

Human Identification of Problematic Handwritten Digits for Pattern Recognition

Nabil Khoury

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements
for the Degree of Masters of Computer Science at
Concordia University
Montréal, Québec, Canada

August 2011

© Nabil Khoury, 2011

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Nabil Houry

Entitled: Human Identification of Problematic Handwritten Digits for Pattern Recognition

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Yuhong Yan

_____ Examiner
Dr. Thomas Fevens

_____ Examiner
Dr. Tonis Kasvand

_____ Supervisor
Dr. Adam Krzyzak

_____ Supervisor
Dr. Ching Y. Suen

Approved by _____
Chair of Department or Graduate Program Director

Dr. Robin A. L. Drew, Dean
Faculty of Engineering and Computer Science

Date _____

ABSTRACT

HUMAN IDENTIFICATION OF PROBLEMATIC HANDWRITTEN DIGITS FOR PATTERN RECOGNITION

Nabil Khoury

After decades of work in pattern recognition, humans are still considered the best recognizers of images and symbols especially in unconstrained everyday applications. This has made the human visual model a major topic of interest in pattern recognition research. A number of studies have presented promising recognition models that incorporate different aspects of the human model such as selective attention, biologically plausible saliency detection and top-down recognition. On the other hand, the last hundred years of research in human eye movement behaviour has revived the ancient philosophical idea that we see in our mind's eye. Several computational models of eye movement control were suggested that successfully predict eye movement behaviour demonstrating a close coupling between eye movements and underlying oculomotor and cognitive processes. In the present study, the author evaluates a combined approach to identifying features of interest for Pattern Recognition applications. In the data collection stage, sixty participants are asked to verbally identify fifty-four problematic and twenty prototypical handwritten digits. Both verbal responses and visual fixations are recorded for further analysis. In the analysis stage, a smaller set of ambiguous digit images is identified based on how often participants change their minds about the numeral they represent. For each digit, visual fixations are grouped based on the numeral that participants called out. Each fixation group is then combined into a single fixation heat map. Results show that by comparing and contrasting heat maps for a given digit the features deemed most disambiguating by the human model can be identified.

ACKNOWLEDGMENTS

The current study was made possible through a concerted collaboration of the Center for Pattern Recognition and Machine Intelligence and the Concordia Vision Lab. I would like to thank Professor Michael von Grünau of Vision Lab for his crucial support and contribution to the design of the eye-tracking experiment, facilitating the ethical approval and participant recruitment process and providing access to the necessary eye-tracking equipment. I would also like to thank Professor Ching Y. Suen for refining and focusing the research objectives to their current form and for his insightful feedback during experiment design. Sincere appreciations go to instructor Afroditi Panagopoulos of Vision Lab for allowing access to potential participants from among her students. Special thanks to Professor Nawwaf Kharma of CENPARMI for suggesting this research topic and last, but not least, to Professor Adam Krzyzak of CENPARMI for his valuable feedback during the preparation of this thesis.

DEDICATIONS

To the peoples of the Arab Spring. Has there ever been greater hope that the legacy of Carthage and Giza, Ma'rib and Ugarit and so many others will soon no longer be *ancient* history

CONTENTS

LIST OF ILLUSTRATIONS	viii
LIST OF TABLES	xii
LIST OF ABBREVIATIONS	xiii
INTRODUCTION	1
Chapter	
1. Background and Literature Review.....	4
Visual Pattern Recognition Progress and Challenges	4
Background in Eye Movements	5
Review of Eye Movement Research.....	10
Eye Movement in Pattern Recognition and Image Analysis	11
2. Present Study.....	21
3. Data Collection.....	24
Lab Setup and Available Tools.....	24
Experiment Design Justification	28
Unconstrained Identification during Normal Viewing	47
Data Collection Process	51
Collected Data Reliability Ranking	53
4. Data Analysis and Discussion.....	55
Preparation of Identification Data.....	55
Preparation of Visual Data.....	60
Analysis of Identification Response Data.....	75
Analysis of Visual Response Data	90

5. Conclusions and Further inquiry	105
Summary and Preliminary Results.....	105
Suggestions for Further Inquiry	109
REFERENCES	111
APPENDIX.....	117
GLOSSARY	169

ILLUSTRATIONS

Figure	Page
1. Early eye movement experiment showing eye scan path.....	8
2. Saliency mapping in eye movement-inspired digit recognition.....	14
3. A top-down augmented recognition model.....	15
4. A bottom-up algorithm defining regions of interest.....	17
5. Common errors made by handwritten digit classifiers.....	21
6. The original seventy-four MNIST digits.....	22
7. Experimental setup at the Concordia Vision Lab.....	24
8. MNIST aliasing and distortions	30
9. MNIST-NIST matching scheme	33
10. Verbal task viewing conditions.....	40
11. Digit luminance contrast across Verbal task viewing conditions	41
12. Summary of collected data and reliability.....	54
13. Output comparison of original and modified EDF2ASC tool.....	63
14. Addition of support columns to facilitate fixation filtering and isolation.....	64
15. Side-by-side comparison of bubble plot and Data Viewer's spatial overlay	67
16. Comparison of Gaussian convolution and Data Viewer's heat maps	70
17. Duration-based heat map and annotation	71
18. Batch-mode selection table	74
19. Identification rates across all viewing conditions of Verbal task.....	77
20. Average response time in Manual and Verbal tasks	79
21. Identification rate in Manual and Verbal tasks	81

22. Discrepancies in identification under Manual and Verbal tasks	84
23. Digit and heat map bounding boxes	90
24. Gaze spatial span in regular and irregular digits across viewing conditions of Verbal task	91
25. Ratio of duration-based to count-based gaze span	93
26. Fixation selection in re-identified digits.....	95
27. Gaze preceding identification as <i>zero</i> and <i>six</i>	96
28. Gaze preceding identification as <i>two</i> and <i>seven</i> I.....	97
29. Gaze preceding identification as <i>two</i> and <i>seven</i> II	98
30. Gaze preceding identification as <i>three</i> and <i>five</i> I	99
31. Gaze preceding identification as <i>three</i> and <i>five</i> II.....	100
32. Gaze preceding identification as <i>four</i> and <i>nine</i> I.....	100
33. Gaze preceding identification as <i>four</i> and <i>nine</i> II	101
A1. Verbal task storyboard in Experiment Builder.....	118
A2. Manual task storyboard in Experiment Builder	119
A3. Session data folder and files under Verbal and Manual tasks.....	120
A4. Participant consent form.....	121
A5. Participant debriefing sheet.....	122
A6. Overview of verbal response isolation tool.....	123
A7. Verbal and Manual task <i>zeros</i>	124
A8. Verbal and Manual task <i>ones</i>	125
A9. Verbal and Manual task <i>twos</i>	126
A10. Verbal and Manual task <i>threes</i>	127
A11. Verbal and Manual task <i>fours</i> I.....	128

A12. Verbal and Manual task <i>fours</i> II.....	129
A13. Verbal and Manual task <i>fives</i>	130
A14. Verbal and Manual task <i>sixes</i> I	131
A15. Verbal and Manual task <i>sixes</i> II	132
A16. Verbal and Manual task <i>sevens</i> I.....	133
A17. Verbal and Manual task <i>sevens</i> II	134
A18. Verbal and Manual task <i>eights</i> I.....	135
A19. Verbal and Manual task <i>eights</i> II	136
A20. Verbal and Manual task <i>nines</i> I.....	137
A21. Verbal and Manual task <i>nines</i> II	138
A22. Side-by-side comparison of Gaussian convolution and Data Viewer's heat maps (large)	139
A23. Side-by-side comparison between count- and duration-based heatmaps from Verbal task.....	140
A24. Individual variations in gaze within and across viewing conditions of Verbal task	141
A25. Response rate across viewing conditions in Verbal task.....	142
A26. Individual variations in correct identification and response time	143
A27. Most identifiable irregular digits in Manual and Verbal tasks (detailed).....	144
A28. Most misidentified irregular digits in Manual and Verbal tasks (detailed).....	145
A29. Most misidentified irregular digits by numeral in Manual and Verbal tasks (detailed).....	146
A30. Most common confusion pairs in Manual and Verbal tasks (detailed).....	147
A31. Most ambiguous irregular digits in Manual and Verbal tasks (detailed)	148
A32. Most re-identified irregular digits in Verbal task (detailed)	149
A33. Gaze preceding identification as <i>zero</i> and <i>six</i> (large).....	162

A34. Gaze preceding identification as *two* and *seven* I (large)..... 163

A35. Gaze preceding identification as *two* and *seven* II (large)..... 164

A36. Gaze preceding identification as *three* and *five* I (large)..... 165

A37. Gaze preceding identification as *three* and *five* II (large) 166

A38. Gaze preceding identification as *four* and *nine* I (large)..... 167

A39. Gaze preceding identification as *four* and *nine* II (large)..... 168

TABLES

Table	Page
1. Most correctly identified digits in Manual and Verbal tasks	82
2. Most misidentified digits in Manual and Verbal tasks.....	83
3. Most misidentified digits by numeral.....	85
4. Most common confusion pairs	86
5. Most confusing digits	87
6. Most re-identified digits	88
A1. Identifiability of irregular digits in Manual task	150
A2. Identifiability of irregular digits in unsmoothed conditions of Verbal task	154
A3. Identifiability of irregular digits in smoothed conditions of Verbal task	158

ABBREVIATIONS

ASC	ASCII or plain text data
BG	Background (of an image)
BU	Bottom-up
<i>bV1</i>	Occurring before first verbal response
<i>bV2</i>	Occurring before second verbal response
C	Correct numeral
EDF	Eye movement data file
ESACF	Enhanced summary autocorrelation function
FG	Foreground (of an image)
FG0	Verbal task viewing condition where the foreground-handwritten digit colour is RGB(0,0,0) and the background colour is set to RGB(240,240,240)
FG120	Verbal task viewing condition where the foreground-handwritten digit colour is RGB(120,120,120) and the background colour is set to RGB(240,240,240)
FG180	Verbal task viewing condition where the foreground-handwritten digit colour is RGB(180,180,180) and the background colour is set to RGB(240,240,240)
FG210	Verbal task viewing condition where the foreground-handwritten digit colour is RGB(210,210,210) and the background colour is set to RGB(240,240,240)
FG228	Verbal task viewing condition where the foreground-handwritten digit colour is RGB(228,228,228) and the background colour is set to RGB(240,240,240)
MNIST	Modified NIST
NIST	National Institute of Standards and Technology
PNG	Portable Network Graphics

ROI	Region of interest (of gaze)
TD	Top-down
<i>tr</i>	Training database
<i>ts</i>	Testing database
<i>V1</i>	First verbal response or occurring during first verbal response
<i>V2</i>	Second verbal response or occurring during second verbal response
<i>V2+</i>	Occurring after second verbal response
VBA	Visual Basic for Applications; a scripting language to program MS Excel macros
VAD	Voice activity detection
XLS	Standard file extension for Microsoft Excel spreadsheet documents.
XLSM	File extension for Microsoft Excel macro-enabled workbooks

INTRODUCTION

After decades of work in pattern recognition, humans are still considered the best recognizers of images and symbols especially in unconstrained everyday applications. This has made the human visual recognition model a major topic of interest in pattern recognition and machine intelligence research (Barriere and Plamondon 1998; Côté and others 1998; Keller and others 1999; Suen and others 2000; Maw and Pomplun 2004). One hundred years of research into human eye movement behaviour, on the other hand, has revived the ancient philosophical idea that we see in our mind's eye. Several computational models of eye movement control were suggested that successfully approximate different aspects of collected human eye movement data and demonstrate a close coupling between eye movements and underlying oculomotor and cognitive processes (Brandt and Stark 1997; Chernyak and Stark 2001; Ojanpää 2006; Paulson and Goodman 1999; Stark and Choi 1996; Reichle, Rayner, and Pollatsek 2003; Reilly and O'Regan 1998). The focus in most of these computational models, however, is on topics of interest in psychology such as psycholinguistics and visual perception. Some computational models in the vision science literature focus on eye movement control during visual search and detection (Rao and others 2002; Rao and others 1996; Zhang and others 2006) While, at least one computational model claims to successfully predict human visual fixations during identification of handwritten Katakana letters (Watanabe, Gyoba, and Maruyama 1983)

Eye movement research inspired a number of studies in image analysis and pattern recognition. Some present scene analysis and visual recognition techniques that

incorporate different characteristics of the human model. Among these, selective attention, biologically plausible saliency detection and top-down recognition have been used with promising results. Some of these techniques exploit the computational efficiency associated with serial recognition (Exel and Pessoa 1998; Itti, Koch, and Niebur 1998; Rybak and others 1998; Salah, Alpaydin, and Akarun 2001; Salah, Alpaydin, and Akarun 2002; Xianglin Meng and Zhengzhi Wang 2009). Others exploit context awareness to guide the recognition process (Chernyak and Stark 2001; Exel and Pessoa 1998). Yet others attempt to predict image coordinates and regions that are most likely to attract human attention (Hacisalihzade, Stark, and Allen 1992; Osberger and Maeder 1998; Privitera and Stark 2000; Yagi, Gouhara, and Uchikawa 1993). However, limitations in biologically inspired saliency detection, and the increased availability of eye tracking equipment motivated a multidisciplinary approach. A number of saliency detection schemes were evaluated, refined or trained using human eye movement and identification data specifically collected for these purposes. These include: (1) algorithms for defining regions of interest in static images (Watanabe, Gyoba, and Maruyama 1983; Privitera and Stark 2000; Yagi, Gouhara, and Uchikawa 1993; Kienzle and others 2007; Schomaker and Segers 1999), (2) task-dependent selective attention in video analysis (Peters and Itti 2007), and (3) identification of informative features in handwriting (Watanabe, Gyoba, and Maruyama 1983; Schomaker and Segers 1999). Despite promising results in detecting features and regions of interest in scenery and video, the author found no mention of similar research in the context of pattern recognition applications.

In the present study, the author evaluates a novel approach that explores the use of human visual fixations and identification data in order to identify features of interest for Pattern Recognition applications. We select handwritten digit recognition as a prototype

application and use seventy-four digit images from the NIST database as stimuli. Fifty-four of these digits are of particular interest because they are reported to be particularly problematic for a variety of classifiers in the literature (Lauer, Suen, and Bloch 2007; Suen and Tan 2005). The other twenty look very prototypical and are used as a reference. In the data collection stage, sixty participants are asked to identify the handwritten digits verbally. Both verbal responses and visual fixations are recorded during the course of the identification task for further analysis. In the analysis stage, a smaller set of ambiguous digit images is identified based on how often participants change their minds about the numeral they represent. For each of these ambiguous digits, visual fixations corresponding to a given response are combined into a single fixation heat map. Preliminary results show that, by comparing and contrasting these maps, the handwritten digit features deemed most disambiguating by the human model can be identified.

We start with an overview of recurrent challenges in the field of visual pattern recognition, to motivate a closer look at the human visual model. We proceed with a presentation of existing methods and computational approaches incorporating different aspects of this model in the eye movement and pattern recognition literatures. A detailed discussion of our data collection methodology and various technical aspects will follow as well as a description of the computational tools we developed to facilitate the use and analysis of collected data. We proceed by ranking the handwritten digits based on a number of difficulty and ambiguity criteria and presenting corresponding fixation heat maps. We conclude with a summary of our findings.

CHAPTER 1

BACKGROUND AND LITERATURE REVIEW

According to Watanabe (1985) a pattern “[is] the opposite of a chaos; it is an entity, vaguely defined, that could be given a name.” More formally, visual pattern recognition is the study of how machines can learn to discern visual patterns of interest from their environment and accurately determine the category to which they belong (Watanabe 1985; Jain, Duin, and Jianchang Mao 2000).

Visual Pattern Recognition Progress and Challenges

The past two decades have seen a resurgence of interest in visual pattern recognition due to the emergence of complex and computationally demanding applications like handwriting recognition, efficient searching of text documents and multimedia databases as well as personal identification based on face and fingerprints (Suen and others 2000; Jain, Duin, and Jianchang Mao 2000). This interest has resulted in significant advances in this research area. In handwriting recognition, for instance, efforts have translated into close to perfect recognition rates in some restricted applications. Yet despite numerous such breakthroughs – thanks to the use of powerful approaches like neural networks, hidden Markov models and support vector machines – many real-life applications of visual pattern recognition remain unreliable. This is partially due to the pitfalls of real-life conditions like optical artefacts and position variations. It is also due to the nature of certain visual patterns exhibiting large within category variations and considerable similarities among different categories. This makes it difficult to define the

most identifying discriminative features for a given category in a reliable way. In fact, in many of the emerging applications, it is clear that no single approach or simple scheme will ever be found that can find such features (Suen and others 2000; Jain, Duin, and Jianchang Mao 2000). The search for an adequate and perhaps multi-faceted scheme is hence ongoing and the human model, widely held as the best existing recognizer thanks largely to its heavy reliance on context, knowledge and experience, is an obvious target of investigation for researchers in this field (Barriere and Plamondon 1998; Côté and others 1998; Suen and others 2000; Jain, Duin, and Jianchang Mao 2000). However, such investigation requires concerted collaboration with other disciplines ranging from biology to cognitive psychology. In these disciplines, human eye movement data has long been used as a metric of processes underlying human visual behaviour (Jain, Duin, and Jianchang Mao 2000). We now turn to a brief background on the human visual system and, in particular, anatomical, behavioural and cognitive aspects that affect human eye movement.

Background in Eye Movements

The human visual system relies on a multi-resolution field of view. At the sensory level, the retina has two kinds of photoreceptors: the rods, sensitive to low illumination, and the cones, sensitive to normal illumination levels. The cones are densely present, and therefore provide higher resolution sampling, at the centre of the field of view (fovea) decreasing rapidly towards the periphery. This decrease in sensory resolution outwards is coupled by an analogous decrease in processing resources represented by the number of neuronal receptive fields and the size of visual cortical area devoted to the transmission and processing of sensory input. The combined effect of sensory density at the fovea and cortical magnification of its signals means that our 180×140-degree field of vision is reduced to the central foveal area with a 2-degree diameter for high visual acuity (Keller

and others 1999; Rybak and others 1998). Under normal reading conditions of font size and reading distance, the multi-resolution field of view breaks down into three regions of visual acuity. The foveal region spanning over the range of 6-8 Latin letters provides the highest sampling resolution. The parafoveal region extending over 15 to 20 Latin letters has been found to possess enough visual acuity to provide *fuzzy* sensory input necessary for efficient recognition (Paulson and Goodman 1999). The third is the peripheral region, spanning over the rest of the visual field, provides the lowest sampling resolution. These characteristics mean that, in order for the mind to study the points of interest of a given scene with enough details for adequate awareness, rapid eye movement shifts (saccades) are required (Osberger and Maeder 1998). Saccades occur every 100 to 500 milliseconds and are guided in a pre-attentive manner. The eye movement literature presents a number of theoretical models and evidence for various factors guiding this behaviour.

Evidence for Low Level Attractors

For a long period, the widely accepted nature of visual processing was predominantly passive, responding to certain physical characteristics and cues present in the surroundings. Even when prior knowledge did bias the visual search process, it was believed to do so by selecting from a list of conspicuous features that had already been detected and stored in a topographically coded map (Saliency Map). According to this model, the selection of target features is done in a largely low-level, fast and feed-forward manner involving relatively little processing. The best evidence for this view came relatively early with neuroanatomical data of a visual pathway containing cells uniquely responsive to different stimulus dimensions. Frequency, line orientation, edge size and direction of motion, for instance, are among the numerous dimensions detectable at the cellular level (Itti, Koch, and Niebur 1998). Other evidence shows the Human Visual System attuned to edge-like and line end features that are bright, contrast rich,

larger or more elongated (Ojanpää 2006; Osberger and Maeder 1998). This was a general view biased towards a separation between the low-level and higher levels of processing and awareness. The biological bias of this view was perhaps compounded by the prevailing Behaviourist paradigm of the time (Paulson and Goodman 1999) emphasizing learning principles and Black Box treatment of organisms. Models based on the Feature Integration Theory proposed a similar view; a vastly parallel pre-attentive saliency mapping guides a time-consuming, serialized object recognition task attending only to a subset of the total retinal projection while attention is used to glue object features hence giving us the illusion of a united whole (Gestalt) (Itti, Koch, and Niebur 1998). These models seemed both biologically plausible and efficient justifying the vastness of human visual processing capabilities.

Evidence for High Level Guidance and Scanpath

Evidence for high-level processes guiding eye movements came as early as a hundred years ago with experimental data showing how the eyes scan text differently when reading foreign and native languages. Other evidence came during viewing and reading experiments where subject eye movements scanning an image or text employ different path patterns depending on the viewing instructions and purpose (Yarbus 1967; Buswell 1937). Some suggested that, depending on the instructions and motivation, a painting could convey different relevant content and different timeline hence guiding subject's eyes in a compatible way to scan a relatively small number of features in a repetitive idiosyncratic manner. The commonality of features attended across subjects as well as the ordered and repetitive nature of the path taken to foveate them provided strong evidence for an object-specific internal representation guiding the scan. Noton and Stark (1971) called it a scanpath and theorized that we learn and memorize images by

associating each with its own representation. Accordingly, such representation is composed and stored the first time we view the image and is later repeatedly invoked every time we encounter it. The nature of the representation is formulated as an ordered sequence of eye muscle motor traces and visual sensory traces residing in our memory. When *played back*, the representation generates a feature template matching route, embedded in the eye scan path, to confirm the hypothesized identity of the image being recognized. (Fig. 1) (Stark and Choi 1996; Noton and Stark 1971).

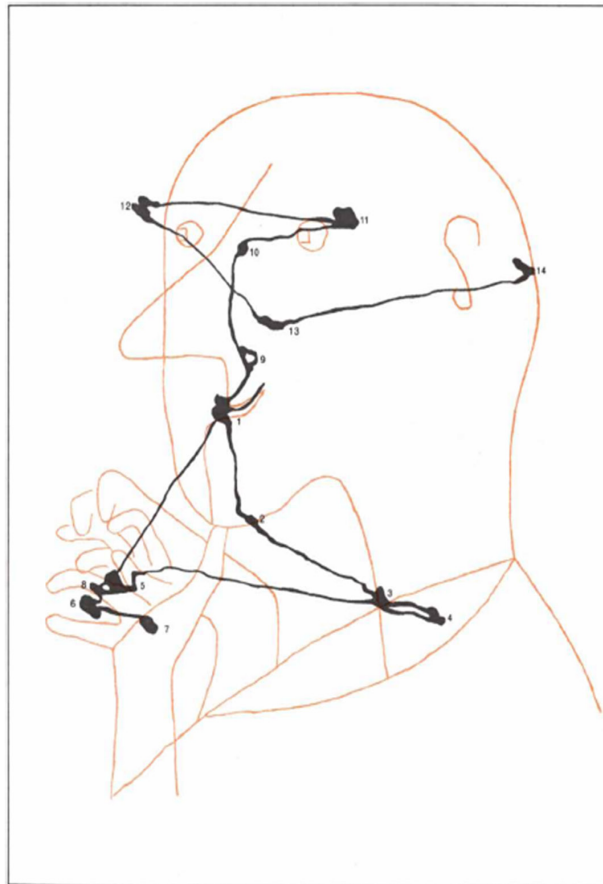


Fig. 1. Early eye movement experiment showing eye scan path. Reprinted, by permission, from Noton and Stark (1971).

Although eye movement researchers never rejected the existence of bottom-up attractors, it was clear from experimental data that these alone could not account for the observed scanning characteristics especially the lack of a common idiosyncratic path across different reading and image viewing tasks (Stark and Choi 1996; Noton and Stark 1971). From this line of reasoning came the idea that other factors, which account for context awareness and higher semantics, have to be considered. As Stark eloquently puts it, “we have to approximate the outside world with an internal representation for our species to have survived.” Such internal representation needs to take account of laws of physical action and reaction and spatial relationship that bottom-up facilities, alone, cannot be expected to embed (Stark and Privitera 1997).

But what evidence do we have that eye movements serve as a template-matching scheme for our internal representations or, to put it differently, that higher level cognitive factors drive our visual search early enough in the human visual process to affect eye movements so significantly?

Two bodies of evidence came to support the acceptance of a scanpath-internal representation model of visual recognition. One came from Imagery experiments showing that eyes will trace similar paths when subjects are imagining or recalling an image as when they are actually viewing it (Brandt and Stark 1997). The other came from anatomical findings by Mishkin of two cortical pathways that, for the most part, presented a concrete and biologically plausible framework for cognitive control of the lower facilities of the visual system. As such, it identifies two pathways for visual processing: the *what* and the *where* pathways dealing with representing object features and spatial information respectively. Perhaps the most significant implication of this framework, as far as scanpath theory is concerned, is that it provides an explicit functional coupling between low-level vision (at the foveal and visual cortex level) and

high-level brain structures involved in visual perception and recognition and points out the role of visual attention in this coupling (Keller and others 1999; Rybak and others 1998).

Review of Eye Movement Research

Can eye movements really serve as a reliable window into cognitive and oculomotor phenomenon and to what extent? A landmark study by Tinker in 1936 investigated the validity of eye movement experiments in reading tasks and concluded that the presence of a tracking camera was not obtrusive to the extent of affecting eye movement performance (Paulson and Goodman 1999; Tinker 1936). In addition, Just and Carpenter (1980) gave credence to two major assumptions stemming from eye movement data and in particular the differing length of fixations during reading tasks. The first is the *immediacy* assumption where word recognition starts as soon as the respective text is focused. The second is the *eye-mind* assumption stating that the eye will fixate a word as long as it is being processed (Paulson and Goodman 1999).

Recent research on eye movement control during reading saw the emergence of many promising computational models based on two contrasting theoretical viewpoints. On the one hand, Reilly and O'Regan (1998) demonstrate that a set of simple oculomotor heuristics can provide a good account of the positioning of subject eye fixations during reading. On the other hand, Reichle, Rayner, and Pollatsek (2003) posit that eye movement is triggered by serial cognitive processing and can account for both location and timing of eye fixations. These two viewpoints roughly correspond to the low-level and high-level control we discussed before. However, the difference among models adhering to the first or the second viewpoints is one of degree rather than kind. In fact, all promising eye movement control models in reading suggest that both oculomotor and cognitive processes interact to guide the reader's eye fixations.

In summary, despite the theoretical contrast within eye movement research, there is wide acceptance regarding the reliability and validity of experimentally collected eye movements as an expression of cognitive and oculomotor processes that drive the execution of a given visual task (Reichle, Rayner, and Pollatsek 2003; Zhang and others 2006).

