception of three-dimensional soups. 1963.

[177] Jimmy Roehrig and Ronald A Castellino. The promise of computer aided detection in digital mammography. European Journal of Radiology, 31(1):35–39, 1999.

[178] Azriel Rosenfeld and Mark Thurston. Edge and curve detection for visual scene analysis. Computers, IEEE Transactions on, 100(5):562–569, 1971.

[179] John C Russ and Roger P Woods. The image processing handbook. Journal of Computer Assisted Tomography, 19(6):979–981, 1995.

[180] Urs E Ruttimann, Richard L Webber, and Edgar Schmidt. A robust digital method for film contrast correction in subtraction radiography. Journal of Periodontal Research, 21(5):486–495, 1986.

[181] Berkman Sahiner, Heang-Ping Chan, Lubomir M Hadjiiski, Philip N Cascade, Ella A Kazerooni, Aamer R Chughtai, Chad Poopat, Thomas Song, Luba Frank, Jadranka Stojanovska, et al. Effect of cad on radiologists' detection of lung nodules on thoracic ct scans: analysis of an observer performance study by nodule size. Academic radiology, 16(12):1518–1530, 2009.

[182] Prasanna K Sahoo, SAKC Soltani, and Andrew KC Wong. A survey of thresholding techniques. Computer vision, graphics, and image processing, 41(2):233–260, 1988.

[183] Mehul P Sampat, Mia K Markey, and Alan C Bovik. Computer-aided detection and diagnosis in mammography. Handbook of image and video processing, 2(1):1195–1217, 2005.

[184] Maurice Samulski, A Hupse, Carla Boetes, G den Heeten, and Nico Karssemeijer. Analysis of probed regions in an interactive cad system for the detection of masses in mammograms. In SPIE Medical Imaging, pages 726314–726314. International Society for Optics and Photonics, 2009.

[185] Maurice Samulski, Rianne Hupse, Carla Boetes, Roel DM Mus, Gerard J den Heeten, and Nico Karssemeijer. Using computer-aided detection in mammography as a decision support. European radiology, 20(10):2323–2330, 2010.

[186] Maurice René Marina Samulski. Computer aided detection as a decision aid in medical screening. 2011.

[187] Stephen J Sangwine and Robin EN Horne. The colour image processing handbook. 1998.

[188] Rajib Sarkar, Sambit Bakshi, and Pankaj K Sa. Review on image sonification: a non-visual scene representation. In Recent Advances in Information Technology (RAIT), 2012 1st International Conference on, pages 86–90. IEEE, 2012.

[189] N Senthilkumaran and R Rajesh. Edge detection techniques for image segmentation–a survey of soft computing approaches. International journal of recent trends in engineering, 1(2), 2009.

[190] Jean Serra. Morphological filtering: an overview. Signal processing, 38(1):3–11, 1994.

[191] Jean Serra and Luc Vincent. An overview of morphological filtering. Circuits, Systems and Signal Processing, 11(1):47–108, 1992.

[192] Yang Shao and Ross S Lunetta. Comparison of support vector machine, neural network, and cart algorithms for the land-cover classification using limited training data points. ISPRS Journal of Photogrammetry and Remote Sensing, 70:78–87, 2012.

[193] Huanfeng Shen, Huifang Li, Yan Qian, Liangpei Zhang, and Qiangqiang Yuan. An effective thin cloud removal procedure for visible remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 96:224–235, 2014.

[194] BG Sherlock, DM Monro, and K Millard. Fingerprint enhancement by directional fourier filtering. In Vision, Image and Signal Processing, IEE Proceedings-, volume 141, pages 87–94. IET, 1994.

[195] Ben Shneiderman. Dynamic queries for visual information seeking. Software, IEEE, 11(6):70–77, 1994.

[196] Richard G Shoup. Color table animation. ACM SIGGRAPH Computer Graphics, 13(2):8–13, 1979.

[197] Daniel J Simons and Christopher F Chabris. Gorillas in our midst: Sustained inattentional blindness for dynamic events. Perception-London, 28(9):1059–1074, 1999.

[198] Wladyslaw Skarbek, Andreas Koschan, Technischer Bericht, and Zur Veroffentlichung. Colour image segmentation-a survey. 1994.

[199] Kenneth R Sloan and Christopher M Brown. Color map techniques. Computer Graphics and Image Processing, 10(4):297–317, 1979.

[200] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In Advances in neural information processing systems, pages 2951–2959, 2012.

[201] Irwin Sobel. History and definition of the sobel operator. 2014.

[202] Exelis Visual Information Solutions. Envi. `http://www.exelisvis.com/`, 2000-2015.

[203] Milan Sonka, Vaclav Hlavac, and Roger Boyle. Image processing, analysis, and machine vision. 2014.

[204] Srinivas Sridharan, Reynold Bailey, Ann McNamara, and Cindy Grimm. Subtle gaze manipulation for improved mammography training. In Proceedings of the Symposium on Eye Tracking Research and Applications, pages 75–82. ACM, 2012.