Eye Movement in Pattern Recognition and Image Analysis

There are two main approaches and respective sets of models that are inspired by eye movement behaviour in the image analysis and pattern recognition literatures. The first set includes recognition models that borrow the concepts of selective attention and serial processing of image parts, chosen based on their perceived importance, to improve recognition performance. In most such models the process starts with a topographic mapping of features of interest detected based on a biologically plausible scheme like presence of contrast, edges or difference in orientation in the subsampled image. This mapping is then used to guide a selective attention template-matching scheme to examine each feature in more details until a certain level of confidence of the image category is reached (Exel and Pessoa 1998; Itti, Koch, and Niebur 1998; Rybak and others 1998; Salah, Alpaydin, and Akarun 2001; Salah, Alpaydin, and Akarun 2002; Xianglin Meng and Zhengzhi Wang 2009; Stark and Privitera 1997). Some of these recognition models are augmented with a higher level of image category awareness to guide the selective attention scheme more *wisely* (Chernyak and Stark 2001; Exel and Pessoa 1998).

The second set is of eye movement prediction techniques that simulate human eye movement behaviour using a variety of saliency detection schemes. Some use biologically plausible low-level saliency detection (Xianglin Meng and Zhengzhi Wang 2009; Hacısalihzade, Stark, and Allen 1992; Osberger and Maeder 1998; Privitera and

Stark 2000; Kienzle and others 2007) or a mathematical model (Watanabe, Gyoba, and Maruyama 1983; Yagi, Gouhara, and Uchikawa 1993) to predict human subject eye fixations while viewing different stimuli. However, limitations in these low-level techniques, and the increased availability of eye tracking equipment motivated a multidisciplinary approach. A number of saliency detection schemes were evaluated, refined or trained using human eye movement and identification data specifically collected for these purposes (Watanabe, Gyoba, and Maruyama 1983; Privitera and Stark 2000; Yagi, Gouhara, and Uchikawa 1993; Kienzle and others 2007; Schomaker and Segers 1999; Peters and Itti 2007).

The main purpose of the eye movement prediction approach is to segment images or other visual media according to regions of perceptual importance. The more important regions can then receive further attention such as: (1) higher sampling rates during image compression or higher bandwidth during broadcasting (Osberger and Maeder 1998; Privitera and Stark 2000), (2) extra processing during image or video analysis (Xianglin Meng and Zhengzhi Wang 2009; Hacisalihzade, Stark, and Allen 1992; Privitera and Stark 2000; Yagi, Gouhara, and Uchikawa 1993; Kienzle and others 2007; Peters and Itti 2007) and (3) identification of informative features in handwriting (Watanabe, Gyoba, and Maruyama 1983; Schomaker and Segers 1999).

Review of Pattern Recognition Related Work

In the two sets of eye-movement-inspired research we briefly summarized above, a number of studies piqued the author's interest due to striking similarities in method or objectives. A study by Salah, Alpaydin, and Akarun (2001) is of particular interest because it is prototypic of other studies that use biologically plausible saliency-based visual attention to train and test pattern classifiers. A common concern among eye

movement prediction studies relates to the validity of saliency detection schemes even in the context of low-level analysis of scenes and videos. We examine a study by Privitera and Stark (2000) because it represents attempts to evaluate a number of such bottom-up schemes based on their ability to accurately predict human-like regions of interest (ROI). A closer look at a study by Kienzle and others (2007) follows since it questions the validity of relying on biologically plausible saliency detection alone. We briefly discuss the proposed use of human eye movement data to compliment bottom-up visual saliency, in order to motivate our own research methodology. If the exclusive reliance on bottom-up saliency techniques is questionable in the context of predicting ROI during general viewing, it is even more limited in the context of task-dependent viewing. We briefly discuss the work of Peters and Itti (2007) which combines human visual fixations recorded during video game play and bottom-up saliency to train a promising top-down-bottom-up ROI estimation scheme.

Selective Attention in Handwritten Digit Recognition

Salah, Alpaydin, and Akarun (2001) present a serial recognition technique whereby an image of a handwritten digit or a face is attended to by examining a sequence of regions defined by a 4x4 grid. The regions are selected based on their perceived saliency using a simplified biologically plausible detection algorithm. This work is particularly interesting in that it borrows significantly from the human visual model by: (1) attending to alignment features – considered an instance of low-level eye movement attractors –, (2) employing selective attention, and (3) using serial processing strategies (Fig. 2). These similarities with the human model allow this method to be scaled to more computationally intensive recognition tasks.

The suggested model has an Attentive, Intermediate and Associative levels. In the

training mode, the Attentive level produces four line maps for each digit image in the training set. This allows the selection of the n most salient points in an image and, consequently, the generation of an ordered sequence from the most to the least salient points, also known as the *where* stream. This stream will guide the selective attention shifts in the Intermediate Level.

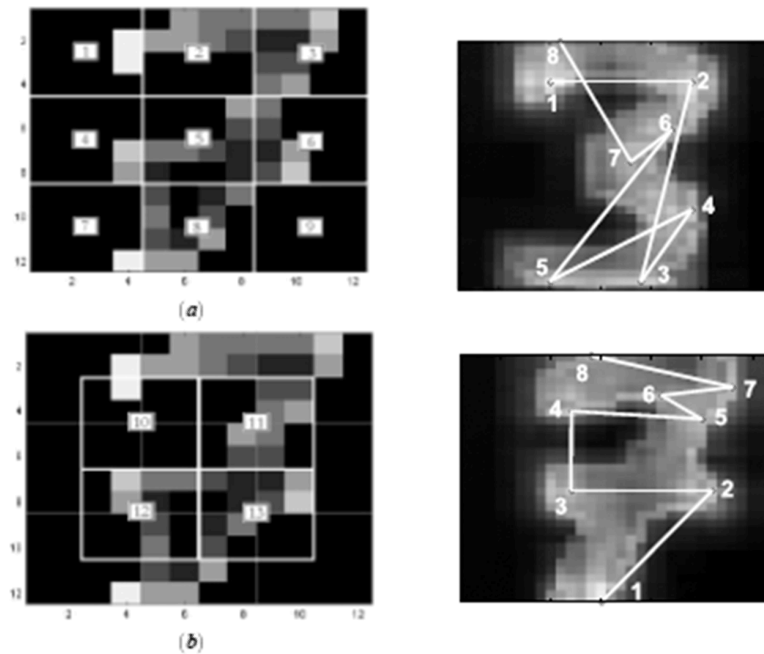


Fig. 2. Saliency mapping in eye movement-inspired digit recognition. *Left*, Attentive Level; *right*, Intermediate Level. Reprinted, by permission, from Salah, Alpaydin, and Akarun (2001).

The Intermediate Level trains, in a supervised manner, a neural network to quantize the regions of the most salient points. To do so, the neural network uses Line Maps data as input and outputs an attribute vector that is clustered into Observation Symbols using k-means clustering. The Associative Level acquires the *what* and the *where* information from the previous levels in the form of quantified Observation Symbols and foveation states respectively to train an Observable Markov Model.

In the recognition mode, the model defines the line maps and foveation states for the digit image to be classified (Attentive Level), quantizes the foveation contents using the same neural-network-k-means clustering procedure to produce Observation Symbols (Intermediate Level) and uses the Observable Markov Model from the Training mode to find the likely digit label.

Selective Attention in Scene Class Recognition

Chernyak and Stark (2001) present a model of scene class recognition augmented with top-down selective attention. The model acquires higher awareness of the scenery image under study by segmenting it into regions of distinct colours. The presence of blue and yellowish-grey regions, for instance, may be an indication that the image is of beach scenery. The model then attempts to increase its level of confidence in the hypothesized scenery category by attending to the respective colour regions in more details until a predefined confidence level is reached (Fig. 3).

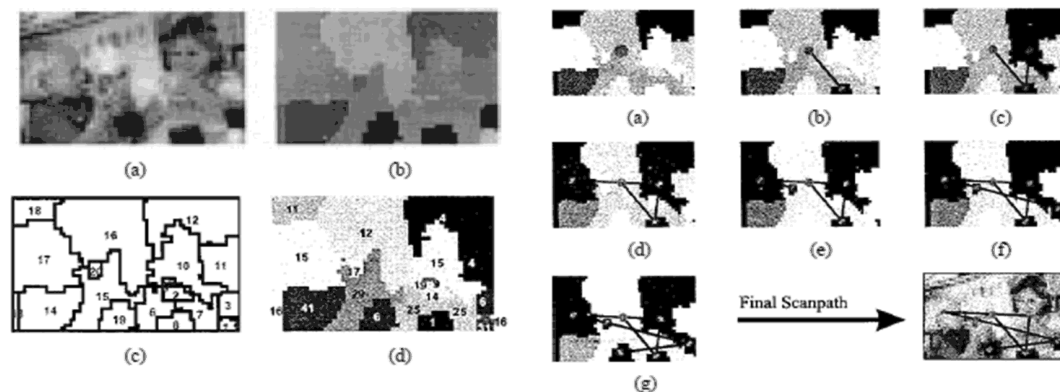


Fig. 3. A top-down augmented recognition model. *Left*, the model segments an image of scenery into colour histograms and identifies the most disambiguating scene segments; *Right*, the model fixates individual segments to verify the hypothesized scene category. © 2001 IEEE. Reprinted, by permission, from Chernyak and Stark (2001).

Bottom-Up Detection of Salient and Perceptually Important Features

The study by Privitera and Stark (2000) represents a significant contribution in the context of our research because it outlines the limitation of relying on a single bottom-up saliency scheme to predict human-like regions of interest (ROIs). The paper describes the process of automatically defining a sequence of ROIs on a given image using a wide range of image processing algorithms and clustering techniques. It also defines a similarity metric, using clustering and string editing, to quantify the difference between two arbitrary sequences of ROIs. This metric allows the evaluation of candidate ROI-detection schemes by measuring the similarity of their output to human visual fixations recorded during the viewing of the same images. The study concludes that a well-chosen image processing algorithm – irrespective of its biological plausibility – can come a long way in predicting human-like regions of high perceptual importance.

Each candidate image processing algorithm is used to define a large set of ROIs that are later clustered using a simple scheme to yield regions of yet *higher* interest. These regions are subsequently compared to eye fixations collected from human participants during free viewing of the same image (Fig. 4).

Two main conclusions can be drawn from this study. First, although image processing techniques by definition do not account for higher-level visual processes, some can be effectively used to predict regions that attract human eye fixations, which are guided, at least in part, by such processes. Second, there is no optimal image processing algorithm that works all the time. Rather, we can only determine which algorithm will work best for a particular class of images, such as paintings, landscapes or terrain photographs, after it has been used on that class and its output compared with human eye movement data.

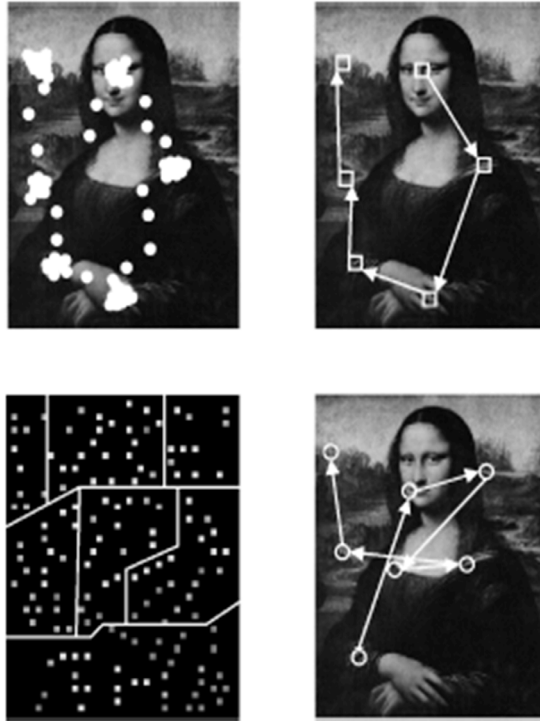


Fig. 4. *Bottom*, A bottom-up algorithm is used to generate fixations similar to those recorded during an eye-tracking experiment (*top*). © 2000 IEEE. Reprinted, by permission, from Privitera and Stark (2000).

Kienzle and others (2007) identify the downsides of manually defining bottom-up saliency detectors in Privitera and Stark (2000) and other works. The paper notes that, irrespective of biological plausibility, a number of parameters still need to be specified manually by the researcher in ways that are often ad hoc and hard to justify. A non-parametric alternative is proposed which starts with a generic saliency model made up of a linear combination of Gaussian radial basis functions. The specific parameters of these functions are then *learned* directly from human eye movement data recorded on the same images. The machine learning model is claimed to predict image ROIs as well as the best biologically motivated models without the questionable assumptions and guesswork associated with the latter.

Top-Down Selective Attention using Eye Movement Data

So far, we presented research that attempts to detect features and define regions based on a bottom-up estimate of their perceptual importance in humans. Such techniques may be suitable for visual media compression, broadcasting, and other applications where the influence of task and context specifics is minimal. However, for applications where the intention is to guide selective attention in visual search or object recognition bottom-up visual saliency alone is very limited. Peters and Itti (2007) present a computational model that combines bottom-up (BU) saliency and dynamic top-down (TD) task relevance in order to predict human visual fixations while playing a video game. The TD model is acquired by learning to associate the signature of a video frame with the corresponding human fixation data. A hybrid biologically plausible saliency map made up of colour and line orientation detectors is used to determine the signature for both BU and TD models. The results show that while the TD model alone performs twice as well as the BU model, a combined model obtained using point-wise multiplication is significantly better than either model on its own.

Other Related Work

A number of other studies have also used human eye movement and identification data to explore aspects of the human model with some relevance to pattern recognition. Watanabe, Gyoba and Maruyama (1983) demonstrated the use of Hayashi's quantification discriminant to predict the confusability of handwritten Katakana letters. They found that features identified by the model were also fixated more frequently by experiment participants. Unfortunately, the details of the study were published in Japanese only and the author found very few related citations (Tappert, Suen, and Wakahara 1990).

Schomaker and Segers (1999) conducted an experiment to identify the types of

geometrical features most attended to by participants during reading of Western cursive handwriting. Participants were asked to identify initially blurred handwritten words under time pressure and could *unblur* a given part of the word using mouse clicks. Their results show that ascenders, descenders, crossings and points of high curvature were most frequently clicked suggesting their informative value.

Maw and Pomplun (2004) studied the role of parafoveal and peripheral vision in the recognition of actors faces using a gaze-contingent mask to hide face parts during an eye tracking experiment. Their findings confirm that extra-foveal vision plays a crucial role in successful recognition. Response time was also reported to increase significantly under more restrictive viewing conditions especially for unfamiliar faces of non-famous actors.

Summary and Conclusions

Eye movement research inspired a number of studies in image analysis and pattern recognition. All start off with a low-level scheme: an algorithm that detects differences in contrast, orientation or colour mapping arguably informative features to (1) guide attentional shifts for recognition (Exel and Pessoa 1998; Rybak and others 1998; Salah, Alpaydin, and Akarun 2001; Salah, Alpaydin, and Akarun 2002; Stark and Privitera 1997), or (2) find human-like features and regions of interest for further analysis (Xianglin Meng and Zhengzhi Wang 2009; Hacisalihzade, Stark, and Allen 1992; Osberger and Maeder 1998; Privitera and Stark 2000; Yagi, Gouhara, and Uchikawa 1993; Kienzle and others 2007; Peters and Itti 2007). A few of these complement the low-level saliency scheme using eye movement data to (1) evaluate the validity of their models (Privitera and Stark 2000), (2) optimize biologically inspired saliency detection parameters (Kienzle and others 2007) or (3) acquire task-dependent strategies (Peters and Itti 2007). Despite promising results of this approach in scene and video analysis, the

author found no mention of similar research in the context of pattern recognition applications. A number of studies, however, are note worthy since they used human data to explore aspects of the human visual recognition model with pattern recognition in mind (Maw and Pomplun 2004; Watanabe, Gyoba, and Maruyama 1983; Schomaker and Segers 1999).

CHAPTER 2
PRESENT STUDY

In the context of handwritten digit recognition, a widely cited study of error analysis identified three categories of errors made by some of the best classifiers in the literature (Suen and Tan 2005).

Category 1, accounting for around a quarter of classification errors, is of digit images that are easily confused with other numerals because of the similarity of their primitives and structures. Images in this category usually belong to these confusing pairs: 4–9, 0–6, and 3–5 some of which are shown in Fig. 5a.

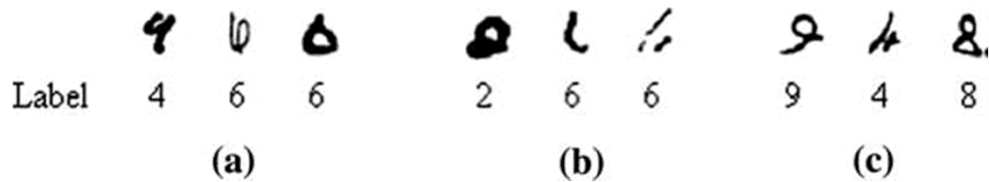


Fig. 5. Samples of misclassified data: (a) Category 1, (b) Category 2, and (c) Category 3. Reprinted, by permission, from Suen and Tan (2005).

Category 2, accounting for around an eighth of all errors, is of digits that are difficult to recognize by classifiers and humans alike because of degradation and distortion due to factors ranging from poor scanners to peculiar writing habits (Fig. 5b).

Category 3, accounting for 62.70% of classification errors, is of digits that humans can recognize without any ambiguity (Fig. 5c.) but are nevertheless misrecognized by classifiers due to the lack of training samples that have the same

prototype (Lauer, Suen, and Bloch 2007; Suen and Tan 2005).

In the present study, we focus on seventy-four MNIST digit images (Fig. 6).



Fig. 6. The original seventy-four MNIST digit images on which our experiment stimuli are based with their respective sequence number and numeral label: (a) fifty-four irregular digits from the MNIST *testing* database and (b) twenty regular digits from the MNIST *training* database. The MNIST index and correct numeral-label are indicated under each digit.

Fifty-four of these digits were selected based on their identification by the literature as commonly misclassified due to one of the above errors (Lauer, Suen, and Bloch 2007; Suen and Tan 2005). We refer to this subset as *irregular* digits. An additional twenty digit images were chosen based on a subjective evaluation favouring the more prototypical among forty randomly selected MNIST images. We refer to this

subset as *regular* digits. The goal of the present study is to build a database of human eye movement collected during the identification of these digits. The purpose of the database is to provide reliable statistics on the regions and features most commonly fixated by the participants during the identification of the selected MNIST digits. Here, we note that while we do not make any specific claims as to the significance of these statistics, our guiding assumption, based on the preceding literature review, is that by properly controlling for various threats to validity, we increase our confidence that the data not only provide reliable recording of participant eye movement, but also that it closely reflect the informative value of digit features they fixate during the digit identification task. As such, our database can serve as a reference to guide future efforts to overcome common errors in handwritten digit recognition by embedding various human-like characteristics and observations. Next, we discuss the data collection methodology.

CHAPTER 3
DATA COLLECTION

Lab Setup and Available Tools

The setup at the Vision Lab features an EyeLink II® head-mounted eye tracker used in at least two other studies reviewed earlier (Maw and Pomplun 2004; Kienzle and others 2007). The system is made up of two PCs: the display and the host (Fig. 7).

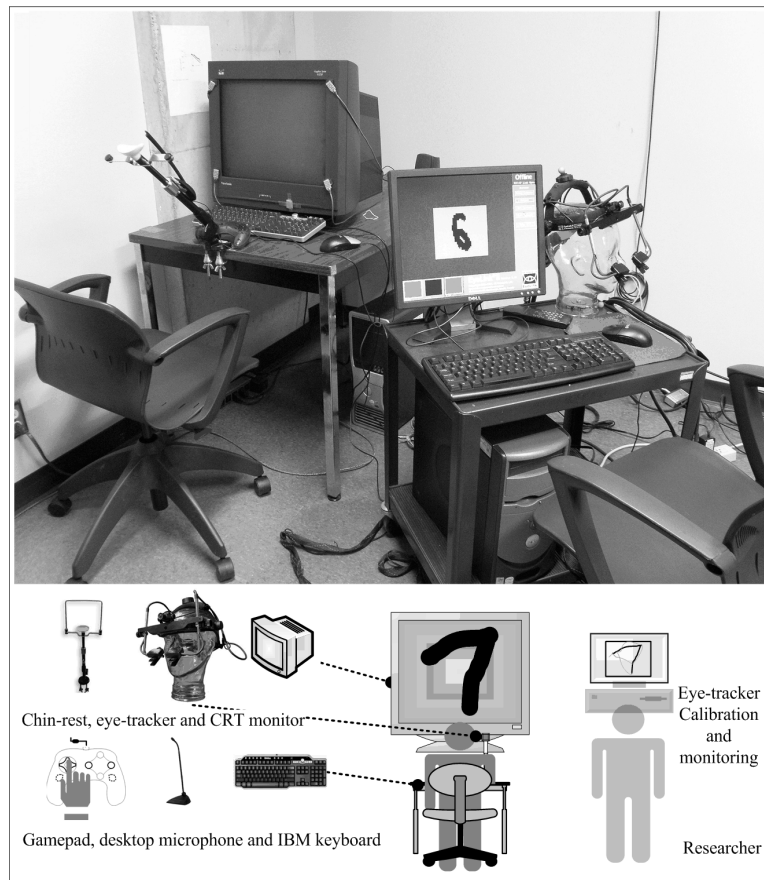


Fig. 7. Experimental setup at the Concordia Vision Lab

On the left, the display PC, where the participant is seated, is delegated the task of presenting audio and visual stimuli and controlling the flow of the experiment. On the right, the host PC, where the experimenter is seated, ensures the proper operation of the eye tracker worn by the participant. The eye tracker connects to the host PC directly and has three infrared video cameras. A head-tracking camera, mounted on the eye tracker's headband near the participant's temple, points forward to capture the IR signature of four infrared markers positioned at the corners of the display monitor. Two adjustable eye cameras, each equipped with a pair of IR illuminators, are mounted at the front of the eye tracker to capture the infrared signal reflected by the participant's eyes (*EyeLink II head-mounted user manual* 2009). The host receives the infrared video from the two eye cameras at a sampling rate of 500 frames per second. A host background algorithm measures the pupil's position as detected in these samples to calculate the instantaneous rotation of each eye. The video from the head-tracking camera allows detection of the position and rotation of the participant's head which is then combined with the eyes' rotation to determine the on-screen gaze coordinates. Another real-time host algorithm uses angular velocity heuristics to parse the resulting stream of gaze coordinates into corresponding saccades and fixations. The EyeLink II (2009) specification claims a binocular eye-tracking capability with a 500 Hz sampling rate, a typical average error of $<0.5^\circ$, and a spatial resolution of $<0.01^\circ$. A gamepad-like button box, used to record participant manual response, is connected directly to a host's USB port to maximize timing accuracy (*EyeLink II head-mounted user manual* 2009).

In order to record various experiment trial events and eye movement data in a synchronized fashion, the host communicates with the display computer constantly during the operation of the eye tracker. A crossover network connection is used to notify the host of changes in stimulus presentation and other trial events on the display side. It is

also used to signal various eye-tracking events captured on the host's side to the display computer. Besides making synchronized recording of trial events and eye movement data possible, the real-time connection also allows the experimenter to monitor both participant gaze and what the participant is viewing via the host's LCD monitor.

The display computer features a pair of desktop speakers to play back audio stimuli and a standard IBM keyboard providing an alternative form of input to the button box. It also features a high-end flat CRT monitor with a 120-Hertz refresh rate at 1024×768 resolution and a 20-inch viewable screen. In order to improve eye-tracking reliability the participant's head is stabilized with a chin guard mounted at the edge of their desk at a standard viewing distance of 57cm from the display monitor. To enable recording of participant voice, a common desktop microphone, attached to the stem of the chin guard, is connected to the display computer audio input port (Fig. 7).

A proprietary software program, Experiment Builder® (2009), is available to create eye-tracking experiments that run seamlessly on the above eye-tracking hardware. The graphical environment comes with a suit of predefined node-like components to facilitate the process of designing and controlling the presentation of audio and visual stimuli during experiments. Action nodes and trigger nodes can be dragged, dropped, linked together and grouped into sequences to define the experiment's workflow (see Fig. A1c).

A typical workflow starts with an introduction in the form of a sequence of nodes (see Fig. A1a). Display action nodes with nested text resources may be used to display introductory slides (see Fig. A2) while key-press trigger nodes allow the participant to move to the next slide in the sequence. The eye tracker setup sequence follows allowing the calibration and testing of the eye tracker. During calibration, the participant, wearing the head-mounted eye tracker, is asked to focus on dots appearing in sequence at different

spots on the display monitor. At the mean time, a host background algorithm uses the sample frames captured by the eye tracker's infrared cameras to determine various parameters necessary for subsequent eye tracking. The next sequence typically represents an experiment block. A block sequence contains a sub-sequence defining the flow of an experiment trial (see Fig. A1c). A display action node with a nested image resource may be used to display a PNG file to cover the full screen while a voice-key and a timer node can trigger a sound action node to play an error sound if the participant fails to identify the image on time. A trial sequence can be executed a predetermined number of times. The experimenter can define its iterative behaviour by attaching a spreadsheet-like data source. Each data row contains stimuli and parameters corresponding to a single trial. The experimenter can also set randomization rules to control the order of trials during the experiment.

When the workflow is completed, it can be compiled and built into an executable experiment. An executable generated by Experiment Builder can, on the one hand, manage the various operations of the eye tracker while, on the other, control the experiment workflow and record participant responses all under the close monitoring of the experimenter. This greatly facilitates the task of running eye-tracking experiments as well as storing the synchronized recordings of participant eye movement, button presses and other event data in a single binary file (EDF). The experimenter also has the option to record select experiment data like key presses and trial-specific parameters to a more readily accessible plain text Results File. Data collected during a single experiment session is saved under a distinct session folder specified at the start of the experiment. A session data folder may contain an EDF file, a Results File or both. Since the actual trial order may be randomized and different from the original order specified in Experiment Builder, a modified version of the data source is generated in the session folder at the end

of the experiment. The new data source has its rows rearranged to reflect the actual trial order.

EDF files can be visualized using a proprietary data analysis program. Alternatively, a free API is also available which allows parsing, reading and exporting of the EDF file contents to plain text.

Experiment Design Justification

Due to the novelty aspect of the present study – the absence of research on identification of a specific category of images in the eye movement literature – a number of new considerations relating to data collection have to be taken into account. Our purpose of research dictates numerous departures from experiment design norms whereby multiple controls are put in place to allow investigation of a stated hypothesis with clearly defined and quantifiable variables. For one thing, we have no prior notion of how humans will examine these or other handwritten digits; indeed, our digits were not selected based on how we expect humans to respond to them but rather how problematic they are to automatic classifiers. It is also worth mentioning that the puzzling complexity of cognitive processes unfolding during handwriting identification is precisely the reason why we chose to exploit the human model in the first place. This makes the general methodology found in a number of studies discussed in the literature review such as: Privitera and Stark (2000), Yagi, Gouhara, and Uchikawa (1993), Kienzle and others (2007), Peters and Itti (2007), Schomaker and Segers (1999) and Noton and Stark (1971), more appropriate than the more rigorous design norms typically followed in eye movement research proper: (Ojanpää 2006; Paulson and Goodman 1999; Legge and others 1997). To put it plainly, since we are trying to learn from humans how to better identify handwritten digits, the best we can do is let them do it as normally and unconstrained as possible. However, such lack of experimental controls comes with a

number of threats to validity especially under eye-tracking conditions.

In the particular case of handwritten character recognition, a plausible threat to validity relates to the size at which single-character images need to be presented under eye-tracking conditions. Individual eye fixations have a perceptual span ranging from 1-5° of the field of vision. Under normal reading conditions, *printed* Latin letters and Arabic numerals occupy a visual angle of around 0.25° (Rayner 1998; Brysbaert 1995). This means that a single fixation is normally enough to process several printed characters (Ojanpää 2006; Legge and others 1997; Duchowski and SpringerLink 2007). In addition, technical specifications of even higher-end eye trackers like the one we are using acknowledge tracking errors may often exceed 0.5° (*EyeLink II head-mounted user manual* 2009). Therefore, in order to record human eye fixations with a level of reliability and details suitable for our purposes in automated recognition, the digits will need to be displayed at a much larger scale. This may represent a serious threat to validity since it constitutes a major departure from viewing conditions under which humans normally recognize digits.

Another threat to validity stems from peculiarities in the widely used MNIST image database from which the seventy-four handwritten digits were chosen. Scaling 20×20-pixel handwritten digits to fit a standard 20-inch display already results in heavily aliased images with prominently jagged contours or jaggies (Fig. 8). Such artefacts may indeed be conspicuous enough to influence eye movement due to sudden changes in orientation and intensity against the plain background (Itti, Koch, and Niebur 1998).

To make things even worse, in the process of size normalization, most MNIST database images were sub-sampled from the original NIST version (LeCun and Cortes 2010). In order to counteract the resulting aliasing, an anti-aliasing algorithm was used to

make MNIST digits look like the original NIST digits when viewed at normal size of 1° or less; however, when these digits are magnified significantly this process may actually add distortions and artefacts to the final image (Fig. 8b).

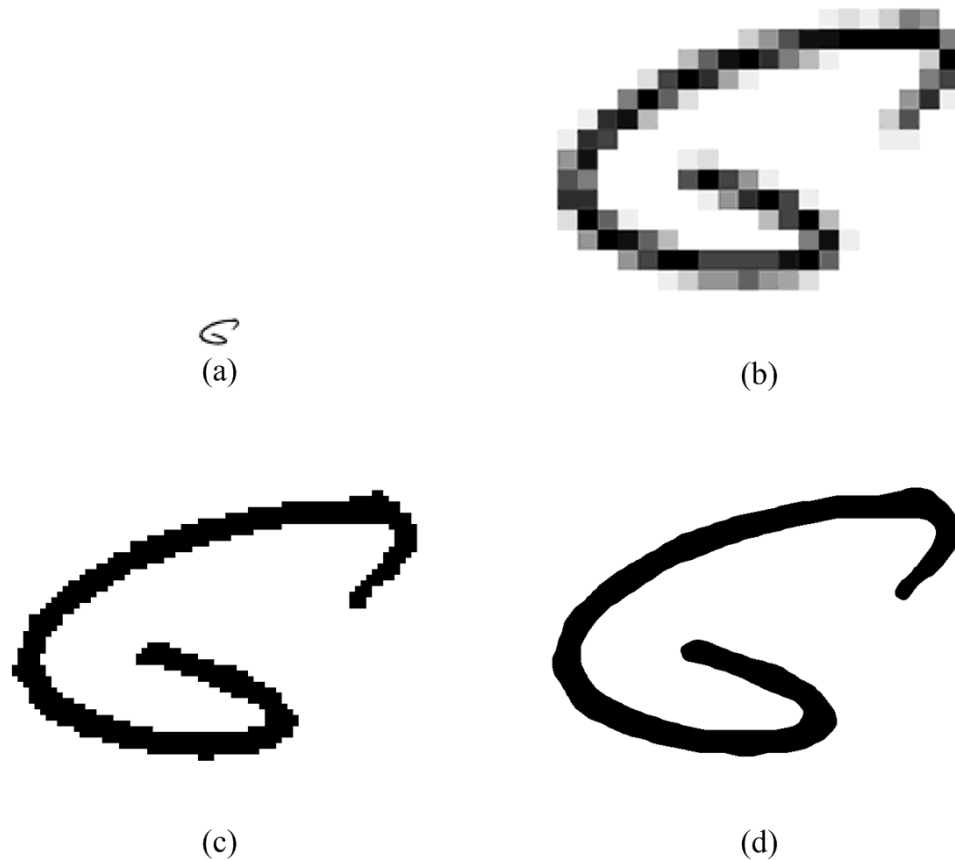


Fig. 8. *Top*, Aliasing artefacts and anti-aliasing distortions in MNIST images: (a) shows a 20×14 -pixel MNIST image of a *six* while (b) shows the same image after scaling to 200×140 pixels using nearest-neighbour interpolation. *Bottom*, Aliasing artefacts in NIST images: (c) shows the original NIST version after scaling to 200×140 using nearest-neighbour interpolation while (d) shows NIST after being scaled to same size using bicubic interpolation and *smoothed* using MATLAB *disk* filter.

When it comes to foreground-background contrast manipulation the presence of greyscale pixels in MNIST images presents an additional problem. According to the literature on eye movement during reading and visual search, both luminance contrast

and size of printed character interact to affect participant performance. A decrease in contrast and character size are reported to significantly decrease performance speed and increase both number of saccades and the duration of individual fixations (Ojanpää 2006; Legge and others 1997). While no data is available on the effect of these dimensions during identification of individual or handwritten digits, it is plausible to assume such effects apply in our experiment. For instance, a lowering of display contrast may force the human visual system to fixate features of interest more closely. Hence, in addition to controlling for the effect of contrast variations on collected eye movement, the use of different contrast conditions can also help us identify guidelines that, when used to display a handwritten digit, may yield eye movement data that is most suitable for pattern recognition applications (Optimal Viewing) for similar research in the future.

However, varying the luminance contrast of greyscale images presents some questions and challenges. For instance, it is quite likely that the visual system perceives luminance variations in a non-linear manner making any such manipulation subjective and difficult to justify.

Therefore, eye movement recorded under these conditions may be largely influenced by the manner in which the digit being identified is displayed. This casts serious concerns about the validity of our methodology: The use of eye movement data to determine the informative value of handwritten digit features during identification. Unfortunately, many of these threats to validity are largely unavoidable: In order to obtain detailed eye movement statistics suitable for character recognition applications, digit images have to be displayed at a much larger scale than during normal reading. Conspicuous aliasing artefacts will follow potentially influencing collected eye movements. However, measures can be taken in designing the experiment that can help control for these and other extraneous variables. Below, we discuss these considerations.

Visual Parameters and Display Considerations

The reduced resolution, increased spatial aliasing and complications relating to image processing manipulation make the use of MNIST database less suitable for the purposes of our research than the NIST images from which it is derived. NIST digits are black and white (binary), and offer higher details when compared to their MNIST counterparts (Fig. 8). We therefore opted for the NIST images instead.

Finding Matching Digits in NIST

Matching MNIST digits into the NIST database from which they are derived is not a task to be taken lightly. The creators of the MNIST database do not provide the detailed image processing algorithm they used to size-normalize NIST digit images and the author found no published mapping between MNIST and NIST images based on indexes or identifiers. Matching by naked eye is not an option since each NIST numeral set contains tens of thousands of handwritten digit images. To find the corresponding twenty regular and fifty-four irregular MNIST digits in the NIST database, the author experimented on a number of image processing transformations using MATLAB Image Processing Toolbox (2009). Fig. 9 summarizes our simple yet effective matching scheme.

First, using a given MNIST index and an open-source MATLAB tool *loadMNIST* (Sirotenko 2009) the digit image is extracted into a matrix object and transposed to compensate for the different coordinate systems. The loaded MNIST image differs from the NIST original in two ways: The MNIST image is size-normalized to fit a 20×20-pixel box then padded by positioning its centre of mass in the middle of a 28×28-pixel canvas; the NIST images, on the other hand, have an arbitrary size and no padding.

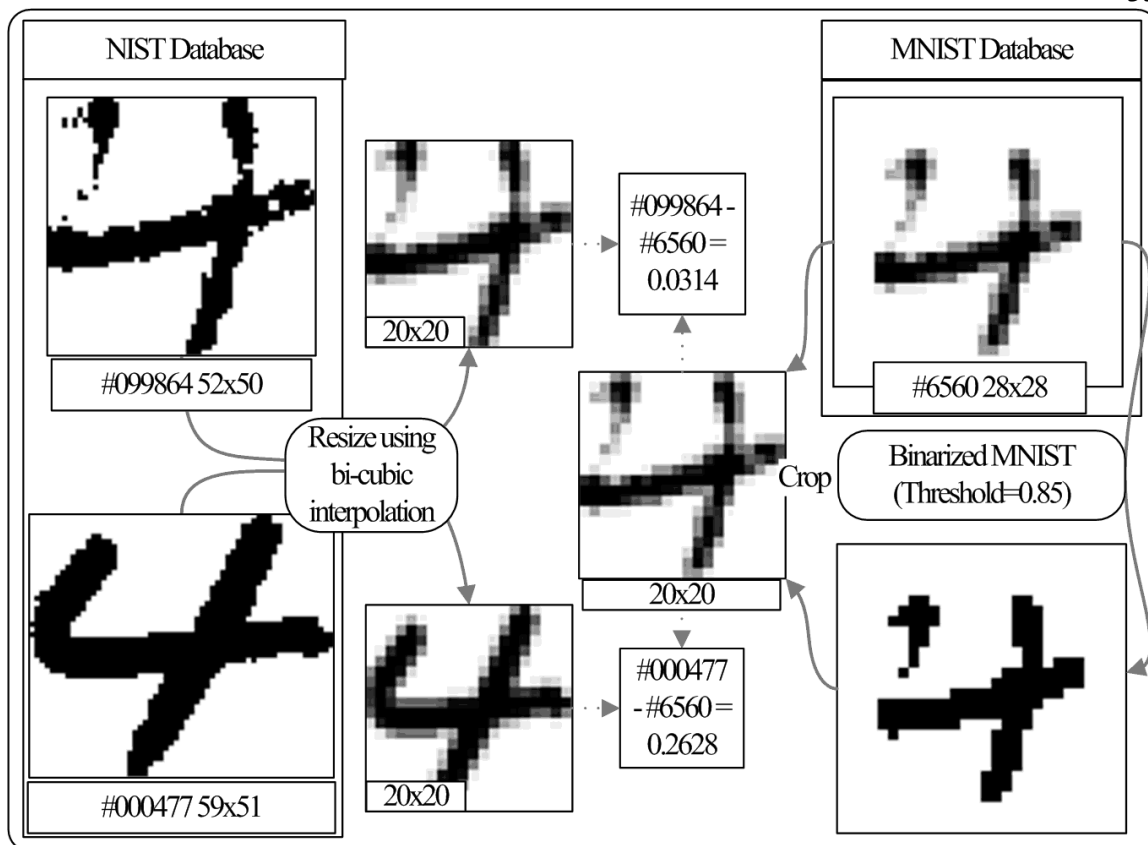


Fig. 9. MNIST-NIST matching scheme

In order for the two image types to be easy to compare they have to have the same size without the padding; however, perhaps due to the anti-aliasing algorithm, the removal of the MNIST white padding often results in an unpadded image of a different aspect ratio compared to its NIST counterpart. Upon further examination of a small set of pre-matched MNIST and NIST digits, the author determined that in order to rectify this discrepancy additional cropping of the lightest grey contours is needed. Based on greyscale index range from black=0.0 to white=1.0, the threshold grey index was determined experimentally as 0.85. Cropping the padding and light grey contours proceeds as follows: (1) the greyscale MNIST images are converted to binary images via thresholding: Pixels with grey indexes of the threshold value or lower become black

while all other pixels become white; (2) the bounding box of the binary image is calculated; (3) this bounding box is used as a guide to crop the initial greyscale MNIST image into an unpadding version with an aspect ratio that is now more true to the original NIST.

The goal of the next step is to produce a similarity ranking for each of the candidates in the NIST database using a very simple scheme. Since both NIST and MNIST images are labelled and the corresponding numeral is known, NIST candidate images are found under the corresponding numeral's folder. Each NIST candidate is loaded then resized to match the dimensions of the target MNIST image. Then, the average per-pixel absolute difference between the two greyscale-indexed matrices is calculated and the result used to rank the similarity of all NIST candidates to the MNIST image in question. The author experimented with a number of resizing interpolation methods available in MATLAB and determined that bicubic interpolation gave a greyscale image that resembled the MNIST image the most; indeed, in most of the fifty-four MNIST digits, the correct NIST match was found among the ten highest ranking candidates according to this scheme. This greatly simplifies the matching process and suggests that the size normalization and anti-aliasing scheme used to create the MNIST database has a somewhat similar effect as the bicubic interpolation resizing implemented in the Image Processing Toolbox.

Once the corresponding NIST digit images have been identified, the process of manipulating contrast and other viewing conditions becomes easier given the binary nature of these images. The higher level of details available in NIST images also significantly alleviates the problems associated with spatial aliasing outlined before. The general process of creating digit stimuli starts with size normalization of NIST images to make the most use of available display area. The images are rescaled such that the digit

bounding box fits in a 575×575 pixel box or roughly $23^\circ \times 23^\circ$ at the standard 57cm viewing distance – compared to a 20×20 -pixel box ($0.8^\circ \times 0.8^\circ$) for MNIST images. This box is then centred on a 1024×768 pixel canvas chosen to cover the entire 20-inch display during the experiment– compared to the 28×28 pixel canvas for MNIST images. The foreground and background greyscale colours can now be chosen based on the desired Michelson contrast. Smoothing algorithms that eliminate spatial aliasing along the rough jagged contours of the digit images can also be applied more conveniently.

Verbal Task Design and Testing

Identification Response Considerations

Manual response using a button box or a computer keyboard is generally much more convenient as far as data analysis is concerned; however, due to the need for ten different buttons corresponding to the ten Arabic numerals, the use of a button box or keyboard would greatly interfere with participant eye movement in our experiment. An on-screen input method where participants can use eye movements to select alternative responses is another possibility but may still present a serious visual distraction when displayed along the handwritten digit. Verbal identification may also interfere with eye movement. For instance, it is plausible to assume that eye movements made during speech have a special significance. However, unlike other response methods, questionable fixations can be reliably isolated by identifying the corresponding speech segments. Therefore, despite the added labour and overhead associated with the labelling and accurate isolation of audio segments, we opted for verbal identification as the superior identification methods in the context of our experiment. As such, we refer to the eye-tracking experiment as the Verbal task.

Verbal Task Temporal Parameters Considerations

Of prime concern to us is the relevance of recorded eye movement to the identification task. Our driving motivation during experiment design is to say with confidence that features most fixated by participants during the presentation of handwritten digits are also most attended to by cognitive processes involved in recognizing the numeral in question. By asking participants to identify the displayed handwritten digit such claim can be made with more confidence. However, since no similar research has been conducted in the past, this leaves the question of how participants should be prompted for their verbal response wide open. For instance, should the participants have all the time they require before they give an answer?

To answer such question a closer look at the intended use of the collected eye movement data is due. One of the principle challenges in automated handwriting recognition is determining the most disambiguating features. For example, in handwritten digit recognition an important challenge is to identify features that can best tell two or more numerals apart. If the human participant is asked to identify the displayed digit once and given an open window to do so, we risk recording visual fixations without the ability to correlate them with the corresponding numerals that the participant is considering as potential answers. This limits the usefulness of collected data especially since many of the digit stimuli resemble two or more numerals.

An alternative approach would be to prompt the participant for their answer once at the start of a trial with a response window that is only sufficient for the participant to give their *quick instinct*. After a predefined period, during which the participant is given ample window to reconsider their earlier response, another beep prompts the participant for their second and final answer. This way, eye movement recorded during the first prompt may be better correlated with their first verbal response; the second set of eye

movement recorded during the intermediate period between the two prompts is likely to correlate with a wider set of features that represent various candidate numerals that the participant's mind is considering while the third set of eye movement captured during the second prompt is likely to correlated with their second verbal response.

This still leaves many temporal parameters wide open. Assigning proper trial events time and duration can be very subjective without proper pilot testing which we discuss next.

Verbal Task Pilot Testing

The pilot test was conducted on ten participants. Five participants were Vision Lab volunteers and another five were students recruited in class.

The purpose of the pilot test was to fine-tune a number of experiment parameters:

1. The first response window
2. Timing of the second prompt
3. Second response window
4. Total trial length
5. Different contrast conditions
6. Description of the task and wording of the introduction
7. Content of pre-recorded sample trials
8. Number of practise trials needed to familiarize participant with the task prior to

data collection

Based on pilot testing, the first prompt was set to occur at the onset of the trial and give the participant 1.5 seconds to provide their verbal response. This response window was determined based on observations made during and interviews conducted after each pilot experiment. The 1-1/2-second period was deemed a good compromise to ensure that participants have sufficient time to give a preliminary identification of the presented

handwritten digit yet still have enough doubt requiring further verification.

The pilot testing also helped determine the appropriate timing of the second prompt and subsequent response window. A period of three seconds was deemed an appropriate delay between the estimated end of the first response and the onset of the second prompt followed by a one-second response window.

A tutorial segment was also added to the experiment based on pilot feedback that task description was a bit vague. The segment came after the introduction and before the practise trials and included four pre-recorded sample trials. The first and second sample trials show an *easy* image of a handwritten numeral *one* with a pre-recorded voice demonstrating how to identify the numeral verbally at each prompt. The third and fourth sample trials show an ambiguous image of numeral *two*, which resembles a *seven*, with a pre-recorded voice identifying it as a *two* at the first prompt then a *seven* at the second prompt. The purpose of the third and fourth sample trials is to impress upon the participant the fact that some of the digits they will be identifying may be so poorly written that they should go with their *instinct*. The participant is instructed as follows:

You should focus on giving a quick answer on the first prompt to avoid hearing the error sound... [And] it's perfectly alright to change your mind on the second prompt if you feel you have a better answer.

Pilot testing also helped determine the number of practise trials needed to familiarize participants with the digit identification task. Six additional NIST digits were handpicked for practise. Unlike test trials, a practise trial is repeated as long as the participant fails to give their verbal answer within the allotted response window. Due to the limited sensitivity of the available desktop microphone, the practise trials are also important in that they train participants to identify the digit more clearly to avoid hearing an error sound after which they would need to repeat the practise trial again. This also makes the labelling of verbal responses easier during data analysis.

During subjective analysis of high contrast condition with foreground RGB(0,0,0) and background RGB(240,240,240), an eye movement pattern emerged in some participants where fixations exhibited excessive central tendency. This observation is consistent with a well-documented correlation between visual span and contrast (Ojanpää 2006; Legge and others 1997). In short, the higher the contrast between stimulus and background the bigger the part of the stimulus that can be perceived with a single fixation and the less eye movement is required to study the whole stimulus. In our experiment this is a serious threat to validity since potentially informative features on the outskirts of a handwritten digit are significantly less likely to be directly fixated regardless of their informative value and how likely they are to be attended to by the participant's visual and cognitive processes. This came as a confirmation of the concern that various contrast conditions need to be tested and led the author to test lower contrast conditions to investigate the increase in the eccentricity of recorded fixations. In total, four levels of Michelson contrast were tested during the pilot. To reduce eye fatigue associated with bright white backgrounds all contrast conditions had a light grey background RGB(240,240,240). The four contrast conditions had grey foreground colours: RGB(0,0,0), RGB(192,192,192), RGB(210,210,210) and RGB(228,228,228). Based on post-experiment interviews and subjective evaluation of recorded eye movement the author determined that foreground RGB210 provides a good compromise by addressing the issue of excessive central fixations without significantly affecting participant identification performance.

Main Verbal Task Viewing Conditions

Seven different PNG image files were generated for each of the seventy-four NIST test digits and six NIST practise digits (Fig. 10). The background colour of all seven PNG versions was set to the special transparency value that allows for more

flexibility during experiment design and deployment. Five of these are scaled using nearest-neighbour interpolation to ensure that the resulting digit images contained the same details as the original digits (Fig. 10a-e). The resizing is done such that the digit bounding box fits in a 575×575 pixel box. This box is then centred on a 1024×768 pixel canvas chosen to cover the entire 20-inch display during the experiment. The foreground colour (FG) is set to one of the following RGB values: (0,0,0), (120,120,120), (180,180,180), (210,210,210) and (228,228,228).

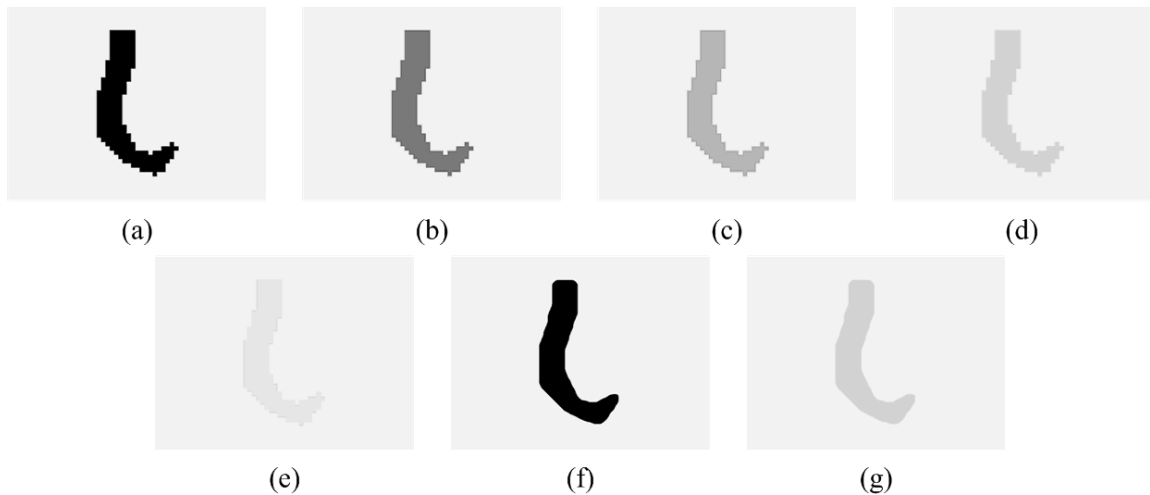


Fig. 10. viewing conditions in Verbal: (a) Unsmoothed-FG0, (b) Unsmoothed-FG120, (c) Unsmoothed-FG180, (d) Unsmoothed-FG210, (e) Unsmoothed-FG228, (f) Smoothed-FG0 and (g) Smoothed-FG210. See Fig. A7-21 For a complete listing of handwritten digit stimuli in all Verbal viewing conditions.

To control for the effect of rough aliased contours we discussed earlier, two additional PNG images were generated using a smoothing scheme (Fig. 10f-g). A widely used MATLAB Image Processing Toolbox filter called *disk* was used to smooth the rough contours. In order to minimize the changes on the resulting image, the original NIST images were size-normalized using bicubic interpolation. This results in significantly less rough contours allowing use of a relatively small 22-pixel disk during

smoothing (Fig. 8d, Fig. 10f-g). The foreground colour for the two smoothed versions was set to one of the following RGB values: (0,0,0) and (210,210,210). Since all seven PNG versions have background pixels set to a special transparency value, the actual background colour (BG) can be changed during experiment design and programming. The experiment background was ultimately set to RGB value (240,240,240) instead of white (255,255,255) based on a common norm in eye-tracking experiments to reduce background brightness (Fig. 10).

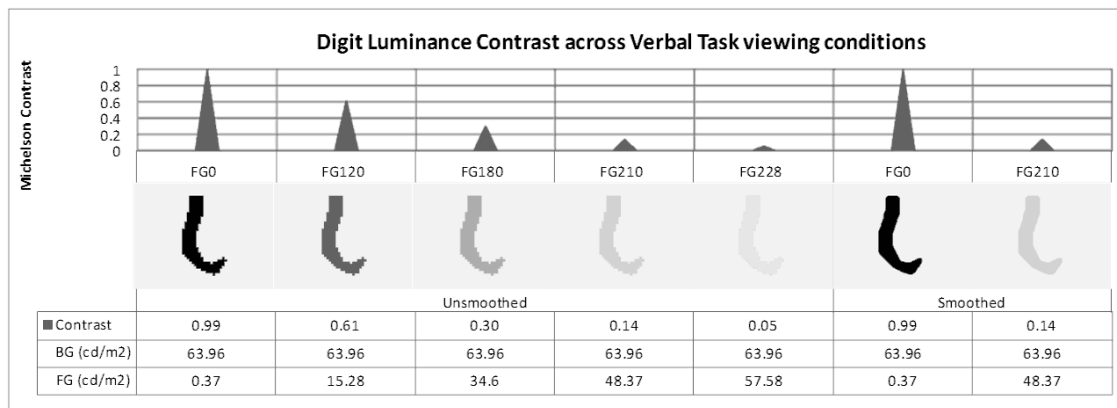


Fig. 11. viewing conditions in Verbal and respective background and foreground luminance measured using a handheld luminance meter. Top table row and graph show corresponding Michelson contrast

defined as $C = (Lum_{BG} - Lum_{FG}) / (Lum_{BG} + Lum_{FG})$

Verbal Task Design and Implementation

The Verbal task was designed and implemented using Experiment Builder. EyeLink II options were set to pupil-only binocular tracking at 500 Hz with high saccadic detection sensitivity. The Verbal task workflow is made up of four major sequences (see Fig. A1a). In the introduction sequence, the participant is presented with a welcome screen introducing the handwritten digit identification task, its general purpose, approximate duration and asking them to advise the experimenter of any discomfort they

may experience with the head-mounted eye tracker. The introduction sequence also explains the use of the button box and microphone and reminds the participant that they are free to discontinue the experiment at any time without adverse consequences.

The second major sequence is the setup sequence that insures accurate calibration of the eye tracker and proper working of the desktop microphone. During calibration, the participant is asked to focus on nine dots appearing in random sequence on a 3×3 grid covering most of the display monitor. The calibration is followed by a similar validation process to ensure accurate eye tracking. The experimenter closely monitors the calibration process and can make adjustments on the eye tracker's headband or eye cameras, assist the participant with detailed instructions or repeat the entire process as they see fit. During the microphone test, the participant is asked to say, "Continue". If the audio capture setup fails to detect the participant's voice, the experimenter can readjust the position of the microphone and ask the participant to repeat the test. The setup sequence completes when the voice-key node detects the participant's voice properly hence triggering the display node, which confirms that audio recording is in working order.