[205] Thomas G Stockham. Image processing in the context of a visual model. Proc. IEEE, 60(7):828–842, 1972.

[206] PM Taylor, J Champness, RM Given-Wilson, HWW Potts, and K Johnston. An evaluation of the impact of computer-based prompts on screen readers' interpretation of mammograms. 2014.

[207] PM Taylor, J Champness, RM Given-Wilson, HWW Potts, and K Johnston. An evaluation of the impact of computer-based prompts on screen readers' interpretation of mammograms. 2014.

[208] Stuart A Taylor, Susan C Charman, Philippe Lefere, Elizabeth G McFarland, Erik K Paulson, Judy Yee, Rizwan Aslam, John M Barlow, Arun Gupta, David H Kim, et al. Ct colonography: Investigation of the optimum reader paradigm by using computer-aided detection software 1. Radiology, 246(2):463–471, 2008.

[209] Matthew J Thurley and Victor Danell. Fast morphological image processing open-source extensions for gpu processing with cuda. Selected Topics in Signal Processing, IEEE Journal of, 6(7):849–855, 2012.

[210] HJ Vala and Astha Baxi. A review on otsu image segmentation algorithm. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2(2):pp–387, 2013.

[211] Rein Van Den Boomgaard and Richard Van Balen. Methods for fast morphological image transforms using bitmapped binary images. CVGIP: Graphical Models and Image Processing, 54(3):252–258, 1992.

[212] Rein Van Den Boomgaard and Richard Van Balen. Methods for fast morphological image transforms using bitmapped binary images. CVGIP: Graphical Models and Image Processing, 54(3):252–258, 1992.

[213] Frans JM Van der Wel, Linda C Van der Gaag, and Ben GH Gorte. Visual exploration of uncertainty in remote-sensing classification. Computers & Geosciences, 24(4):335–343, 1998.

[214] JL Van Genderen, BF Lock, and PA Vass. Remote sensing: statistical testing of thematic map accuracy. Remote Sensing of Environment, 7(1):3–14, 1978.

[215] Luc Vincent. Morphological grayscale reconstruction in image analysis: applications and efficient algorithms. Image Processing, IEEE Transactions on, 2(2):176–201, 1993.

[216] David CC Wang, Anthony H Vagnucci, and CC Li. Digital image enhancement: a survey. Computer Vision, Graphics, and Image Processing, 24(3):363–381, 1983.

[217] Samuel John Welch, Arko Lucieer, and Ray Williams. Interactive visualisation techniques for data mining of satellite imagery. 2007.

[218] Frank Wilcoxon. Individual comparisons by ranking methods. Biometrics bulletin, pages 80–83, 1945.

[219] Xiaoying Wu and Ze-Nian Li. A study of image-based music composition. In Multimedia and Expo, 2008 IEEE International Conference on, pages 1345–1348. IEEE, 2008.

[220] Mussarat Yasmin, Muhammad Sharif, Saleha Masood, Mudassar Raza, and Sajjad Mohsin. Brain image enhancement-a survey. World Applied Sciences Journal, 17(9):1192–1204, 2012.

[221] Danial Yazdani, Hadi Nabizadeh, Elyas Mohamadzadeh Kosari, and Adel Nadjaran Toosi. Color quantization using modified artificial fish swarm algorithm. AI 2011: Advances in Artificial Intelligence, pages 382–391, 2011.

[222] Woon Seung Yeo and Jonathan Berger. A framework for designing image sonification methods. In Proceedings of International Conference on Auditory Display, 2005.

[223] Byoung-Woo Yoon and Woo-Jin Song. Image contrast enhancement based on the generalized histogram. Journal of electronic imaging, 16(3):033005–033005, 2007.

[224] Tsubasa Yoshida, Kris M Kitani, Hideki Koike, Serge Belongie, and Kevin Schlei. Edgesonic: image feature sonification for the visually impaired. In Proceedings of the 2nd Augmented Human International Conference, page 11. ACM, 2011.

[225] XD Yue, DQ Miao, LB Cao, Q Wu, and YF Chen. An efficient color quantization based on generic roughness measure. Pattern Recognition, 47(4):1777–1789, 2014.

[226] Bin Zheng, Marie A Ganott, Cynthia A Britton, Christiane M Hakim, Lara A Hardesty, Thomas S Chang, Howard E Rockette, and David Gur. Soft-copy mammographic readings with different computer-assisted detection cuing environments: Preliminary findings 1. Radiology, 221(3):633–640, 2001.

[227] Bin Zheng, Richard G Swensson, Sara Golla, Christiane M Hakim, Ratan Shah, Luisa Wallace, and David Gur. Detection and classification performance levels of mammographic masses under different computer-aided detection cueing environments 1. Academic radiology, 11(4):398–406, 2004.