The third major sequence is a collection of five tutorials that were added based on pilot testing. The first tutorial familiarizes the participant with a recurrent display called *drift correction* that precedes each of the six practise trials and seventy-four test trials. The purpose of drift correction is to compensate for any drift in eye tracking that may result from variations in participant pupil size or other parameters effecting eye tracker calibration. It can also correct for small slippage in the head-mounted eye tracker. During the drift correction display, the participant sees a blank screen with a black dot located at the centre. She is then required to press a button on her gamepad while focusing her eyes on the central dot. If the drift angle separating the fixation locus from the central dot is

within the acceptable range (typically $< 5^\circ$), the eye-tracking host software uses the drift angle to update calibration parameters. If, however, the drift angle is too large the experimenter is alerted on their display monitor and a recalibration process is required before the experiment can go on. When in doubt about the quality of calibration, the experimenter can pause the experiment between trials to carry out eye-tracking validation or repeat the entire calibration process as they see fit.

Tutorials two through five present four pre-recorded sample trials. These tutorials are played back like screencasts allowing the participant to move through textual descriptions and pre-recorded trial segments using their gamepad. The first and second sample trials show an *easy* image of a handwritten numeral *one* with a pre-recorded voice demonstrating how to identify the numeral verbally at each prompt. The third and fourth sample trials show an ambiguous image of numeral *two* that resembles a *seven* with a pre-recorded voice identifying it as a *two* at the first prompt then a *seven* at the second prompt. The purpose of the third and fourth sample trials is to impress upon the participant the fact that some of the digits they will be identifying may be so poorly written that they should go with their *instinct*. The participant is reminded that they should focus on giving a quick answer at the first prompt to avoid hearing the error sound and because it is perfectly alright for them to change their mind at the second prompt if they feel they have a better answer.

The fourth and last major sequence is the trial block sequence. Experiments are often divided into a number of trial blocks. This division helps separate practise trials from test trials and trials corresponding to one experimental condition from trials of other conditions. Since Experiment Builder's abstraction of a sequence allows for nesting and looping we can think of a block sequence as a *nested for loop* statement. The block sequence corresponding to an outer *for* loop executes a number of times corresponding to

the number of experiment blocks. The trial sequence corresponding to an inner *for* loop executes a number of times depending on the current block. In the Verbal task, the block sequence executes twice: first for the practise block with six practise trials and next for the test block with seventy-four test trials (see Fig. A1b). The workflows of practise and test blocks differ in their introduction slides but share the same trial sequence (see Fig. A1c).

After completing the tutorial sequence, the participant is introduced to the practise part and reminded that they can rest their eyes after each handwritten digit. During the drift correction display, participants have the opportunity to pause the experiment, rest their eyes and resume the experiment when they are ready. This is particularly important in eye-tracking experiments since the presence of the high-energy infrared illuminators of the eye tracker can cause the eyes to get dry and teary which, beside the unpleasantness, can affect eye-tracking reliability. After completing the six practise trials, the participant is advised that the *main* part is about to start. They are reminded of how to rest their eyes and to inform the experimenter in case they feel any discomfort due to the head-mounted eye tracker. Once the seventy-four main test trials have been completed the participant sees the concluding slide advising them that the eye-tracking part of the experiment is over and thanks them for their valuable contribution.

As noted earlier a trial sequence executes a predetermined number of times. In our experiment, this iterative behaviour is defined by attaching a spreadsheet-like data source (see Fig. A1b). Each data row contains stimuli and parameters corresponding to a single trial. These column fields include the trial index, the name of a handwritten digit image file, the corresponding numeral label, whether the digit is regular or irregular, practise or test and other details of secondary interest. Data source columns also allow setting the trial randomization rules. The Boolean practise column is used as a blocking criterion. As

such data rows with practise value *'true'* are used for the practise block while data rows with practise value *'false'* are used for the main test block. The trial index column, from 1 to 80, is used to define within-block randomization behaviour with the experiment start time as the seed value for the pseudo-random number generator (see Fig. A1b).

The trial sequence defines the operations to be executed just before, during and right after each trial. Two pre-trial operations are essential to the reliable recording of eye movement. The first operation, executed using the *prepare sequence* action node, ensures that all graphical and audio resources are optimally cached to avoid processing overhead during the actual trial. The second operation, executed using the drift correction node, ensures that small drift in eye-tracking calibration is corrected before the recording of trial eye movement begins. In our experiment, a third operation is added that initializes the audio file to which participant sound will be recorded during the trial.

The actual trial lasts six seconds during which a full-screen handwritten digit image is presented and the participant's speech is recorded to a trial-specific audio file. The digit PNG files used for a given participant corresponds to one of the seven viewing conditions of contrast and smoothing we discussed earlier and remains the same throughout the Verbal task. As such, in unsmoothed viewing experiments the foreground colour of handwritten digits is fixed to one of the RGB values (0,0,0), (120,120,120), (180,180,180), (210,210,210) or (228,228,228). While in smoothed viewing experiments, the foreground colour of handwritten digits is fixed to one of the RGB values (0,0,0) or (210,210,210). The background colour of digit images as well as all experiment displays and slides is always set to RGB value (240,240,240).

The logic and timing of experiment trials is the same regardless of selected viewing condition. At the start of the trial, coinciding with the onset of the digit, the participant hears a 100-millisecond beep prompting for their first verbal identification

with a 1-1/2-second response window. At the end of this window, 1.5 seconds into the trial, an error sound is given if the participant has failed to call out their first answer. During the next three seconds, the participant has the opportunity to take a closer look at the digit. At the end of the three-second window, 4.5 seconds into the trial, the participant hears another 100-millisecond beep prompting for their second verbal identification with a one-second response window. At the end of this window, 5.5 seconds into the trial, an error sound is given if the participant has failed to call out an answer within the previous second. The handwritten digit is replaced with a blank display six seconds into the trial. The trial ends when the recording of eye movement and participant voice stops a few milliseconds after the blank display.

As we discussed before, it is important to ensure that participants give a quick first response even if they are unsure. The six practise trials are designed so that each is repeated until the participant enunciates their response in a timely fashion. To this effect, the trial sequence keeps track of the timeliness of participant verbal responses using a counter (see Fig. A1c). During the post-trial part of the trial sequence, the workflow proceeds by evaluating the timeliness counter to decide whether to repeat the practise trial and which feedback message to display. Practise trials completed on time are followed by feedback: “Good Speed!” or instruction to “Try to call out the number clearly and as fast as possible after each beep” otherwise. Main test trials, however, are never repeated and have no post-trial feedback. The identity of the displayed digit is hence never divulged.

Verbal Task Output Format and Data Files

Data collected during a Verbal task session is saved under a distinct folder specified at the start of the task session (see Fig. A3a). The folder name contains the

participant's initials and the date of the experiment. A session data folder holds a binary eye movement data file (EDF) of the same name. The EDF contains a sequence of time-stamped records of participant eye movement samples, saccades and fixations in addition to corresponding NIST images and trial events. The data folder also contains the participant's verbal identification in the form of trial-specific wave audio files of six seconds each. In addition, a modified version of the original data source, which is attached to the trial sequence in Experiment Builder, is also generated. The new data source has its rows rearranged to reflect the actual trial order in the current session. An optional subjective evaluation file (*notes.txt*) may also be added by the experimenter at the end of the session. The text file primarily contains eye-tracking calibration or participant alertness issues as observed by the experimenter or reported by the participant during the course of the session. It is used to help determine the reliability rating of session data during data analysis (see p. 53 below).

Unconstrained Identification during Normal Viewing

In order to control for the effects of large digit size and double-prompt digit identification used in the Verbal task due to eye-tracking considerations, a new digit identification task is required that features normal digit size and unconstrained digit identification.

Manual Task Design and Testing

Manual Task Design Considerations

Since no eye movement is recorded and given the overhead associated with processing and labelling verbal response during data analysis, the author opted for manual input using the numeric keypad of a standard IBM keyboard instead. To facilitate

future discussion, we refer to this task as the Manual task. This task is conducted on the same participants and right after they have completed the Verbal task. The purpose of the Manual task is hence to keep a record of how participants would identify the same seventy-four digits with no timing constraints and under normal viewing conditions of scale and contrast. Such data can serve as a reference to assess the extent to which eye-tracking conditions during the Verbal task, like large scale and time constraints, affect identification performance. Furthermore, since manipulation of aliasing and contrast is no longer a requirement, the author opted to use the MNIST version of the digit images instead of NIST for two reasons. First, MNIST images are already size-normalized for normal viewing conditions under which they are virtually indistinguishable from the NIST version. Second, since MNIST digits are widely used in classifier performance benchmarks, human identification data on these digits makes for a more relatable reference. To insure normal viewing conditions during identification, the MNIST digits need to be presented at the centre of the display without scaling. In their original size-normalized dimensions, fitting a 20×20 pixel box or less, MNIST digits span around 0.8° of the participant's visual field when displayed at 1024×768 resolution on a 20-inch monitor at 57cm viewing distance. To ensure that the task is as unconstrained as possible, any temporal controls are removed leaving identification response time completely up to the participant.

Manual Task Pilot Testing and Modifications

The pilot testing was conducted on six participants recruited in class. The purpose of the pilot test was to uncover design flaws and to address any threats to validity that may arise. Based on post-experiment interviews and subjective evaluation, the author determined that the number of practise trials initially set to three was insufficient. This

came after a number of complaints of manual entry errors especially when using the keyboard's numeric keypad. As a result, fourteen additional practise trials were introduced for a total of seventeen in the final version of the Manual task. The Manual task part of the experiment, which takes up to five minutes to complete, was conducted on participants after they have completed the twenty-five minute long Verbal task.

Manual Task Design and Implementation

The Manual task was also designed and implemented using Experiment Builder (see Fig. A2). This part of the experiment has a simpler workflow compared to that of the Verbal task. The first sequence is the introduction explaining the handwritten digit identification task using the numeric keypad of a standard IBM keyboard. The second sequence is the block sequence that, together with the nested trial sequence, yields a practise block with seventeen trials and a main test block with seventy-four trials. The practise and test trials are randomized in the same way as in the Verbal task but unlike the latter, only one viewing condition is used. The original greyscale MNIST digit images are presented to span roughly 0.8° degrees of the participant's field of view on a light grey background with no modification in scale, contrast or smoothing.

Just like in the Verbal task, the trial sequence defines the operations to be executed just before, during and right after each trial. During pre-trial, the *prepare sequence* action node, ensures that all graphical and audio resources are optimally cached to avoid processing overhead during the actual trial. Unlike Verbal task trials, Manual task trial duration depends on how fast the participant identifies the handwritten digit. The actual trial starts with a standard fixation screen showing a cross at the centre of the display. One second into the trial, the cross disappears and the handwritten digit is shown for a maximum of ten seconds. The participant can identify the digit by selecting one of the numerals using the numeric keypad or the alphanumeric keys. Upon identification,

the handwritten digit disappears and a right-wrong feedback sound is played back indicating whether the participant has selected the correct numeral followed by a new trial.

The decision to add right-wrong feedback is mainly due to an important shift in emphasis compared to the preceding task. In the Verbal task, a participant is trained to give a somewhat *rushed* first response followed by an amply delayed second and *final* response. In the Manual task, conducted right after the Verbal task, a participant is asked to give her best and final answer during the first and only response. Therefore, the author opted for right-wrong feedback during Manual task trials to remind participants of the shifting emphasis from speed of identification to correctness of identification.

The workflow during the post-trial segment is slightly different in practise and test trials. The main objective of practise trials is to ensure that participants are comfortable using the numeric keypad. The seventeen practise digits are selected subjectively based on ease of identification. Therefore, erroneous identification made during practise trials is likely due to entry errors. Practise trials incorrectly identified are hence recycled for repetition until their correct identification while test trials are never repeated.

Manual Task Output Format and Data Files

Data collected during a Manual task session is saved under a distinct folder specified at the start of the task session. The folder name contains the participant's initials and the date of the experiment. A session data folder holds the plain text *Results File* containing a set of trial records (see Fig. A3b). Records are tab-separated and contain the participant's manual identification response in the form of a timestamp, the selected numeral in addition to the corresponding MNIST image. An optional subjective

evaluation file (notes.txt) may also be added by the experimenter at the end of the session. The text file primarily describes any numeric keypad entry issues reported by the participant during the post-experiment interview. Like the Verbal task notes.txt file, it is used to help determine the reliability rating of session data during data analysis.

Data Collection Process

Ethical Approval

The ethical approval to recruit and conduct our study was obtained under the banner of the Concordia Vision Lab due to similarity to other ongoing research in general objectives and experimental conditions.

Participant Recruiting

Participant recruiting and testing was conducted during the months of February, March and April 2010. In total, the author recruited seventy-seven volunteers for the main phase of data collection. Sixty-one participated in the Verbal task under one of the seven aforementioned viewing and contrast conditions (see p.53 below). The remaining sixteen completed a drastically modified version of the Verbal task experiment featuring gaze-contingent extra-foveal masking that will not be discussed in this paper. The Manual task part was introduced a week after the start of Verbal task data collection and was completed by sixty-one volunteers only. The author solicited participants via a webpage listing of the university participant pool and through in-class presentations in two undergraduate Psychology courses. When a potential participant expressed their interest in volunteering in the experiment, the author contacted them via email and offered a number of available time slots from which to choose. A confirmation email followed advising them of the time of their participation and illustrated directions to the Concordia Vision Lab. The robustness of the EyeLink II meant that it could reliably track

eye movement of most participants with corrective vision. As such, there were no exclusion criteria for people wearing prescription glasses or contact lenses.

Most participants were female in their early twenties and registered in preapproved *participant pool credit* courses, which entitled them to half a mark bonus on their final course grade. Students whose participation went over one hour were entitled to a full bonus mark. After the experiment, the student submits the participant pool credit form, which has been filled out and signed by the author-experimenter, to the department of Psychology for granting of course bonus. Two participants accepted a nominal monetary compensation of ten dollars as an alternative to course bonus. Three participants failed to complete the experiment due to technical difficulties but were still entitled to the credit or compensation.

Testing Process

The experiment starts with the greeting of the participant and signing of the consent form describing the risks associated with eye-tracking experiment such as the exposure to safe levels of infra-red radiation and advising the participant that they can discontinue the experiment at any time without adverse consequences (see Fig. A4).

The experiment is composed of two handwritten digit identification tasks. The first part is the Verbal task, which uses an eye tracker and a microphone to record participant identification behaviour and lasts around twenty-five minutes. The preparations for this task include seating the participant, fitting of the head-mounted eye tracker, readjustment of eye cameras to ensure proper framing and detection of pupils. The participant is then given the gamepad-like button box and advised of the big button to press in order to advance through the experiment. Next, they are notified of the presence of the desktop microphone and instructed to rest their chin on the chin guard. The room lights are turned off shortly after in order to improve eye-tracking reliability

and reduce visual distractions.

The Verbal task is followed by the Manual task, which uses a keyboard to record identification response and lasts five minutes or less. The preparations for this task include removing the eye tracker, collapsing of the chin guard, turning the lab room lights back on and pulling the display PC's keyboard closer to facilitate the participant's access to it. After the completion of the Manual task, the author-experimenter completes the participation credit form and conducts a short interview with the participant. The primary purpose of the post-experiment interview is to obtain participant self-report. Questions can range from “Did you feel that there were some number entry errors during the second part?”, “Did you change your mind on the second answer because you felt it was another possibility or a better answer?” or “Did you notice that the numbers were the same in both experiments?” to more general enquiries like “How was the experiment?” After the interview, the participant is debriefed about the purpose of the study, how it relates to Handwritten Character Recognition, and a short list of relevant references in the literature (see Fig. A5). After the participant is shown out, the author-experimenter completes a subjective evaluation of the two parts of the experiment. The evaluation record is kept in two *notes.txt* text files along with the participant's experiment data. Subjective evaluation is used to help determine the reliability rating of experiment data, which we discuss next.

Collected Data Reliability Ranking

Fig. 12 summarizes the collected data and its reliability. Sixty-one participants underwent both main Manual and main Verbal tasks. All sixty-one participations in the Manual task were ranked as good with a negligible number of participant-reported manual entry errors. In the Verbal task, on the other hand, fifty-three participations were ranked as *good* with seven and one participations assigned *fair* and *poor* ranking respectively and subsequently excluded from analysis. Five and two participations were

ranked as fair due to experimenter-reported calibration-recalibration issues and participant-reported fatigue respectively. The one poor-ranked participation was due to eye-movement data file corruption resulting in processing errors during conversion to plain text format to be discussed in the next chapter. Overall, 4514 Manual trials, 3883 Verbal trials and 7766 verbal responses (7702 non-empty responses) made it to the data analysis phase.

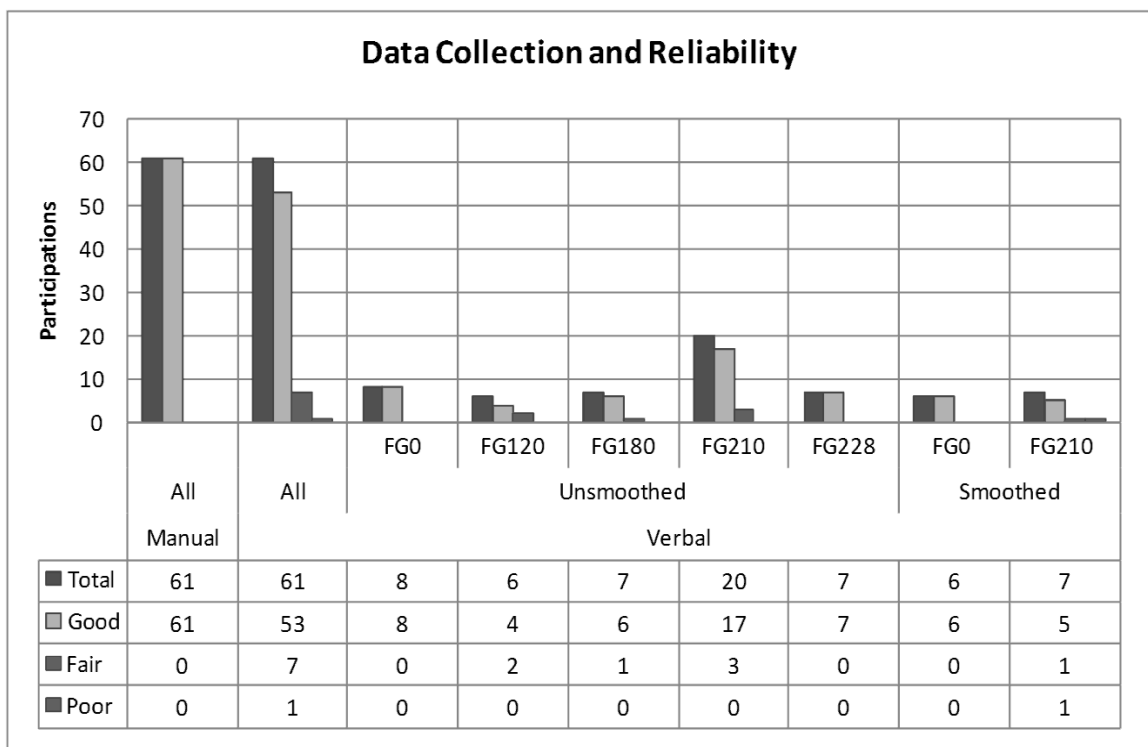


Fig. 12. Summary of collected data and reliability. Sixty-one and 53 participations were considered during analysis of Manual and Verbal respectively. Eight participations in Verbal were assigned *poor* or *fair* ranking and excluded from analysis due to: (1) data processing errors, (2) questionable calibration reliability, or (3) participant-reported fatigue.

CHAPTER 4

DATA ANALYSIS AND DISCUSSION

Before the analysis of collected data can proceed, a number of crucial steps need to be taken in order to consolidate and facilitate access to all relevant data. We start this chapter with a discussion of a number of steps taken and tools developed to prepare Manual and Verbal tasks data for analysis. We then move on to the analysis of identification and visual response data.

Preparation of Identification Data

Verbal Response Isolation and Labelling

As we discussed before, the decision to opt for verbal identification stems from the requirement to collect ten possible responses (0 to 9) which makes manual input inappropriate in the context of eye tracking. Verbal identification, however, has its own challenges; indeed, the most accurate way to identify the numeral in a verbal response is by ear. In addition, it is quite possible that eye movement made during or right before the actual verbal response have special significance requiring accurate isolation. To do so necessitates pinpointing the start and the end of the verbal response. Even though experiments programmed using Experiment Builder have a voice-key trigger node that records the start of a participant's oral response (see Fig. Ac), the author found no discussion or available software components to record the end of the oral response. Furthermore, a participant often produces inadvertent sounds while clearing their throat or saying “ah” prior to enunciating the numeral making any response-timing data

inherently unreliable. The quality and sensitivity of the desktop microphone used in the Verbal task only adds to the problem; during data collection, the author often noticed that the late response error was sounded even when participants clearly gave their response on time. Alternatively, when the author-experimenter increased the voice-key trigger threshold or placed the microphone closer to the participant's mouth to compensate for sensitivity limitations, the system often mistook ambient noise or non-Verbal sounds made by the participant to an actual verbal response and failed to sound the late response error. This makes verbal identification response time as recorded in the EDF file highly unreliable and unsuitable for our purposes.

Although listening to thousands of six-second trial audio files is not very time consuming, the process of accurately isolating verbal response time is. In addition, the process of manually labelling then consolidating the labelled responses of all participants across different viewing conditions is highly error-prone. These issues cast concerns not just over the quality of our own data analysis but also over the feasibility of further efforts to use eye movement for pattern recognition applications.

Evaluation of Available Tools

The search for automated or semi-automated tools to facilitate the processing of verbal response data led the author to evaluate a number of available tools. Among these, an open-source sound recording and editing program called Audacity (Brubeck, Haberman, and Mazzoni 2010) was most promising. The program (Windows version 1.3 Beta) came with an impressive suit of features that allows loading several trial audio files at a time as well as carrying out noise reduction and removal of audio recording artefacts like *clicks*. This, in addition to a labelling feature that allows the user to graphically select audio segments of interest, label their selection, then export the labels in the form of an accessible spreadsheet containing the selection label, start time and end time. A *Sound*

Finder add-on based on amplitude heuristics is also available which automatically identifies and creates a label tag for the corresponding sound segments. Despite striking similarities to features of interest to us, the Sound Finder feature was highly unreliable perhaps due to the presence of breathing noise in many of the trial recordings making solely amplitude-based algorithms all but unusable. The manual labelling feature greatly facilitates labelling and consolidation of verbal identification data into a single spreadsheet; however, the process was still extremely time-consuming and quite error-prone especially when it came to precisely selecting verbal response endpoints then tagging the selected audio segments.

However, the program offers a number of built-in spectral analysis options one of which, *Enhanced Autocorrelation*, the author discovered by accident to be a reliable way to discriminate between spoken audio and breathing noise. The author later discovered that the function was an implementation of the Enhanced Summary Autocorrelation Function (ESACF) described in Tolonen and Karjalainen (2000) and widely cited as an efficient multi-pitch detection technique. Since much of human voice, especially during vowel sounds, has a highly auto-correlating spectral pattern in the form of a fundamental frequency ranging from 80Hz to 350Hz (Plannerer 2005), the ESACF plot shows big even spikes in that range only when speech segments are analyzed. Unfortunately, this scheme often fails to detect fricative sounds, which lack a fundamental frequency like the *s* sound at the beginning and the end of *six* making it limited on its own. A combination of ESACF and amplitude-based heuristics seemed like a very plausible alternative; however, the Sound Finder add-on cannot be customized to include criteria other than amplitude and the author found no implementation of ESACF in the Nyquist scripting language used to program Audacity add-ons.

A common technique in voice-activity detection (VAD) uses amplitude and zero-

crossing rate heuristics and is widely shared in the open-source community under both C++ and MATLAB implementations. Most of the implementations the author evaluated seemed geared to efficient and approximate detection of voice or voice-like audio segments rather than accurate isolation of spoken segments per se. In addition, none offered the manual editing means available in Audacity in case the user chooses to do some fine-tuning.

Development of Verbal Response Isolation and Labelling Tool

Given the wide availability of audio analysis implementations in the form of MATLAB toolboxes, the author chose the high-level programming language to explore alternative solutions more suited to the research requirements. The author used windowing, spectral analysis, and zero-crossing rate functions in the VOICEBOX toolbox (Brookes 2010) to implement a new speech isolation scheme which combines a new, simplified, implementation of ESACF with the amplitude-zero-crossing-rate VAD scheme discussed above (see Fig. A6).

To determine the precise start and end points of an isolated voice segment, the program also detects extrema in the spectral distance and amplitude curves. When a specific number of extrema overlap or are adjacent, a candidate block is marked. The closest candidate within a maximal interval before or after a detected voice segment is then merged with the segment to form the isolation label. A graphical user interface is also available which allows the user to click on a given isolation and enter the numeral it represents. Since a trial audio file may be missing one or both verbal responses, if the participant failed to identify the handwritten digit, the user has the option to create a new isolation label and annotate it with $()$ denoting a missing-empty response. To facilitate further fine-tuning, candidate blocks can be removed or added using mouse clicks.

A batch-mode feature was also added to allow loading and isolation of all trial audio in a given participant's data folder. When the user has completed editing and annotating all isolation labels, two export functions, reminiscent to the ones in Audacity, are used. The first, exports all labels into separate verbal response audio files that can then be conveniently loaded into an audio player play-list and played back for a quick final verification. The second export function converts all labels into a text file of tab-separated rows. Each row represents a single label and contains the trial number in which the verbal response is given, whether it is the first or second response, the start and the end of the response in trial time as well as the numeral identified. The response start and end times are expressed in seconds and precise to 11 milliseconds with an average error of 33 milliseconds (author's rough estimate).

Consolidation of Manual and Verbal Identification Data

Development of Consolidation Tools

The spreadsheet-like text file containing a participant's verbal response labels is copied and pasted into a template-based MS Excel workbook named using the participant's unique ID. Since the trial order is randomized and differs from one participant to another, the participant's own trial data source is needed in order to determine the corresponding handwritten digit. Once copied over into the workbook, the trial numbers and handwritten digits are automatically matched and merged into a new table now containing all verbal response data necessary for analysis as follows:

1. The name of the PNG image file containing the handwritten digit being identified and its true numeral label.
2. Which numeral the participant called out during their first and second responses.

3. The start and the end of the first and second responses in trial time.

When this process is completed for all participants, the individual verbal response workbooks are copied to a special folder for further consolidation.

In the next phase, the author implemented two MS Excel VBA macros embedded in a macro-enabled workbook (XLSM). The two macros use each of the seventy-four handwritten digits to locate and retrieve the corresponding identification response for each participant automatically one viewing condition at a time. At the end of each batch job, a log table reports the count of unsuccessful lookups for each participant to facilitate verifying that all available data has been successfully retrieved. The first macro is used to consolidate the participants' verbal response workbooks while the second is used to consolidate their manual response Results Files described on p. 50 above. The two macros use a consolidation table to combine the identification responses as they are being retrieved one row at a time. Each table row contains the two verbal and one manual identification response details for a given participant on a given handwritten digit. This structuring facilitates cross analysis of first versus second verbal responses, on the one hand, and verbal versus manual responses on the other. To facilitate analysis of verbal identification or manual identification as a whole, a similar approach is used to consolidate identification responses individually into distinct rows.

The verbal and manual identification data is now consolidated and can be conveniently accessed using pivot tables. These powerful data-processing tools are particularly suited for the task of custom filtering, summarizing and visualizing large amounts of tabulated data.

Preparation of Visual Data

The EyeLink system comes with a powerful data analysis tool (Data Viewer) to analyze, filter, consolidate and visualize the eye movement data in the binary EDF files

(*EyeLink data viewer* 2010). However, as we argued during the data collection discussion, the potential usefulness of eye movement data for pattern recognition applications is quite limited without the possibility to correlate it with the corresponding identification response. Since there is no viable way to isolate and label a participant's verbal responses online (i.e., during the experiment), the recorded EDF files have no reliable record of verbal identification of which to make use. One possibility is to use Data Viewer's filtering feature to select visual fixations individually or based on specified periods of interest. However, such manual process is highly error-prone and extremely time-consuming. Another downside to this option is the lack of a batch-mode in Data Viewer's fixation *heat-mapping* feature (version 1.10.1). Of particular interest to us, heat mapping is widely used to visualize a set of related visual fixations of an arbitrary size. It can be best thought of as a colour-coded probability density map representing the average share of gaze that various parts of a given image receive. The source of gaze data can be an arbitrary number of participants viewing the image over an arbitrary period. The ability to generate custom heat maps for the seventy-four handwritten digits would be a great boon to this study since they intuitively communicate the attractiveness of various digit features. However, since Data Viewer only allows the creation of one map at a time, our ability to generate heat maps based on a variety of criteria and periods of interest is severely limited.

The alternative is to implement consolidation, visualization and heat-mapping tools while focusing on a subset of Data Viewer features that are of primary interest. In order to facilitate the use of the tabulated verbal identification data as selection criteria, the author opted to consolidate the visual fixations into a similar tabulated form in MS Excel.

Import and Consolidation of Visual Fixation Data

The manufacturer of the EyeLink system provides a C-based EDF access API and example source-codes demonstrating how to use it. One of the examples, EDF2ASC, is a command-line utility that allows the conversion of a set of binary EDF files into ASCII files (ASC). As such, an ASC is a text file containing an ordered sequence of time-stamped event records of a single eye-tracking session. Given this sequential form, data corresponding to the start, end, duration and screen-coordinates of a single fixation is located in different records throughout an ASC. Since time stamps used in EDF are expressed in experiment-time, trial-specific events, spanning over thousands of other event records, are also required to determine a fixation's start in trial time.

Development of Conversion and Consolidation Tools

In order to facilitate the process of consolidating visual fixations into tabulated form, the author carried out the following modifications on the way EDF2ASC processes a set of EDF files as summarized in Fig. 13:

1. *Fixation Record Creation:* All event records corresponding to individual fixations need to be detected and combined into a single tab-separated record containing the participant ID (name of the EDF file) and fixation-specific data: Trial number, right or left eye, start and end time, duration and screen coordinates.

2. *Trial Record Creation:* All event records corresponding to individual experiment trials need to be detected and combined into a single tab-separated record containing the participant ID and trial-specific data: (a) trial number, (b) trial index, (c) handwritten digit image file name, (d) its true numeral label, (e) trial start time, (f) trial end time, (g) fixation count and (h) start and end times of the trial audio recording. The latter is needed since it differs from the start and end times of the actual

trial by a few milliseconds.

3. *Consolidation of Fixation and Trial Records*: In order to facilitate the process of mapping fixations to trials, the two sets of fixation and trial records need to be consolidated separately in two spreadsheets (XLS). As such, instead of one ASC per EDF the conversion produces two consolidated XLS files for all EDFs (Fig. 13).

The figure displays two views of EDF2ASC output. The top view is a Notepad window showing raw EDF data for participant 'ab_mar09'. The bottom view is an Excel spreadsheet showing consolidated trial and fixation data for the same participant.

Top View (Notepad): Shows raw EDF data with various event types and timestamps. Key events include:

- MSG 11914425 12 !V ARECSTART 0 1 2 ts_044986.tif.png.wav
- MSG 11914477 !MODE RECORD P 500 0 0
- MSG 11914482 LEFT RIGHT EVENTS
- MSG 11914490 !V APLAYSTART 0 8 library\audio\ding100ms.wav
- MSG 11914537 -7 ES_DIGITSCREEN
- MSG 11914538 -6 !V DRAW_LIST .../runtime/dataviewer/ab_mar09/g
- MSG 11914546 ES_RECORDERAUDIO
- MSG 11914655 -14 !V APLAYSTOP 5491 8 library\audio\ding100ms.wav

Bottom View (Excel): Shows consolidated trial and fixation data for participant 'ab_mar09'.

Trial Data Table:

PARTICIPANT	TRIAL No.	TRIAL Start	TRIAL End	TRIAL Stim	TRIAL Index	TRIAL Label	AUDIO Stc	AUDIO Na	FIXATION Count
ab_mar09	1	11914482	11922063	2_nist044	1	2_11914425_11920416_1_2_ts_04			42
ab_mar09	2	11928912	11926275	2_nist044	1	2_11914836_11934847_2_2_ts_04			37
ab_mar09	3	11939100	11946257	8_nist328	5	8_11939047_11945037_3_8_tr_32			29
ab_mar09	4	11948648	11955821	4_nist235	3	4_11948597_11954588_4_4_tr_23			38
ab_mar09	5	11958162	11965377	9_nist140	6	9_11958108_11964098_5_9_tr_14			39

Fixation Data Table:

PARTICIPANT	TRIAL No.	FIX Event	EYE	FIX Start	FIX End	FIX Durati	FIX X	FIX Y	FIX Acceleration
ab_mar09	1	EFIX	L	11914490	11914800	312	512.4	381.1	4450
ab_mar09	1	EFIX	R	11914490	11914800	312	510.2	381.6	4311
ab_mar09	1	EFIX	L	11914838	11914968	132	535.3	278.7	4427
ab_mar09	1	EFIX	R	11914838	11914968	132	522.4	270.1	4213
ab_mar09	1	EFIX	L	11915022	11915176	156	374	469.8	4270
ab_mar09	1	EFIX	R	11915022	11915188	168	346	468.3	4250
ab_mar09	1	EFIX	L	11915354	11915644	292	477.5	348.7	4409
ab_mar09	1	EFIX	R	11915342	11915644	304	480.9	337.7	4227

Fig. 13. Output comparison of original and modified EDF2ASC tool. *Top*, two views of the same ASCII eye movement data of a single participant after conversion from EDF using original EDF2ASC. *Bottom*, the corresponding visual fixation and trial data organized in tabulated form along with data of all participants in Verbal task output by modified EDF2ASC.

Visual Fixation Filtering and Isolation

Visual Fixation Selection Scheme

One of the advantages of consolidating large amounts of data records of the same type in MS Excel spreadsheets is the availability of powerful selection tools like *Advanced Filters*. This feature allows convenient filtering of tabulated data based on multiple criteria at a time. Each criterion can be set to match a specific value or a range of values in the respective table column. Therefore, in order to increase the number and variety of possible selection criteria, the fixations table needs to consolidate as many fixation-related details as possible at the risk of some redundancy. To that end, the author carried out the following additional consolidations to the tabulated fixation data as outlined in Fig. 14:

Fixation Data

B	D	Q	R	S	T	U	V	W	X
PARTICIPANT	TRIAL No.	EYE	FIX Start	FIX End	FIX Durati	FIX X	FIX Y	FIX Acceleration	
ab_mar09	9	R	12161708	12161986	280	516.9	385.9	4205	
ab_mar09	9	L	12161708	12161990	284	517.5	384.1	4511	
ab_mar09	9	L	12162016	12162186	172	494	339.9	4479	
ab_mar09	9	R	12162016	12162186	172	487.6	342.7	4110	
ab_mar09	9	L	12162220	12162406	188	488.2	477.1	4511	
			12162224	12162422	200	467	475.5	4321	

Trial Data

PARTICIPANT	TRIAL No.	TRIAL Start	TRIAL End	TRIAL Stim	TRIAL Inde	TRIAL Lab	AUDIO Sta	AUDIO Stc	AUDIO Na	FIXATION Count
ab_mar09	9	12161700	12168839	8_nist012	77	8	12161648	12167638	9_8_ts_01	34

Verbal Response Data

Lookup	Participant	Condition	Trial Audi	Stimulus1	Stimulus	Irregular	Label	V1	V1 Correct	V2	V2 Correct	V1-V2	V1 Start	V1 End	V2 Start	V2 End
ab_mar099	ab_mar09	FG210	9_8_ts_01	1071	8_nist012	0	(c)	8	1	8	1	1	0.725333	1.162667	5.03725	5.638542

New Fixation Data

Look	CON	DIGIT ID.	IRREGULAR	LABEL	V1	V1 Correct	V2	V2 Correct	V1-V2	TRIAL Tim	AUDIO Tim	FIX T. Lab	FIX Lapse
ab_mar099	FG210	1071	0	(b)	8	1	8	1	1	(8)	(60)	(bV1)	(-665.333)
ab_mar099	FG210	1071	0	(d)	8	1	8	1	1	(a)	(b)	(g)	(-633.333)
ab_mar099	FG210	1071	0	8	8	1	8	1	1	316	368	bV1	-357.333
ab_mar099	FG210	1071	0	8	8	1	8	1	1	520	572	bV1	-153.333
ab_mar099	FG210	1071	0	8	8	1	8	1	1	524	576	bV1	-149.333

Fig. 14. Addition of support columns to facilitate fixation filtering and isolation. Due to figure width constraints, the fixation table was split into two views: *Top*, Fixation Data and *Bottom*, New Fixation Data. The actual order of table columns can be deduced from displayed column letters.

1. *Fixation trial time*: Add a formula column to determine fixation start in trial time as follows: Number of milliseconds elapsed between trial start and fixation start (Fig. 14a).

2. *Fixation audio time*: Add a formula column to determine fixation start in trial-audio time as follows: Number of milliseconds elapsed between the start of the recording of trial audio and fixation start (Fig. 14b).

3. *Fixation Viewing Condition*: Add a formula column to determine corresponding viewing condition by looking up the participant ID in a table listing all participants and their corresponding viewing conditions (Fig. 14c).

4. *Fixation Digit*: Add three formula columns to determine handwritten digit ID, whether it is regular or irregular and its correct numeral label by looking up the image file name in a table containing handwritten digit details (Fig. 14d).

5. *Fixation-to-Verbal Lookup*: In order to facilitate future matching with corresponding verbal identification data, add a formula column that concatenates participant ID and trial number. Add similar column in all verbal identification data tables. We refer to these as lookup columns (Fig. 14e).

6. *Fixation Verbal Answers*: Add five formula columns that use the lookup columns to determine, for each fixation, the first verbal answer and whether it is correct, the second verbal answer and whether it is correct and whether the two answers are the same (Fig. 14f).

7. *Fixation Time Label*: Add a formula column that uses the lookup columns and fixation audio time to determine the fixation's trial time label based on when the fixation starts relative to the two verbal responses: Before the first (*bV1*), during the first (*V1*), between the first and second (*bV2*), during the second (*V2*), or after the second (*V2+*) verbal responses (Fig. 14g).

8. *Fixation Lapse*: Add a formula column that uses the lookup columns, and fixation audio time to determine the fixation time lapse in milliseconds relative to the start of the respective verbal response. During analysis, an Advanced Filter can be applied by specifying a criterion on this column along with another criterion on the Fixation Time Label to select all fixations that occurred within a specific interval of a given verbal response. For example, criterion =bV1 along with criterion >-1000 selects all fixations that occurred within 1000 ms prior to the first verbal response; Criterion =V2 along with criterion <100 selects all fixations, if any, that occurred during the first 100 ms of the second verbal response and so on (Fig. 14h).

Fixation Lapse data are of particular interest since they facilitate selection of fixations based on their temporal relation to the respective verbal answer. The visual fixations are now fully consolidated and ready for selection.

Eye Movement Visualization Scheme

An intuitive and commonly used eye movement visualization scheme displays trial fixations and saccades overlaid on top of the visual stimulus presented to the participant during that trial. Fixations are presented as dots or circles while saccades are represented by lines or arrows linking the fixations. In EyeLink Data Viewer (2010), this scheme is referred to as *spatial overlay* (Fig. 15b). Fixation duration is expressed by proportionately varying the circle's diameter while the circle's label displays other fixation data of interest.

Since we are not interested in saccades or the order of fixations in the present study, MS Excel 2007 *Bubble Charts* can yield similar results for our purposes. Plot bubbles can be parameterized using four cell ranges for bubble *x- and y-* coordinates, widths (or areas) and labels corresponding to fixation coordinates, duration and start time respectively. A bubble chart can also have a background image that can be set manually

or programmatically. By formatting the chart's plot area properly the spatial overlay scheme is effectively reproduced (Fig. 15a).

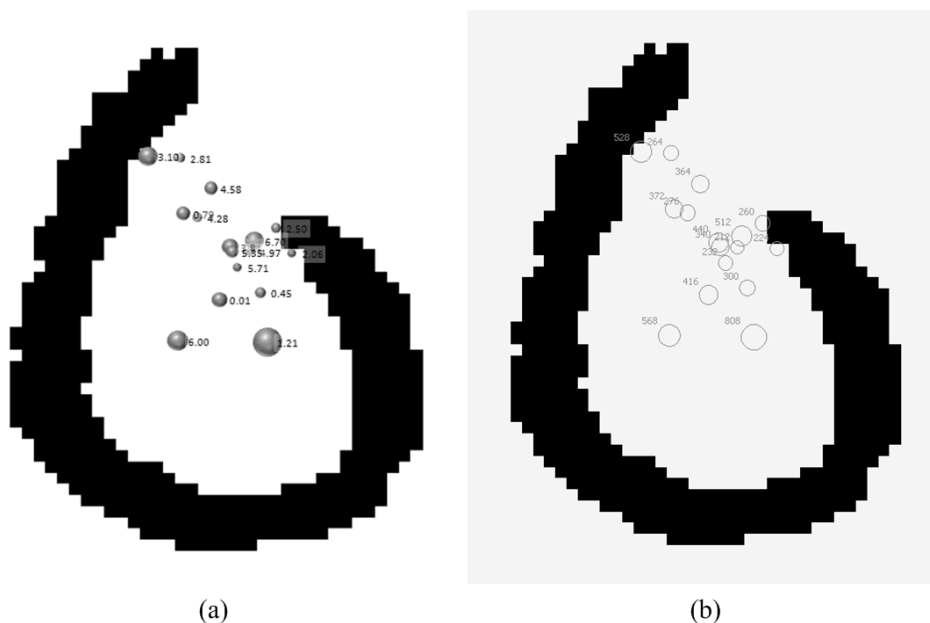


Fig. 15. Side-by-side comparison of (a) bubble chart and (b) Data Viewer's spatial overlay showing visual fixation data from the same trial. Bubble and circle sizes are both proportional to fixation duration. Spatial overlay circles are also labelled with fixation duration in milliseconds while bubbles are labelled with fixation trial time in seconds.

Selective Plotting of Visual Fixations

To accelerate the process of visualizing a set of trial fixations, the author implemented a simple MS Excel VBA macro. In one common scenario the user picks a participant ID, viewing condition and handwritten digit criteria from the respective criterion drop-down lists. When a button is clicked, the macro applies the Advanced Filter set of criteria and copies resulting fixations over to the plot worksheet. At this point four *Dynamic Named Range* objects automatically expand or contract to span the newly selected fixation data. The bubble chart, whose data source points to these range objects, is updated and new bubbles are plotted according to the selected x - and y - coordinates,

fixation durations and fixation start times. Finally, the macro uses the specified digit ID and viewing condition to locate and set the corresponding image file as the chart's background image hence producing the effect of spatial overlay (Fig. 15a). In some scenarios, the user may wish to select fixations from multiple trials at a time. In this case, general statistics about the resulting trials and fixations may be of interest. To that end, the macro also calculates the number of unique participant IDs, the total row count and maximum duration in the selection results. Three worksheet cells are then used to report these statistics in the form of trial count, fixation count and duration of the longest fixation respectively. Given the potentially large number of selection criteria applied at a time, the macro parses the entire row of Advanced Filter criteria summarizing it into a more compact form in a fourth worksheet cell.

Development of Heat-Mapping Tool

The use of fixation plots is restrictive in two important ways. First, plots fail to communicate the repetitiveness and combined duration of gaze when any more than a few fixations are displayed. This holds true in most cases since some image parts are often revisited while most receive little to no direct gaze. In a fixation plot, this yields a set of crowded and overlapping circles virtually hiding, instead of emphasizing, image regions of high importance. Second, fixation plots only communicate the parts of the image directly fixated, which represents a small fraction of the region actually perceived by the visual system. Therefore, in order to effectively communicate how much, how long and how often various parts of an image are perceived a visualization scheme must be able to combine and smooth fixation data in a biologically plausible manner.

Creation of Fixation Heat Maps

In the EyeLink Data Viewer Manual (2010) the process of creating a *fixation map*

(heat map) describes applying a 2D Gaussian at the location of each fixation then adding the results to the corresponding region of an internal map. The map is then normalized and applied to a colour or brightness scale to create the heat map. The Gaussian is set to a default standard deviation $\sigma=1^\circ$, based on a 2° -foveal field of view, while the height of the Gaussian depends on the type of heat map selected by the user. In count-based heat maps, used to express the repetitiveness of gaze in various parts throughout a displayed stimulus, the Gaussian is described to be weighted equally at each fixation. In duration-based heat maps, used to express the combined duration of gaze in various parts throughout a displayed stimulus, the Gaussian is described to be weighted by the duration of individual fixations. Caldara and Miellet (2011), describe a similar method for creating statistical fixation maps. Initially, each fixation is represented by a point with the same x - and y - coordinates and an intensity value proportional to the fixation duration. Then, a Gaussian convolution filter is applied on the intensity image resulting in the equivalent fixation heat map. The author used the MATLAB Image Processing Toolbox (2009) to implement a heat map-generating script that follows the same process and gives maps very similar to those generated using Data Viewer (Fig. 16).

The script takes the following list of parameters:

1. *Fixation details*: Three real-valued vectors of equal size describing duration, x - and y - coordinates of an arbitrary set of fixations.
2. *Image file paths*: Two character arrays specifying (a) the location of the handwritten digit image to use as the heat map background and (b) the path to the output folder where the generated heat map will be saved.
3. *Heat map description*: A character array describing the source of fixations including (a) the list of selection criteria and (b) a scalar value describing the number of trials from which the fixation set was obtained.

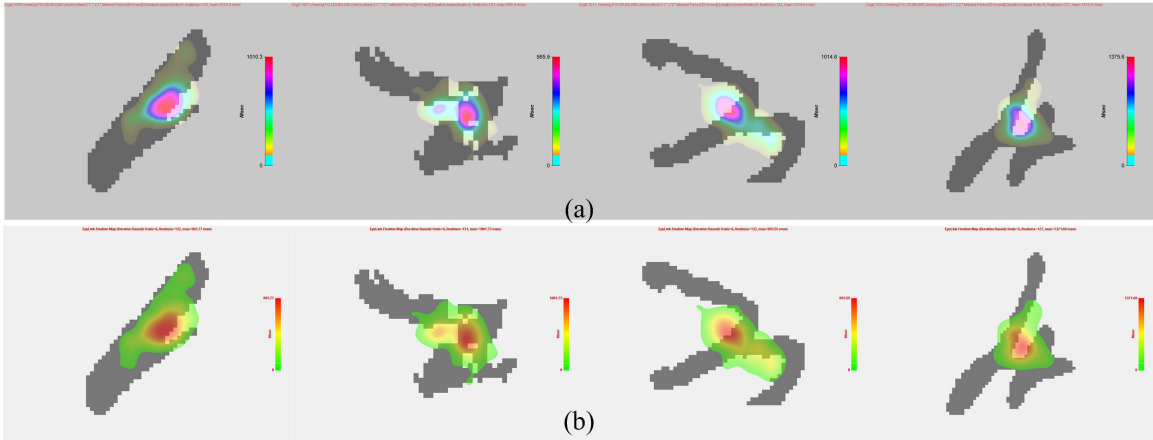


Fig. 16. Comparison of duration-based heat maps generated using: (a) Gaussian convolution and (b) SR Research Data Viewer on the same set of digit stimuli and visual fixations taken from all 6 participants in U-FG120 (see Fig. A22 for a larger version).

The heat-mapping script proceeds as follows:

1. *Creation of fixation matrix*: Use x - and y - coordinates of the set of fixations to obtain their respective subscripts into a 2D matrix of the same size as the digit image (1024×768). Then, set matrix values at these subscripts to one, for count-based heat maps, and to the duration of the respective fixation in milliseconds, for duration-based heat maps. Use function *accumarray* to handle overlapping fixations.
2. *Gaussian convolution*: Create a two-dimensional Gaussian low-pass filter with $\sigma=24$ (number of pixels in 1°) using function *fspecial*. Then, perform convolution filtering using function *imfilter* and the Gaussian filter on the fixation matrix to obtain the filtered matrix.
3. *Creation of gaze map*: Normalize the resulting matrix to the greyscale index range from black=0.0 to white=1.0. To impose an activity cut-off limit similar to the one used in Data Viewer, whereby the bottom 10 percent least perceived regions are ignored, suppress all greyscale indexes <0.1075 .
4. *Creation of overlay heat map*: Use a MATLAB toolbox that specializes in

image overlaying (SC) (Woodford 2007) and select the probability density theme that uses a hue-saturation-value-based (HSV) colour map. Using this colour theme, the gaze map channel is plotted in the foreground by modulating the hue while the handwritten digit image is plotted in the background by modulating the greyscale intensity.

5. *Heat map annotation:* Prior to displaying and saving the heat map to the specified output path, the script annotates the heat map with the specified descriptive text, specified number of trials and an estimated peak value in the count- and duration- based heat maps (Fig. 17).

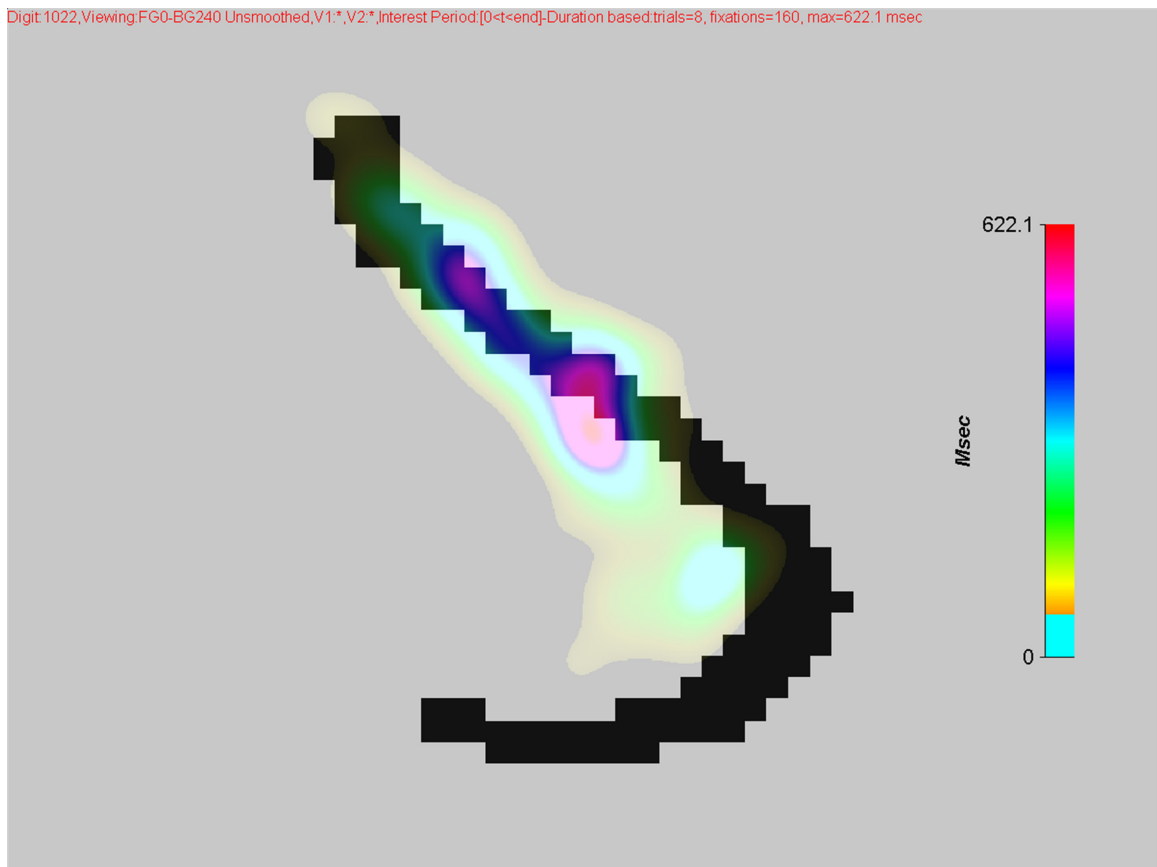


Fig. 17. Duration-based heat map and annotation describing source and statistics of used visual fixations: *Digit ID 1022 (MNIST no. 1394) in Unsmoothed-FG0 with any first response and any second response from trial start to trial end based on 160 fixations in 8 trials. Region colour-coded in red received a combined gaze of 622.1ms (peak value).*

The author steered away from the red-yellow-green colour scale used in Data Viewer (Fig. 16b) primarily because the orange and green shades, which convey very different levels of gaze activity, are hard to discern for people with partial colour blindness (of which the author is one.) This colour scale is also non-standard and the author found no corresponding colour map in MATLAB.

Heat Maps and Gaze Span

In addition to count- and duration- based heat maps, we are also interested in studying eye movement eccentricity across different criteria. To this end the script also calculates and returns the following scalar values as crude measures of gaze spatial span and fixation spread (here, heat map refers to the colour-coded parts of a heat map image):

1. *Filled heat map area*: The filled area of the count- and duration- based heat maps: First, the indexed heat map image is converted into a black and white (binary) image. Second, the MATLAB Image Processing Toolbox function *regionprops* is used to determine the area of each connected region of the heat map image including any holes they may contain. Last, areas of individual regions are added to find the total filled area of the heat map in square pixels.

2. *Heat map bounding box area*: The bounding box area of the count- and duration- based heat maps. To calculate the bounding box of a given heat map, the indexed heat map image is first convert into a binary image. Since a heat map may contain more than one connected region, the total bounding box of the binary image is determined by finding the indexes of the first and last rows and columns with a non-zero pixel. The resulting width and height are then multiplied to determine the heat map bounding box area in square pixels.

3. *Heat map standard deviation*: The standard deviation of the two-dimensional fixation matrices used during the creation of the count- and duration- based heat maps.

Selective Heat Mapping of Visual Fixations

In order to make use of the fixation filtering and selection features implemented in MS Excel, the author deployed the heat-mapping script in the form of an Excel add-in using MATLAB Builder EX (2009). This allows the use of selected fixation data in a worksheet to generate custom heat maps at the click of a button. When the user is done specifying fixation criteria, the *Select Fixations* button is pressed to apply filter criteria producing a new set of selected fixations. To generate the corresponding heat map, the user then clicks on the *Heat Map* button, which calls on the heat map add-in with the required parameters discussed before: (1) Selected fixation details, (2) paths to the digit image and heat map output folder, (3) a short description containing summary of criteria and number of selected trials. At this point the embedded MATLAB script generates, displays and writes the heat map to the output folder then returns heat map area and standard deviation statistics to the MS Excel add-in. The add-in then reports the statistics on the worksheet using six reserved cells.

Batch-Mode Selective Heat Mapping

The suit of MS Excel VBA macros and MATLAB scripts implemented for consolidation, selection and heat mapping allows us to analyze fixations based on a wide range of selection criteria. Unavailable in EyeLink Data Viewer or any other analysis tool the author has come across, these criteria enable selection of fixations based on the identity of and time difference from the associated verbal responses. However, the sheer complexity of cognitive processes unfolding during the Verbal task suggests that any applicability to Pattern Recognition may require analysis of a wide range of selection criteria. Without the ability to automate the process of heat mapping based on several sets of selection criteria at a time, such exhaustive analysis comes at a great cost in error-prone manual labour. Here, once again, the availability of versatile automation features in

the spreadsheet world opens new possibilities in the form of batch-mode heat mapping.

To that end, the author implemented another VBA macro that uses a table of selection criteria to generate custom heat maps automatically one row at a time. When the heat maps to be generated have many criteria in common, as is often the case, the use of a spreadsheet also accelerates the creation of the selection table while minimizing room for error (Fig. 18).

A	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
J	PARTICIPANT			Digit ID.								TRIAL Time				EYE					
	F:\kb_feb23			1001								<6000				R					
	F:\kb_feb23			1002								<6000				R					
	F:\kb_feb23			1003								<6000				R					
	F:\kb_feb23			1004								<6000				R					
	F:\kb_feb23			1005								<6000				R					
	F:\kb_feb23			1006								<6000				R					

B	C	D	E	F	G
CountHeatArea	CountHeatBBox	CountHeatSD	DensHeatArea	DensHeatBBox	DensHeatSD
45430	99428	0.10192405	23998	51597	0.07556436
38642	88995	0.097284927	31488	72225	0.079244548
43466	114540	0.088351996	39046	106602	0.086772054
43590	66856	0.098197495	36725	63624	0.081389225
46932	73458	0.102995593	38146	64922	0.085144065
35703	281996	0.091390259	32063	275880	0.087077861

Fig. 18. A snippet of a batch-mode selection table. Due to figure width constraints the table was split into two views: *Top*, Advanced Filter columns where each row is set to produce count- and duration- based heat maps for a given participant based on a given handwritten digit using right-eye fixations during the six-second trial period; *Bottom*, heat map statistics columns where each row shows the corresponding output heat map statistics; *left to right*: count-based heat map area, its bounding box area, its standard deviation, duration-based heat map area, its bounding box area, and its standard deviation.

In addition to the Advanced Filter selection criteria, the selection table features a column specifying the output path where the generated heat maps will be stored. It also has six initially empty columns reserved for storing statistics returned by the embedded heat-mapping script. When the macro is done selecting fixations and generating the count- and duration- based heat maps corresponding to a row of criteria, it uses these free

fields to store corresponding heat map statistics before moving on to the next row. When the batch job completes and all rows have been processed, the selection table can serve an additional function: Table contents can now be archived along with selection tables from other batch jobs for the purpose of analyzing fixation spread across a wide range of viewing conditions and criteria.

Analysis of Identification Response Data

Before we delve into the analysis of identification response data, we reiterate the main purposes of the various experimental conditions used during data collection. The Manual task was introduced in order to evaluate the extent to which eye-tracking considerations such as large-scale digits, aliasing artefacts and double-prompt identification *fundamentally* affect the way participants identify handwritten digits. The five unsmoothed contrast conditions of the Verbal task, on the other hand, were added to explore the effect of luminance contrast changes on the eccentricity and spread of visual fixations. Last, the two smoothed conditions of the Verbal task were introduced in order to evaluate the effect of aliasing jaggies combined with changes in luminance contrast on the way participants identify handwritten digits and patterns of their visual fixations. Data collected under all seven Verbal task viewing conditions can also serve to explore Optimal Viewing guidelines for similar research in the future. Hence, the purpose of these experimental conditions is primarily to serve as references to evaluate the validity and reliability of collected data rather than to investigate a cause-effect relationship between a specific display parameter and a particular response measure. As such, the author is careful not to make any conclusive claims in the identification response analysis presented next.

Identification Rate and Response Time

We start by evaluating the effect of contrast and smoothing on identification performance in Verbal task.

Identification Rate across Viewing Conditions of Verbal Task

Fig. 19 shows the correct identification rate for regular, irregular and all seventy-four digits across viewing conditions of the Verbal task. All viewing conditions show much higher identification rates for regular compared to irregular digits and slightly higher rates during second compared to first response. In Fig. 19a, the overall correct identification including both responses show very small variations across unsmoothed conditions with an important drop under smoothed viewing particularly for irregular digits. This is perhaps due to unintended distortions in the smoothing process (See Fig. A7-21). Fig. 19b shows no evidence of luminance contrast level affecting correct identification during the first verbal response with fluctuations possibly due to individual differences (see Fig. A26a).

In Fig. 19c a peculiar pattern emerges whereby correct identification in second verbal responses shows a slight increase with decreasing luminance contrast. This is somewhat counterintuitive. A possible – albeit rather convoluted – explanation relates to the perceived task difficulty and confidence in first response. The fifty-four irregular digits may be challenging to pattern classifiers but, as results clearly indicate, most are quite recognizable by humans. As luminance contrast drops, however, participants become less confident with their first *rushed* response and may hence be less reluctant to change their answer when change is warranted. Although these small variations may be due to individual differences, it is worth noting that, in overall Verbal response, lower contrast conditions under smoothed (FG210) and unsmoothed (FG210, FG228) viewing show slightly better identification rate compared to counterparts in higher contrast

conditions.

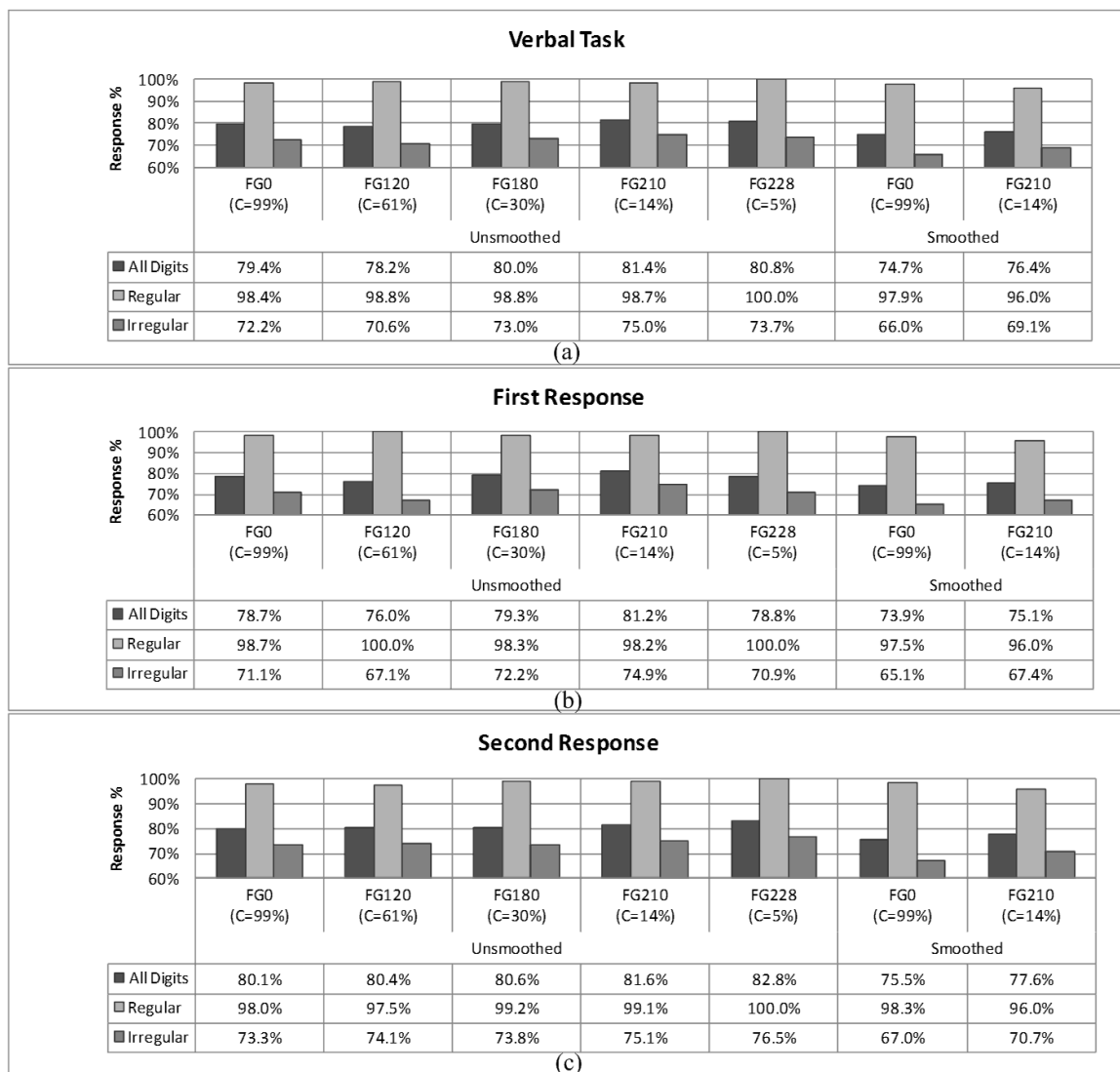


Fig. 19. Percentage of correct identification for regular, irregular and all 74 digits across viewing conditions of Verbal. Horizontal labels show viewing condition name and luminance contrast. Data based on: 8 participants in U-FG0, 4 in U-FG120, 6 in U-FG180, 17 in U-FG210, 7 in U-FG228, 6 in S-FG0, and 5 in S-FG210. Empty verbal responses are considered incorrect and represent: 0.87% in U-FG0, 2.08% in U-FG120, 0.93% in U-FG180, 1.36% in U-FG210, 0.93% in U-FG228, 1.23% in S-FG0 and 0.19% in S-FG210 (see Fig. A25 for detailed response rate statistics).

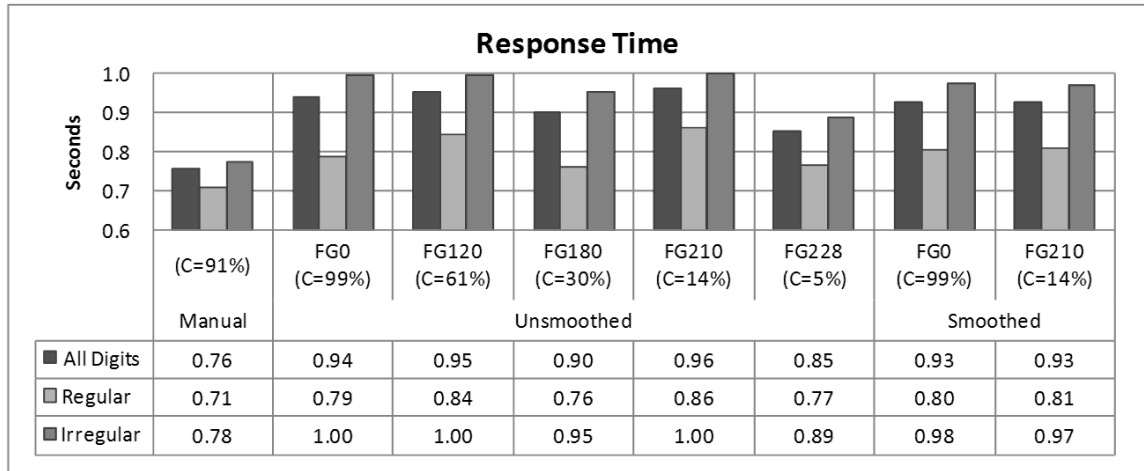
Identification Speed across Viewing Conditions of Verbal Task

Another identification performance metric is speed. Since participants were instructed to delay their second verbal response until they have heard the second prompt 4.5 seconds into the trial, only the first verbal response time can be considered a valid measure of identification speed. Fig. 20a shows average first response time in seconds across viewing conditions while Fig. 20b presents the corresponding correct identification rate for reference purposes. Here, empty responses are excluded from analysis. Just like in identification rate data, response time data shows no clear relationship between display contrast and performance and variations are possibly due to individual differences (see Fig. A26b).

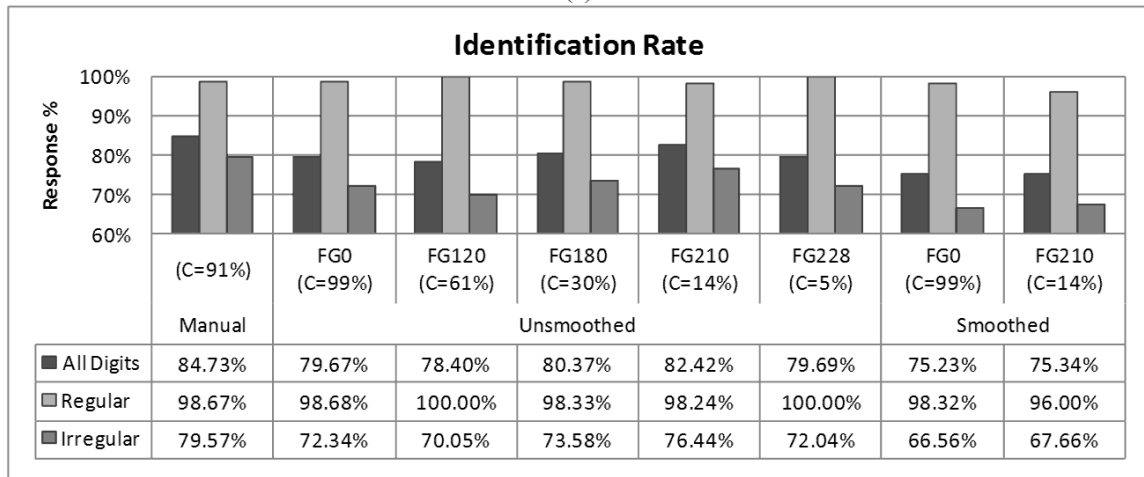
One striking exception is the response time in the lowest contrast condition (F228) which shows a sharp drop (0.85 seconds) compared to the highest contrast condition (0.94 seconds) and the average of unsmoothed conditions (0.92 seconds). This is both counterintuitive and difficult to justify since the literature on contrast and visual perception reports slower reading and visual search speeds with decreasing luminance contrast (Ojanpää 2006; Legge and others 1997; Ojanpää and Näsänen 2003). Furthermore, this increase in identification speed is: (1) most pronounced for irregular digits, (2) comes at little cost in terms of correct identification rate, and (3) is too big to dismiss based on individual differences alone.

A possible explanation may relate to an interaction between luminance contrast and aliasing artefacts. For instance, it is plausible that the presence of jaggies presents an important distraction in higher contrast conditions particularly for irregular digits and that such effect diminishes with decreasing contrast. This may also explain two additional observations: (1) FG210 performance on irregular digits in the first response has a significantly higher correct identification rate (76.4% compared to 72.3%) with the same

response time compared to FG0 under unsmoothed viewing but not under smoothed viewing; (2) response time on irregular digits is slightly lower (faster) under smoothed viewing than in the same luminance contrast under unsmoothed viewing (0.98 and 0.97 compared to 1.00 seconds).



(a)



(b)

Fig. 20. (a) Average response time and (b) percentage of correct identification for manual response and first verbal response across all viewing conditions; horizontal labels show condition name and luminance contrast. Data based on: 61 participants in Manual, 8 in U-FG0, 4 in U-FG120, 6 in U-FG180, 17 in U-FG210, 7 in U-FG228, 6 in S-FG0, and 5 in S-FG210. Empty first verbal responses (1.42%) are not counted.

Identification Performance in Manual and Verbal Tasks

Data collected under the Manual task is a valuable metric to gauge the extent to which Verbal task constraints like large-scale digits, aliasing artefacts and double-prompt identification interfere with the way participants identify handwritten digits. Given the relatively small differences in correct identification rates across contrast conditions, we focus our comparison on identification performance under the Manual task to the overall performance for first and second verbal responses under smoothed and unsmoothed viewing. Here, it is worth reiterating two additional differences between the two tasks. First, Manual task stimuli were obtained from the lower resolution MNIST database whereas Verbal task stimuli were derived from the NIST database; however, as we mentioned before, the two sets of digits are virtually indistinguishable under normal viewing size. Second, participants completed the Verbal task under one of the seven viewing conditions prior to undergoing the Manual task. Although when asked, most exclaimed that they did not realize the digits were the same in the two tasks, one cannot dismiss the possibility that familiarity may have affected task performance in some way.

Having said that, Fig. 20 shows very strong evidence that identification speed is significantly higher under Manual task compared to either response under Verbal task for regular and irregular digits alike. This is hardly surprising since Verbal task digits are much larger than under normal viewing possibly forcing the visual system to execute some eye movements before the digit's identity can be reached with some confidence. Fig. 21 shows that Manual task performance is also significantly higher in terms of correct identification rate for irregular digits compared to either response under Verbal task. When it comes to correct identification of regular digits, on the other hand, Manual task and unsmoothed conditions of the Verbal task are very similar. These results are a reassuring sign that despite important differences in some performance metrics, the

identification task remains fundamentally the same irrespective of eye-tracking design considerations.

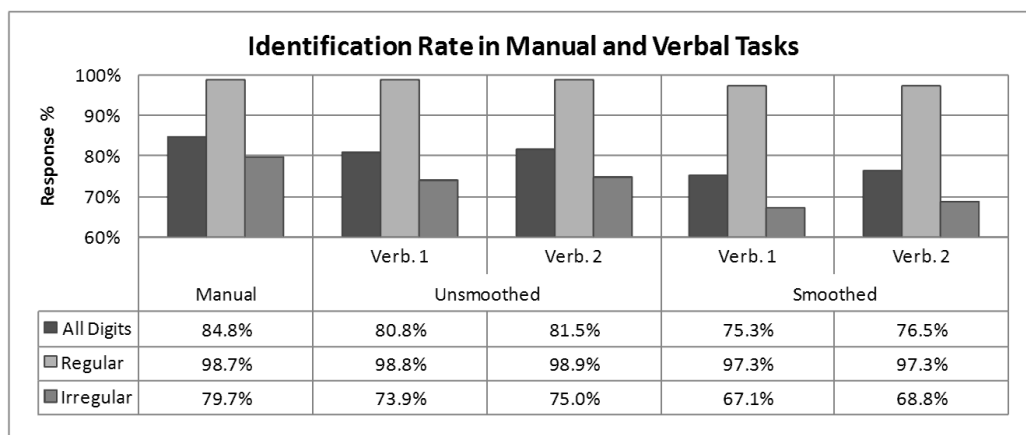


Fig. 21. Digit identification rate in Manual compared to those in first and second verbal responses under smoothed and unsmoothed conditions of Verbal. Data based on: 61 participants in Manual, 8 in U-FG0, 4 in U-FG120, 6 in U-FG180, 17 in U-FG210, 7 in U-FG228, 6 in S-FG0, and 5 in S-FG210. Empty first verbal (1.42%) and empty second verbal (0.26%) responses are not counted.

Digit Identifiability and Ambiguity





































As noted previously and supported by data analysed so far, many of the fifty-four digits that motivated this study are primarily irregular-ambiguous from a machine learning point of view and not for man. Identifying the irregular digits that were also challenging to participants may greatly help us focus our efforts on a smaller set of eye movement data of prime interest; hence, before we move on to the analysis of visual fixations, an evaluation of digit ambiguity and the lack thereof is in order. Next, we break down our evaluation into one of the following metrics: (1) Most and least identifiable digits – digits that were most and least correctly identified, (2) most misidentified digits in each numeral, (3) confusion pair digits – digits that best represent the most common confusion-error in each numeral, (4) most confusing digits – digits that were confused with the most numerals and (5) most re-identified digits – digits for which participants

changed their answer the most. For the exception of the last ambiguity category, we also present the corresponding ambiguous digits under Manual task for reference purposes. Given the relatively small differences in identification performance across contrast conditions, we focus our evaluation on ambiguity under Manual task, unsmoothed and smoothed viewing of Verbal task. Since regular digits are rarely misidentified, we exclude them from the evaluation and focus on presenting irregular digit statistics.

Most and Least Identifiable Handwritten Digits

We start the ambiguity evaluation with the most identifiable of digits. Table 1 shows some of the most correctly identified irregular digits under Manual task and unsmoothed and smoothed conditions of Verbal task.

Table 1. Most correctly identified irregulars and how they were identified in Manual, unsmoothed and smoothed conditions of Verbal.

Manual												
MNIST no.	#4206	#3781	#6560	#248	#1183	#1983	#1045	#1879	#4880	#2897	#4762	#2940
												
Numeral	2	4	4	4	6	6	6	8	8	8	9	9
Response	2	4	4	4	6	6	6	8	8	8	9	9
Correct	C	C	C	C	C	C	C	C	C	C	C	C
Correct%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Verbal (Unsmoothed)												
MNIST no.	#2714	#4206	#3942	#6560	#3781	#1183	#1045	#1040	#1879	#3024	#2897	#4880
												
Numeral	0	2	4	4	4	6	6	7	8	8	8	8
Response	0	2	4	4	4	6	6	7	8	8	8	8
Correct	C	C	C	C	C	C	C	C	C	C	C	C
Correct%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Verbal (Smoothed)												
MNIST no.	#2714	#4206	#3942	#248	#1183	#1983	#2136	#1015	#1040	#3226	#8409	#2897
												
Numeral	0	2	4	4	6	6	6	6	7	7	8	8
Response	0	2	4	4	6	6	6	6	7	7	8	8
Correct	C	C	C	C	C	C	C	C	C	C	C	C
Correct%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%































Notes: Sorted by numeral. Data based on: 61 participants in Manual, 42 in unsmoothed and 11 in smoothed. Top left, In Manual task, digit #4206 is a 2 and was identified correctly in all responses. See Fig. A27 and tables A1-3 for further details

All of the above digits were identified correctly by all participants under different tasks. A complete listing of all digits and their identifiability can be found in appendix tables A1-3.

The identifiable nature of table 1-digits and their recurrence across tasks is hardly surprising. One exception is *zero*-digit no.2714 (MNIST sequence number), which has perfect identification under unsmoothed and smoothed Verbal conditions alike but was misidentified in about 10% of responses under the Manual task as we see shortly.

In the category of least identifiable digits, on the other hand, table 2 shows some important differences in the ten most misidentified digits across tasks in terms of both extent and order of ambiguity.

Table 2. Most incorrectly identified irregulars and how they were most commonly misidentified in Manual, unsmoothed and smoothed conditions of Verbal.

Manual										
MNIST no.	#5655	#8377	#4741	#2131	#2655	#9730	#9506	#3423	#1261	#1902
										
Numeral	7	1	3	4	6	5	7	6	7	9
Response	2	6	5	9	1	6	2	0	1	4
Response%	95%	57%	82%	75%	54%	72%	66%	61%	51%	48%
Incorrect%	95%	87%	82%	77%	77%	72%	67%	62%	52%	52%
Verbal (Unsmoothed)										
MNIST no.	#948	#9730	#5655	#1261	#1394	#2463	#1902	#3423	#2655	#2598
										
Numeral	8	5	7	7	5	2	9	6	6	5
Response	9	6	2	1	6	0	4	0	1	3
Response%	99%	96%	92%	91%	27%	65%	76%	60%	53%	56%
Incorrect%	99%	98%	94%	91%	87%	87%	80%	63%	61%	59%
Verbal (Smoothed)										
MNIST no.	#2463	#9730	#948	#1261	#9665	#2655	#1394	#4177	#583	#1233
										
Numeral	2	5	8	7	2	6	5	2	8	9
Response	0	6	9	1	7	1	1	8	6	4
Response%	100%	100%	100%	100%	95%	91%	41%	45%	62%	82%
Incorrect%	100%	100%	100%	100%	100%	95%	95%	90%	86%	82%

Notes: Sorted by identifiability in ascending order. Empty responses are ignored. Data based on: 61 participants in Manual, 42 in unsmoothed and 11 in smoothed. *Top left*, in Manual task, digit #5655 represents a 7 but was mistaken for a 2 in 95% of responses. See Fig. A28 and tables A1-3 for further details.

Eight-digit no.948, for instance, was misidentified in almost all verbal responses but only 41% of Manual task responses. *Two*-digit no.2463 is another good example; misidentified in 87% of unsmoothed and 100% of smoothed verbal responses it is incorrectly identified in only 16% of Manual task responses. Furthermore, while the most common misidentification for no. 2463 is 0 in Verbal task (65% of responses), it is 8 in Manual task (11% of responses). While such discrepancies may be due to differences between MNIST and NIST databases, a more likely explanation lies in differences relating to display scale between Manual task and Verbal task as shown in Fig. 22.

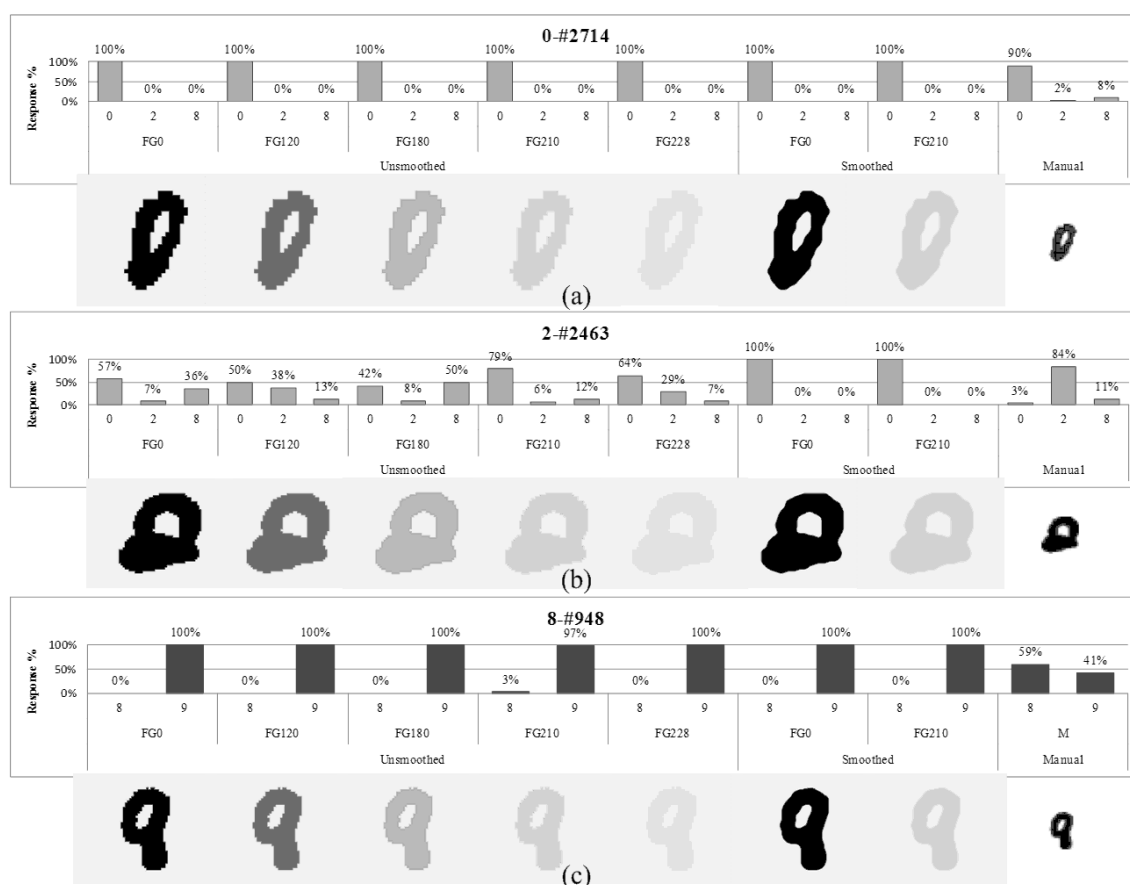
































Fig. 22. Discrepancies in identification for three irregular digits under Manual and Verbal: (a) A 0 which has perfect identification under unsmoothed and smoothed Verbal conditions alike but was misidentified in about 10% of responses under the Manual task; (b) a 2 is misidentified in 87% of unsmoothed and 100% of smoothed verbal responses but only 16% of Manual task responses; (c) an 8 was misidentified in almost all verbal responses but only 41% of Manual task responses.

In all three cases, the visual system seems more likely to perceive loop-like features at small scale than at large scale. A plausible justification for such bias relates to the way the visual system may be dealing with ambiguity at different scales: At small scale, the visual system is more likely to assume that features like loop holes are present but are simply hard to discern due to perceived lack of detail; hence, it is more likely to *compensate* for ostensibly missing features than at very large scale.

In the category of least identifiable digits by numeral, the commonalities across tasks are more pronounced as can be seen in table 3.

Table 3. Most incorrectly identified irregulars by numeral and how they were most commonly misidentified in Manual, unsmoothed and smoothed conditions of Verbal.

Numeral	0	1	2	3	4	5	6	7	8	9
Manual										
MNIST no.	#2714	#8377	#4177	#4741	#2131	#9730	#2655	#5655	#948	#1902
										
Response	8	6	8	5	9	6	1	2	9	4
Response%	8%	57%	15%	82%	75%	72%	54%	95%	41%	48%
Incorrect%	10%	87%	36%	82%	77%	72%	77%	95%	41%	52%
Verbal (Unsmoothed)										
MNIST no.	#1622	#8377	#2463	#4741	#2131	#9730	#3423	#5655	#948	#1902
										
Response	6	6	0	5	9	6	0	2	9	4
Response%	17%	33%	65%	22%	26%	96%	60%	92%	99%	76%
Incorrect%	17%	51%	87%	25%	27%	98%	63%	94%	99%	80%
Verbal (Smoothed)										
MNIST no.	#1622	#8377	#2463	#4741	#9793	#9730	#2655	#1261	#948	#1233
										
Response	6	6	0	5	9	6	1	1	9	4
Response%	23%	41%	100%	48%	45%	100%	91%	100%	100%	82%
Incorrect%	27%	45%	100%	57%	45%	100%	95%	100%	100%	82%

Notes: Sorted by numeral in ascending order. Data based on: 61 participants in Manual, 42 in unsmoothed and 11 in smoothed. Empty responses are ignored. *Top left*, in Manual task, digit #2714 is a 0 and was identified otherwise in 10% of responses and as an 8 in 8% of responses. See Fig. A29 for further details.































Nevertheless, there are some important discrepancies particularly in terms of the extent of digit ambiguity. However, error variations in digit nos.: 2714, 8377, 2463, 4741, 2131,

9730, 948 and 1902 (representing numerals: 0, 1, 2, 3, 4, 5, 8 and 9 respectively) can all be explained in terms of a decreased perception of loop-like features at large display scale.

Most Common Confusion Pairs

Another way to consider ambiguity is based on the most common error in all digits of a given numeral rather than in each digit on its own. We refer to ordered pairs of the form (*numeral, error*) as *confusion pairs* and present them in table 4 in descending order of error. For instance, in Manual task, irregular 3s were most often confused with numeral 5. Confusion pair (3-5) represents 41% of responses on irregular 3s all of which occurred on digit no.4741.

Table 4. Most common confusion in each numeral and the irregular digit that best represents that confusion in Manual, unsmoothed and smoothed conditions of Verbal.













Manual										
Numeral	3	1	7	5	9	4	6	2	8	0
Response	5	6	2	6	4	9	0	7	9	8
Confusion%	41%	35%	32%	14%	11%	10%	8%	7%	5%	4%
MNIST no.	#4741	#8377	#5655	#9730	#1902	#2131	#3423	#9665	#948	#2714
										
Digit%	100%	55%	49%	100%	73%	96%	95%	65%	100%	100%
Verbal (Unsmoothed)										
Numeral	7	9	5	1	2	8	3	0	6	4
Response	2	4	6	6	0	9	5	6	0	9
Confusion%	27%	25%	24%	17%	13%	13%	11%	8%	8%	4%
MNIST no.	#5655	#1902	#9730	#8377	#2463	#948	#4741	#1622	#3423	#2131
										
Digit%	58%	50%	79%	63%	100%	85%	100%	100%	93%	84%
Verbal (Smoothed)										
Numeral	9	2	5	3	7	1	6	8	0	4
Response	4	7	6	5	2	6	1	9	6	9
Confusion%	37%	25%	24%	23%	20%	18%	15%	12%	11%	6%
MNIST no.	#1233	#9665	#9730	#4741	#5655	#8377	#2655	#948	#1622	#9793
										
Digit%	37%	75%	85%	100%	54%	75%	77%	92%	100%	91%

Notes: Sorted by identifiability in ascending order. Empty responses are ignored. Data based on: 61 participants in Manual, 42 in unsmoothed and 11 in smoothed. *Top left*, in Manual task, irregular 3s were most often confused with 5; confusion 3-5 represents 41% of responses on irregular 3s all of which occurred on digit #4741. See Fig. A30 for further details.

Most Confusing Handwritten Digits

Digits that were misidentified-confused with the most numerals are presented in table 5. In this category, a number of error patterns are quite striking across tasks and merit a closer look: In all but two digits at least one response can only be considered plausible if the visual system identified the digit after flipping it upside-down or sideways (flip-identification). The following digits were clearly flip-identified: 2655(2, 7), 4177(5, 9), 3809(5), 1394(6, 7), 4498(4, 9) and 9680(2). Although the data in table 5 and tables A1-3 suggests that significant flip-identification occurred only on digits that are legitimately ambiguous, it does expose a flaw in experiment design under both tasks: Participants were not informed that all digits were going to be presented right side up.

Table 5. Irregulars that were confused with the most numerals in Manual, unsmoothed and smoothed conditions of Verbal.



















Manual															
MNIST no.	#2655				#4177				#3809			#1394			
															
Numeral	6				2				7			5			
Response	1	7	2	4	0	8	7	5	9	2	1	5	8	1	7
Response %	54%	10%	7%	3%	3%	15%	11%	5%	3%	33%	5%	3%	16%	5%	3%
Incorrect%	77%				34%				41%			25%			
Verbal (Unsmoothed)															
MNIST no.	#1394							#2655				#4498			
															
Numeral	5							6				8			
Response	6	1	()	2	0	8	7	1	()	0	7	7	9	4	()
Response %	23%	20%	17%	14%	7%	5%	4%	50%	6%	4%	2%	20%	17%	4%	4%
Incorrect%	89%							62%				44%			
Verbal (Smoothed)															
MNIST no.	#1394				#583			#4177			#9680			#5655	
															
Numeral	5				8			2			6			7	
Response	1	7	6	2	6	2	7	8	7	()	1	2	4	2	1
Response %	11%	6%	5%	2%	16%	4%	2%	11%	8%	2%	7%	2%	2%	17%	2%
Incorrect%	24%				22%			21%			12%			19%	

Notes: Sorted by most confusing. Responses occurring only once are ignored. Empty responses are considered and are denoted using (). Data based on: 61 participants in Manual, 42 in unsmoothed and 11 in smoothed. *Top left*, in Manual task, digit #2655 is confused with five numerals: 1, 7, 2, 4 and 0; the digit is misidentified in 77% of all responses and 1 represents 54% of responses on this digit. See Fig. A31 and tables A1-3 for further details.

Most Re-identified Handwritten Digits during Verbal Task

As discussed on p. 36 above, the double-prompt design of the Verbal task was based on the assumption that collected visual fixations not only reflect the most informative features of a digit but that they may also correlate with the numeral the visual system is considering and hence the verbal response about to be given. In order to evaluate this assumption, and to make better use of collected data, identifying ambiguous digits of a new type may be particularly helpful; hence, our last ambiguity category is of digits for which participants changed their verbal answer the most. Table 6 shows the most ambiguous digits based on their most common re-identification in descending order. For instance, digit no.4370 has been re-identified as a 4 and a 9 in 37% of trials, more frequently than any re-identification in any digit under unsmoothed viewing. In the next section, the author uses these digits to focus the analysis of visual fixation data on a smaller set of trials, verbal responses and corresponding fixations.

Table 6. Irregulars that were re-identified the most in unsmoothed and smoothed conditions of Verbal based on their most common re-identification pair.

Verbal (Unsmoothed)									
MNIST no.	#4370	#2131	#1233	#2036	#2598	#9506	#3809	#9665	#1622
									
Numeral	9	4	9	5	5	7	7	2	0
Re-ID Pair	4, 9	4, 9	4, 9	3, 5	3, 5	7, 2	2, 7	2, 7	0, 6
Re-ID Pair%	37%	29%	29%	26%	24%	22%	21%	19%	19%
Verbal (Smoothed)									
MNIST no.	#4370	#1902	#2036	#1622	#2598	#5655	#9793	#3423	#9506
									
Numeral	9	9	5	0	5	7	4	6	7
Re-ID Pair	4, 9	4, 9	3, 5	0, 6	3, 5	2, 7	4, 9	0, 6	2, 7
Re-ID Pair%	55%	45%	45%	45%	45%	45%	36%	27%	27%

Notes: Sorted by most re-identified. Trials with an empty response are ignored. Data based on: 42 participants in unsmoothed and 11 in smoothed. *Top left*, in unsmoothed conditions of Verbal, digit #4370 is a 9 and is identified as 4 and 9 in 37% of trials (regardless of order). See Fig. A32 for further details.

Concluding Remarks

1. Despite significant differences between Manual task and Verbal task, our results also suggest some striking similarities particularly with unsmoothed conditions in terms of: (1) identification rates of regular digits, (2) most identifiable digits and (3) most misidentified digits by numeral. Based on these observations, the author concludes that the identification task remains fundamentally the same irrespective of eye-tracking considerations like double-prompt identification and large-scale digits.

2. Comparative analysis of identification performance suggests striking similarities across different contrast conditions of the Verbal task when it comes to identification rate of regular and irregular digits alike with the most significant differences emerging in response time for irregular digits under the lowest contrast condition. This may be due to a decrease in the distracting effect of spatial aliasing with decreasing luminance contrast but it may also be due to individual variations among participants.

3. Although lower contrast conditions do show a slight increase in identification rate over higher contrast conditions, fluctuations in both identification rate and response time prevent the author from making recommendations as to Optimal Viewing guidelines for future research. A more rigorous experiment using within-subject design and specifically tailored to studying the effect of luminance contrast on identification response in similar tasks is recommended to that end.

4. Comparative analysis of identification performance shows significant difference in identification rate between smoothed and unsmoothed conditions suggesting that a more nuanced smoothing scheme should be explored in future research.

5. Comparative analysis of the most misidentified digits in Manual and Verbal tasks, suggests that the visual system is more likely to perceive loop-like features at small scale than at very large scale.

6. Analysis of the most confusing digits suggests that significant flip-identification occurred on a handful of irregular digits. This may be avoided in future studies by advising participants that digits (or characters) will be presented right side up.

Analysis of Visual Response Data

Gaze Span across Viewing Conditions of Verbal Task

Although the heat map bounding box statistics calculated by the heat-mapping script (see pp. 72-75 above) are a crude approximation of gaze span, the 10% low activity cut-off imposed during heat map creation adds to the robustness of this measure by eliminating outliers (i.e. regions receiving low gaze frequency or short gaze duration). Given the absence of a general-purpose technique to quantify the differences among arbitrary sets of visual fixations, the author relied on heat map bounding box areas (Fig. 23) in order to evaluate gaze span variation across viewing conditions and digit types.

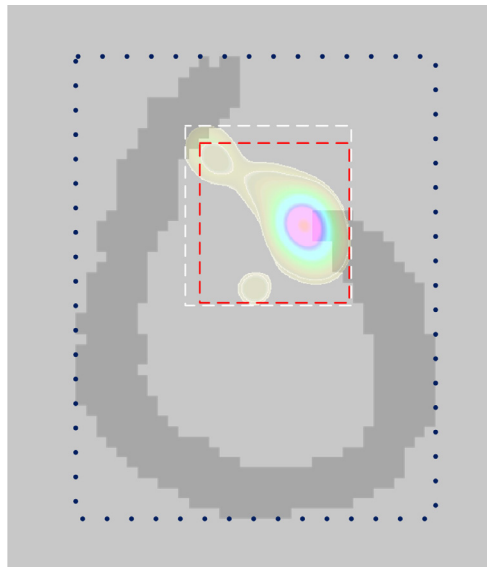


Fig. 23. Digit and heat map bounding boxes. The dark blue, white and red squares represent the bounding boxes of the digit, its count-based heat map and duration-based heat map respectively. Data based on 14 fixations in one trial in U-FG210.

To that end, count- and duration- based heat maps were generated for each of the 3883 trials of the Verbal task individually using the batch-mode heat-mapping scheme described on p. 73 above. The count- and duration-based bounding box areas returned by the heat-mapping script are used along with the trial digit's bounding box area to determine the gaze span:

$$\text{Gaze Spatial Span} = \frac{\text{Heatmap Boundingbox Area}}{\text{Digit Boundingbox Area}}$$

Trial gaze spatial span are averaged for regular and irregular digits in each viewing condition as shown in Fig. 24.

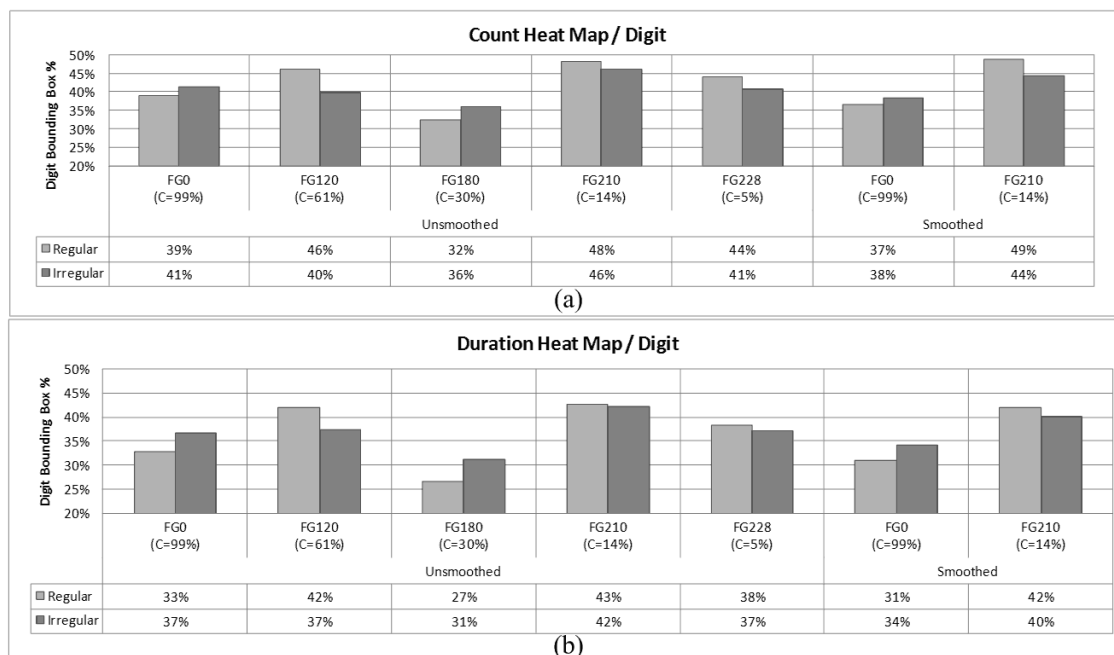


Fig. 24. Gaze spatial span in regular and irregular digits across viewing conditions of Verbal. For each trial, count-based and duration-based heat maps are generated and their bounding box areas are calculated. Gaze spatial span is defined as the ratio of heat map bounding box area to digit bounding box area averaged over a given set of trials. Here, gaze spatial span is shown for regular and irregular digits across viewing condition using (a) count-based and (b) duration-based heat maps. Horizontal labels show condition name and luminance contrast. Data based on: 42 participants in unsmoothed and 11 in smoothed.

Although the average of lower contrast conditions under unsmoothed viewing (FG210 and FG228) and smoothed viewing (FG210) is greater than in higher contrast conditions, the results show no clear relationship between luminance contrast and gaze span. Moreover, a quick visual examination of generated maps shows sweeping variations in terms of both gaze pattern and span among participants within and across viewing conditions (see Fig. A24). A more striking observation is the lack of a clear relationship between digit type (i.e. regular-irregular) and gaze span. Since an important number of irregular digits have been shown to be quite challenging to participants, one would expect the visual system to examine a larger set of features before it can arrive at a conclusive identification.

Before we can address this discrepancy, a closer look at the two heat map types is in order. As results in Fig. 24 show, on average, count-based maps span a greater area than duration-based maps. Upon visual examination of generated trial heat maps one can quickly surmise that the two trial maps mostly overlap with duration-based generally spanning a sub-region of count-based heat maps (see Fig. A23). To put it differently, regions that receive enough gaze duration to be colour-coded in the duration-based map are almost always fixated frequently enough to make it into the count-based map too. The inverse, however, is not as likely: A few regions that are fixated frequently enough to make it into the count-based map are not fixated long enough to make it into the duration-based map. Since it is plausible to assume that more challenging or informative digit features fall into the former category, we may be able to approximate their percentage simply by calculating the ratio of duration-based gaze span to count-based gaze span. Results in Fig. 25 seem to support this notion: Even though the portion of an irregular digit that receives significant gaze may not be bigger than that in regular digits, more of it is challenging-informative than in the latter. Although the variations are quite small, (3-

6%) they are perhaps so due to the relatively low 10% activity cut-off used during heat map generation. The author leaves the manipulation of this parameter as a suggestion for further inquiry into determination of regions of interest.

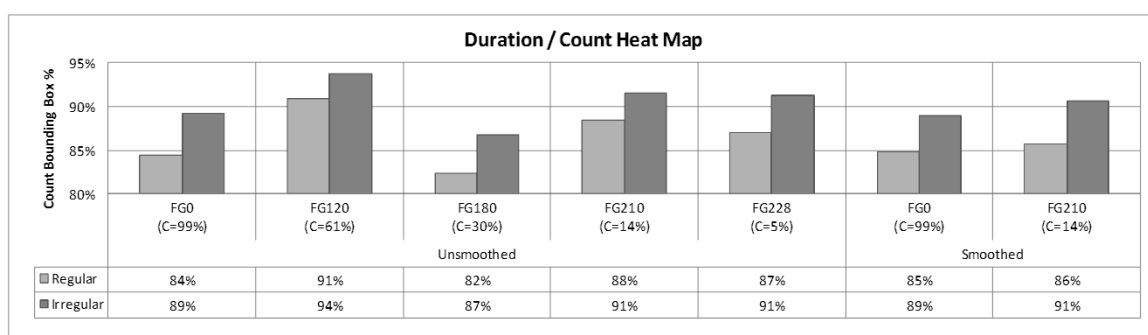


Fig. 25. Ratio of duration-based to count-based gaze span based on data from Fig. 24.

Visual Fixation Selection Criteria

Viewing Conditions Selection Criteria

After a comparative analysis between the seven viewing conditions of the Verbal task based on a number of metrics in verbal identification and visual response data, the author finds no evidence suggesting that change in luminance contrast affects task performance fundamentally and most differences can be attributed to individual variations. Comparison between smoothed and unsmoothed conditions, on the other hand, shows important differences when it comes to correct identification rate particularly in irregular digits (see p.76 above). As noted earlier, these differences are probably due to distortions during smoothing. Hence, while the author sees no basis for restricting analysis of gaze to individual contrast conditions, we opt to combine visual fixation data collected under smoothing conditions separately. An added advantage to this approach is that it allows us to evaluate the effect of smoothing distortions on gaze

activity in the same digit.

Selection Criteria for Trials and Periods of Interest

As we discussed before, the double-prompt design of the Verbal task was based on the assumption that collected visual fixations not only reflect the most informative features of a digit but that they may also correlate with the numeral the visual system is considering and hence the verbal response about to be given. Although this assumption remains to be demonstrated, a crucial advantage of comparing visual fixations within trial is the possibility of controlling for the big gaze variations observed across participants (see Fig. A24). Moreover, when the selection is made in trials where digits were re-identified, a comparison of resulting heat maps may expose not only informative features correlating with each answer but also features that are likely related to the change in answer.

In order to evaluate the above statements, the author selected and compared visual fixations based on the following selection template:

1. Select a commonly re-identified digit and identify its most common re-identification pair from table 6.
2. Select all trials on this digit in all contrast conditions under unsmoothed or smoothed viewing where the verbal responses correspond to the re-identification pair.
3. Select all fixations that started within one second prior to the onset of the verbal response that matches one of the answers in the re-identification pair.
4. Select another set of fixations that started within one second prior to the onset of the verbal response that matches the other answer in the re-identification pair.
5. Generate duration-based heat maps for the two sets of fixations.
6. Compare and contrast the two heat maps to identify features of interest.

Steps 3 and 4 in the above template are made possible thanks to selection criteria Fixation Time Label and Fixation Lapse introduced on p. 64 above. For instance, to carry out the following selection:

Select all right-eye fixations in all trials under unsmoothed viewing where digit no.4370 was identified as 4 then 9 or 9 then 4 and where the fixation started within the 1000-ms period preceding onset of response 4.

We use the Advanced Filter criteria outlined in Fig. 26.

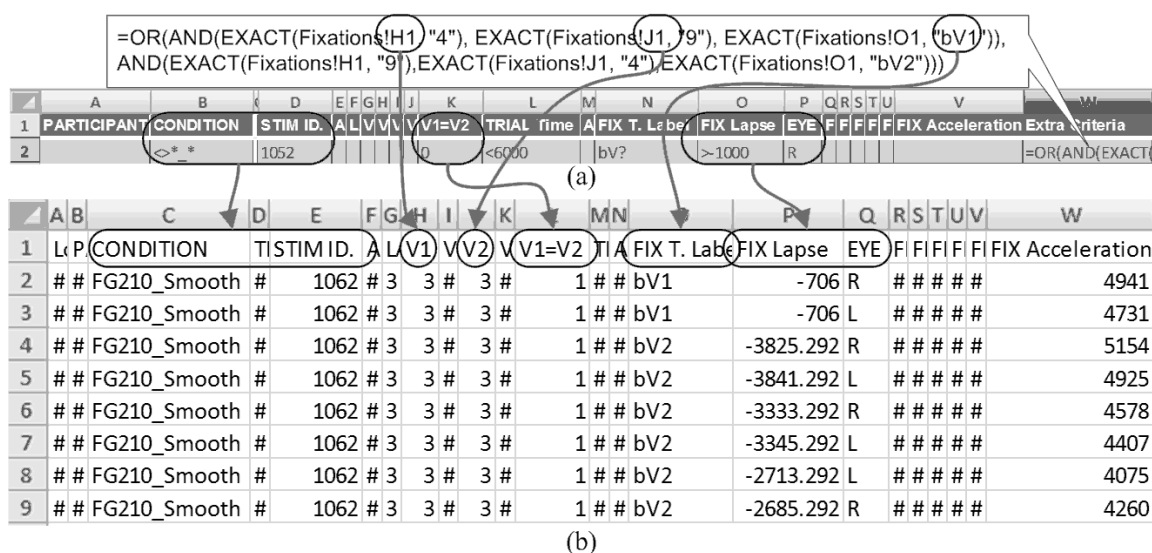


Fig. 26. Fixation selection in re-identified digits: (a) A set of Advanced Filter criteria are used to select all right-eye fixations in all trials under unsmoothed viewing where digit ID 1052 (MNIST no.4370) was identified as 4 then 9 or 9 then 4 from fixation table in (b) where fixation started within the 1000-ms period preceding onset of first or second response 4.

Gaze during Re-Identification

We proceed with a comparison of visual fixations in the most re-identified digits of the Verbal task using the above scheme. To facilitate discussion flow, we keep the size of heat map illustrations in this section to a minimum and leave the full-page versions to the appendix (see Fig. A33-39). To simplify the comparison, we present the duration-based heat maps only.

Gaze during Zero-Six Re-Identification

Fig. 27 presents a comparison between gaze preceding verbal response in trials where a 0 (digit no.1622) was identified as 0 and 6 under unsmoothed and smoothed viewing. In both cases, the heat maps convey that much of the gaze time is spent on repeated attempts probing for *missing* features along a loop-completing path more consistent with the upcoming verbal response.

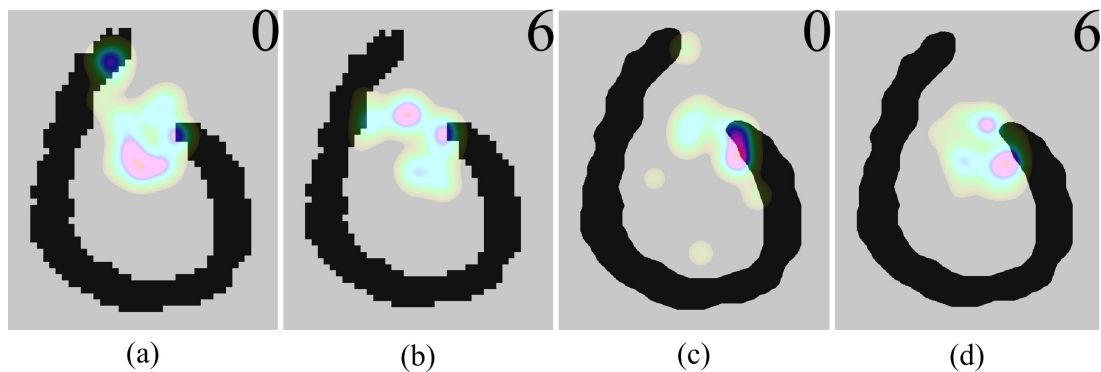


Fig. 27. Four duration-based heat maps showing gaze density during 1-second period preceding verbal identification of 0-digit no.1622 as 0 and 6. Heat maps were generated based on: (a) 17 fixations in 6 trials, (b) 21 fixations in 8 trials, (c) 16 fixations in 5 trials, and (d) 16 fixations in 5 trials (see Fig. A33 for a full-page version).

Gaze during Two-Seven Re-Identification

Fig. 28 presents a comparison between gaze preceding verbal response in trials where a 7 (digit no.9506) was identified as 2 and 7 under unsmoothed and smoothed viewing. Here, the results show more subtle differences than in Fig. 27 and merit closer attention to the shape and placement of colour-coded regions particularly near potential curvatures, which distinguish 2 from 7. Indeed, both unsmoothed and smoothed 2-maps (Fig. 28a,c) show significantly more gaze at the inside of the top corner and less overlap with digit strokes than their counterpart 7-maps (Fig. 28b,d). In contrast, the highest gaze patches in the 7-maps overlap and parallel the vertical stroke. This seems plausible and

suggests that in order to confirm curvatures – here consistent with a 2 – examining the edges of a stroke may be more conclusive than examining the stroke itself which may be better suited to confirming stroke continuation or girth variation – here consistent with a 7. This also explains the placement of the high-gaze patch inside the bottom corner in the smoothed 2-map (Fig. 28c).

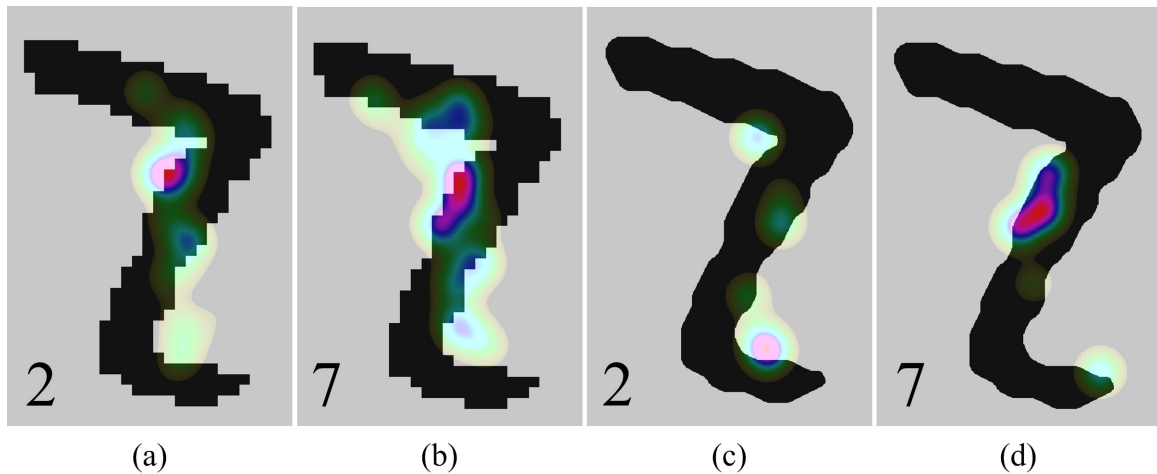


Fig. 28. Four duration-based heat maps showing gaze density during 1-second period preceding verbal identification of 7-digit no. 9506 as 2 and 7. Heat maps were generated based on: (a) 25 fixations in 10 trials, (b) 26 fixations in 10 trials, (c) 6 fixations in 3 trials, and (d) 7 fixations in 3 trials (see Fig. A34 for a full-page version).

Fig. 29 presents a comparison between gaze preceding verbal response in trials where a 2 (digit no.9665) and a 7 (digit no.3809) were each identified as 2 and 7 under unsmoothed conditions. Here, a similar pattern to that observed in Fig. 28 is evident: The 2-maps (Fig. 29a,c) show much higher gaze at the contours of the top curve-like feature than the 7-maps (Fig. 29b,d). Moreover, the first 2-map (Fig. 29a) shows a high-gaze patch right at the inside of the bottom loop-like feature – consistent with a 2 – On the other hand, the first 7-map (Fig. 29b) shows central gaze activity overlapping and paralleling a potential horizontal stroke consistent with a 7. In both 7-maps (Fig. 29b,d) a

high-gaze patch seems to get very close to the loop-like feature at the bottom of the digit without actually reaching it.

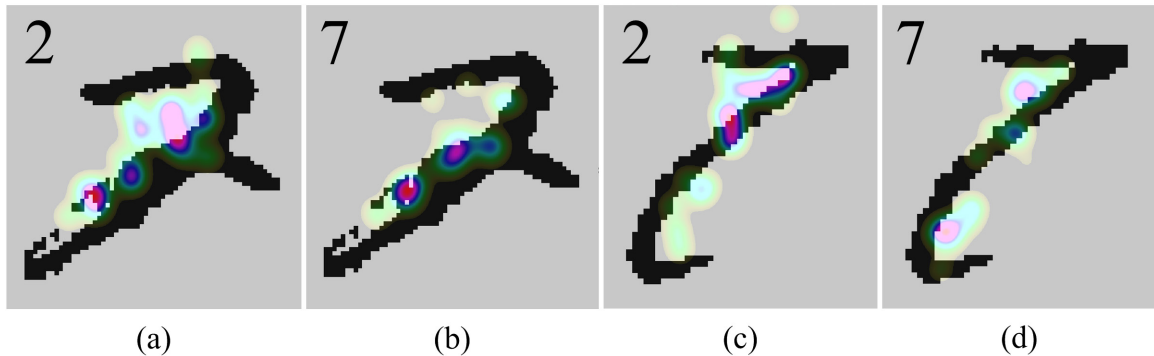


Fig. 29. Four duration –based heat maps showing gaze density during 1-second period preceding verbal identification of 2-digit no.9665 (*left*) and 7-digit no.3809 (*right*) as 2 and 7. Heat maps were generated based on: (a) 26 fixations in 8 trials, (b) 23 fixations in 8 trials, (c) 22 fixations in 9 trials, and (d) 23 fixations in 9 trials (see Fig. A35 for a full-page version).

Gaze during Three-Five Re-Identification

Fig. 30 presents a comparison between gaze preceding verbal response in trials where a 5 (digit no. 2036) was identified as 3 and 5 under unsmoothed and smoothed viewing. The 3-maps (Fig. 30a,c) show much higher gaze inside the top left corner – more consistent with a 3 – than counterpart 5-maps (Fig. 30b,d). The 5-maps, on the other hand, show more gaze inside the top right corner – more consistent with a 5. However, gaze density at the top left corner in 3-maps is a lot higher than at the top right corner in 5-maps. This is quite plausible subjectively since the digit as a whole does look more like a 5 than a 3. To confirm this we look at the identification rate of this digit in tables A2 and A3. Under unsmoothed conditions, this digit is identified as a 5 in 63% of responses, and as a 3 in 35% while under smoothed conditions, it is identified as a 5 in 68% of responses and a 3 in 32%. This suggests that the amount of gaze needed to examine features may relate not only to their informative value but also to the difficulty

in reconciling them, or the digit as a whole, with a given numeral.

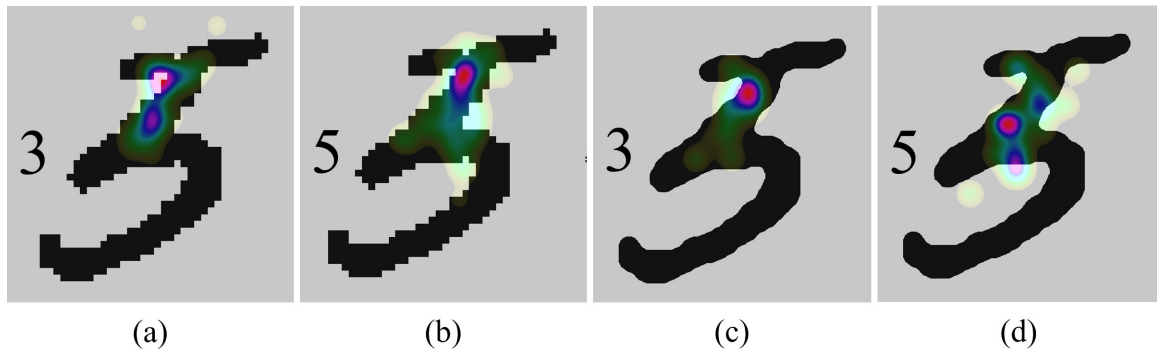


Fig. 30. Four duration-based heat maps showing gaze density during 1-second period preceding verbal identification of 5-digit no. 2036 as 3 and 5. Heat maps were generated based on: (a) 27 fixations in 11 trials, (b) 32 fixations in 11 trials, (c) 14 fixations in 5 trials, and (d) 15 fixations in 5 trials (see Fig. A36 for a full-page version).

Fig. 31 shows a comparison between gaze preceding verbal response in trials where a 5 (digit no. 2598) was identified as 3 and 5 under unsmoothed and smoothed viewing. Under unsmoothed viewing (Fig. 31a,b), heat maps show little difference in terms of which features received gaze; instead, the 5-map (Fig. 31b) shows more gaze at the inside of the central corner than the 3-map (Fig. 31a) conveying more difficulty in reconciling this feature with a 5 than a 3. Indeed, this is also consistent with the overall identification rate for this digit in unsmoothed conditions where identification as 3 is more common than 5 (55% and 40%). Under smoothed viewing (Fig. 31c,d), heat maps show one crucial difference in gaze; namely, while the high-gaze patch remains between the top horizontal stroke and the slanted stroke in the 5-map (Fig. 31d), it decisively overlaps the top stroke in the 3-map (Fig. 31c). This conveys repeated attempts to confirm that the top stroke stretches to meet the top of the slanted stroke in a way consistent with a 3.

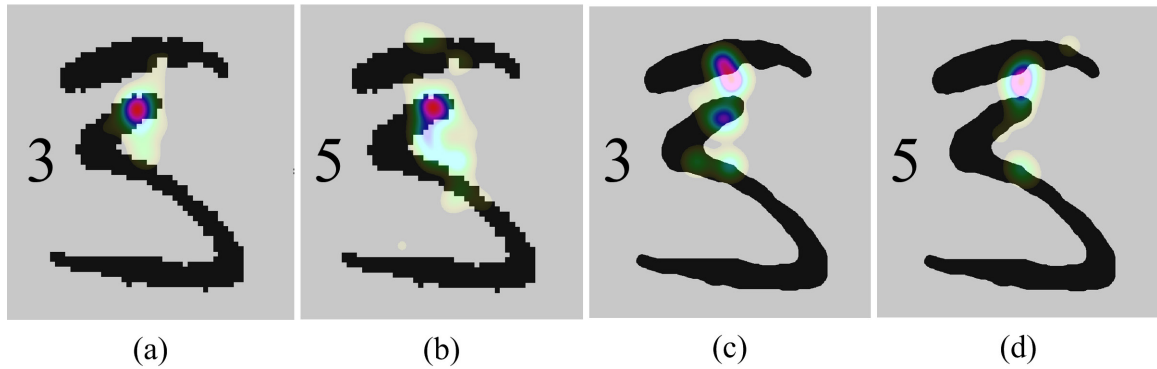


Fig. 31. Four duration-based heat maps showing gaze density during 1-second period preceding verbal identification of 5-digit no. 2598 as 3 and 5. Heat maps were generated based on: (a) 26 fixations in 10 trials, (b) 25 fixations in 10 trials, (c) 11 fixations in 5 trials, and (d) 13 fixations in 5 trials (see Fig. A37 for a full-page version).

Gaze during Four-Nine Re-Identification

Fig. 32 shows a comparison between gaze preceding verbal response in trials where two 9s (digit nos.1233 and 1902), and a 4 (digit no.2131) were each identified as 4 and 9 under either unsmoothed or smoothed viewing.

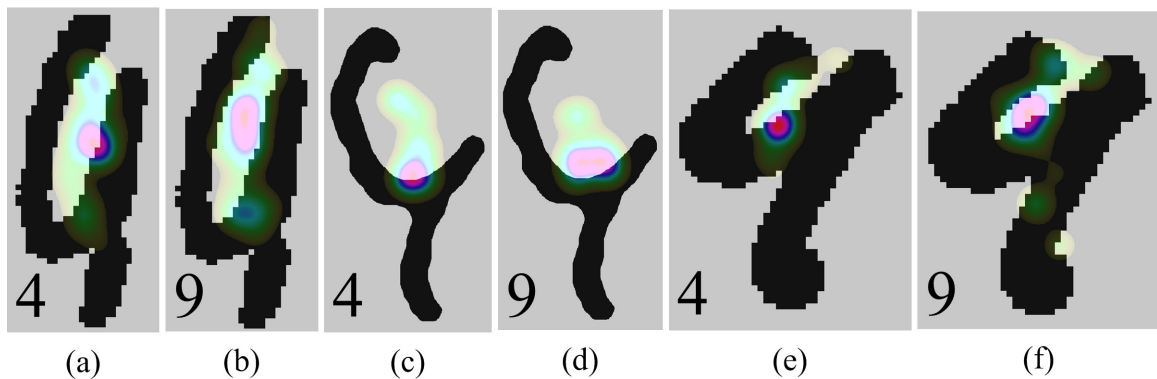


Fig. 32. Six duration-based heat maps showing gaze density during 1-second period preceding verbal identification of 9-digit no.1233 (*left*), 9-digit no.1902 (*centre*) and 4-digit no.2131 (*right*) as 4 and 9. Heat maps were generated based on: (a) 26 fixations in 12 trials, (b) 27 fixations in 12 trials, (c) 10 fixations in 5 trials, (d) 9 fixations in 5 trials, (e) 26 fixations in 12 trials, and (f) 36 fixations in 12 trials (see Fig. A38 for a full-page version).

A striking commonality emerges in shape and placement of the high-gaze patch in all

three digits; namely, while the patch is located at roughly the same place in each digit, it is both more elongated and has less overlap with the target stroke in 9-maps (Fig. 32b,d,f) than in 4-maps (Fig. 32a,c,e). As observed before, this conveys repeated confirmations of curvature. Moreover, the fact that the patch is located in the same place in each digit suggests that the most decisive aspect during identification is whether the *same* feature happens to be more of a corner than a curve, consistent with a 4, or the inverse, consistent with a 9. Another pattern of interest emerges in 9-maps in Fig. 32b,f, where more gaze is present at the top of the digit than in counterpart 4-maps (Fig. 32a,e). This seems to convey repeated probing for *missing* loop features consistent with a 9.

Fig. 33 shows a comparison between gaze preceding verbal response in trials where a 9 (digit no.4370) was identified as 4 and 9 under unsmoothed and smoothed viewing.

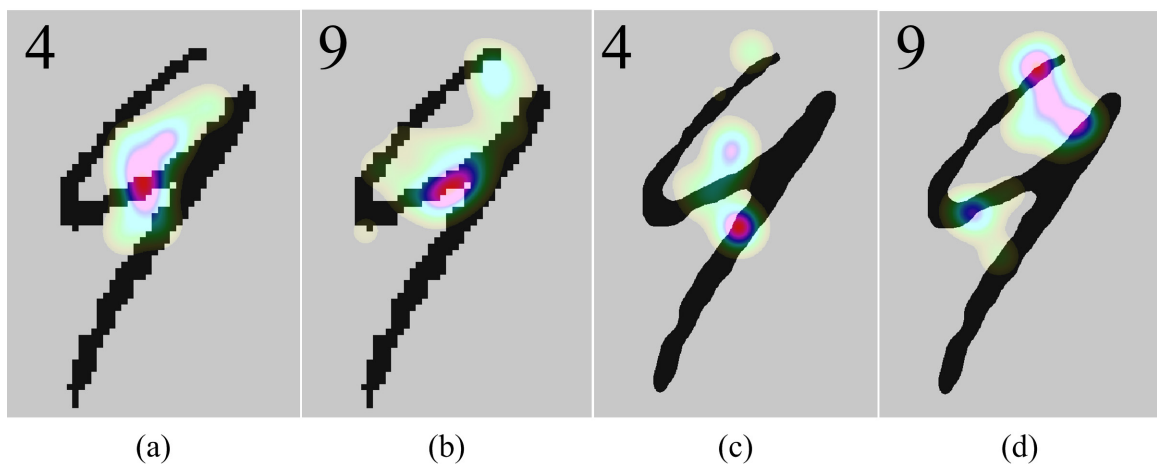


Fig. 33. Four duration-based heat maps showing gaze density during 1-second period preceding verbal identification of 9-digit no.4370 as 4 and 9. Heat maps were generated based on: (a) 40 fixations in 15 trials, (b) 41 fixations in 15 trials, (c) 16 fixations in 6 trials, and (d) 14 fixations in 6 trials (see Fig. A39 for a full-page version).

Here, 9-maps (Fig. 33b,d) show a gaze pattern that is very consistent with our last observation; namely, in both heat maps, gaze density is very high at the top of the digit compared to counterpart 4-maps (Fig. 33a,c). This conveys a number of attempts to confirm a closed loop consistent with a 9. Under unsmoothed viewing (Fig. 33a,b), the high-gaze patches are located at roughly the same place reminiscent of Fig. 32. Here, however, the 9-patch overlaps and parallels the target stroke conveying a number of attempts to confirm that it is straight rather than curved or angled – a straight line here is more consistent with a 9 than a 4. The unsmoothed 4-patch, on the other hand, crosses the target stroke conveying repeated attempts to confirm a point feature such as a corner, hence consistent with a 4. The lack of analogous gaze pattern under smoothed viewing (Fig. 33c,d) suggests that the smoothing process may have affected the way the digit was perceived. Indeed, the most obvious distortion is the disappearance of the small horizontal tip at the top end of the digit making identification as 9 less plausible. This is consistent with overall correct identification of this digit in smoothed (36%) compared to unsmoothed (70%) conditions. It also accounts for the higher gaze activity at the top of the smoothed 9-map (Fig. 33d) compared to its unsmoothed counterpart (Fig. 33b): Reconciling the top of the digit with a 9-loop is more challenging in the smoothed 9 hence requiring a bigger share of gaze.

Concluding Remarks

In summary, analysis of visual fixations in the most re-identified digits of the Verbal task provides evidence suggesting the following:

1. Gaze duration density strongly correlates with informative features especially when feature characteristics are difficult to determine.
2. When identification is consistent with the presence of a loop, repeated attempts and long gaze can be observed probing for features that complete the loop in a way that is

consistent with the verbal identification (Fig. 27, Fig. 32b,f, Fig. 33b,d).

3. When identification is consistent with the presence of a curvature, repeated attempts and long gaze can be observed along the edge of the target stroke rather than the stroke itself (Fig. 28a,c, Fig. 29c, Fig. 32b,d,f). Moreover, when the curve resembles a corner, the gaze becomes more concentrated. (Fig. 28a,c, Fig. 29c).

4. When identification is consistent with the presence of a corner, high gaze density can be observed only when the digit is difficult to reconcile with the verbal answer. Moreover, gaze is more likely to overlap the target stroke than in the case of a curve (Fig. 30a,c, Fig. 32a,c,e, Fig. 33a).

5. When identification is consistent with the presence of a straight stroke or variations in the girth of the stroke, repeated attempts and long gaze can be observed along the stroke itself (Fig. 28b,d, Fig. 29b, Fig. 33b).

In conclusion, much of the gaze activity preceding verbal response in re-identified digits seems to target features *consistent* with the verbal response about to be given. Hence, the strategy appears to be predominantly of *confirmation* rather than *exploration*. Although an examination of visual fixations during the entire trial and in other digits is in order before further claims can be made, this suggests that the visual system hypothesizes the identity of the digit prior to fixating particular features, which implies that extra-foveal vision plays a crucial role in guiding the observed visual response. This preliminary observation is consistent with a number of studies on eye movement during image recognition presented in the literature review (Maw and Pomplun 2004; Brandt and Stark 1997; Chernyak and Stark 2001; Stark and Choi 1996; Privitera and Stark 2000; Noton and Stark 1971). As far as applicability to Pattern Recognition is concerned, preliminary results are consistent with a number of premises upon which our approach rests; namely, that visual fixations not only correlate with the informative-disambiguating

value of features fixated but that they also reflect the numeral the visual system is hypothesizing.

Comparative analysis of gaze spatial span across Verbal task viewing conditions yielded the following observations:

1. Although the average of lower contrast conditions under unsmoothed viewing and smoothed viewing is greater than in higher contrast conditions, the results show no clear relationship between luminance contrast and gaze span. This is perhaps largely due to sweeping variations both in terms of gaze pattern and span among participants within and across viewing conditions.

2. The average portion of an irregular digit that receives significant gaze may or may not be bigger than in the case of a regular digit depending on viewing condition and individual variations. However, irregular digits yield a higher duration-based to count-based gaze span ratio than regular digits. This is consistent with the intuitive notion that irregular digits have a higher portion of challenging-informative features than regular digits.

CHAPTER 5

CONCLUSIONS AND FURTHER INQUIRY

Summary and Preliminary Results

In this thesis, the author evaluated a novel approach that explores the use of human visual fixations and identification data in order to identify features of interest for Pattern Recognition applications. We selected handwritten digit recognition as a prototype application and used seventy-four digit images from the NIST database as stimuli. Fifty-four of these digits were reported to be particularly problematic for a variety of classifiers in the literature. The other twenty look very prototypical and were used as a reference. In the data collection stage, sixty participants were asked to verbally identify each handwritten digit twice at very large scale under one of seven different contrast and smoothing conditions (Verbal task). Both verbal responses and visual fixations were recorded during the course of the identification task for further analysis. Participants were then asked to identify the same digits manually under normal viewing for reference purposes (Manual task). Below, we summarize the outcome of this study.

Development of Software Tools

The author developed the following software tools to select, modify and present handwritten digit stimuli for use during data collection:

1. A script to match MNIST handwritten digit images into the NIST database using MATLAB Image Processing Toolbox.
2. An image scaling and smoothing script to reduce spatial aliasing in large-scale

NIST handwritten digits using MATLAB Image Processing Toolbox.

3. The two handwritten digit identification tasks using SR Research Experiment Builder.

The author developed the following suit of software tools to facilitate analysis of verbal, manual and visual response data:

1. Verbal response isolation and labelling tool combining amplitude, zero-crossing rate and ESACF heuristics using MATLAB and VOICEBOX.

2. A modified version of SR Research's C-based EDF2ASC to import visual fixations and trial details from all EDF files into MS Excel format.

3. A suit of MS Excel VBA macros to consolidate manual, verbal and visual response data into a single MS Excel workbook for convenient access.

4. Two MS Excel VBA macros to select and visualize visual fixations.

5. A MATLAB script using Image Processing Toolbox Gaussian convolution to create count- and duration- based fixation heat maps from an arbitrary set of visual fixations.

6. An MS Excel Add-on to embed the MATLAB heat-mapping functionality into an MS Excel macro-enabled workbook.

Results in Identification Performance

In addition to the identification rate data and digit ambiguity ranking (see tables A1-3 and Fig. A27-32), the analysis of handwritten digit identification performance yielded the following preliminary results:

1. Despite significant differences between identifying handwritten digits under unconstrained identification in normal viewing (Manual task) and double-prompt identification in large-scale viewing (Verbal task), results show some crucial similarities suggesting that the identification task remained fundamentally the same.

2. Comparative analysis of identification performance suggests striking similarities across contrast conditions of the Verbal task when it comes to correct identification rate with the most significant differences emerging in identification response time for irregular digits under the lowest contrast condition (Michelson contrast =5%). This may be due to a decrease in the distracting effect of spatial aliasing with decreasing luminance contrast.

3. Although lower contrast conditions of the Verbal task do show a slight increase in identification rate over higher contrast conditions, fluctuations in both identification rate and response time prevent the author from making recommendations as to Optimal Viewing guidelines for future research.

4. Comparative analysis of identification performance in Verbal task shows significant differences in correct identification rate between smoothed and unsmoothed conditions suggesting that a more nuanced smoothing scheme should be explored in future research.

5. Comparative analysis of the most misidentified digits in Manual and Verbal tasks, suggests that the visual system is more likely to perceive loop-like features at small scale than at very large scale.

6. Analysis of the most confusing digits suggests that significant flip-identification occurred on a handful of irregular digits. This may be avoided in future studies by advising participants that handwritten digits (or characters) will be presented right side up.

Results in Gaze Spatial Span

Comparative analysis of gaze spatial span across Verbal task viewing conditions yielded the following observations:

1. Although the average of lower contrast conditions under unsmoothed viewing

and smoothed viewing is greater than counterparts in higher contrast conditions, the results show no clear relationship between luminance contrast and gaze span. This is perhaps largely due to great variations in both gaze pattern and gaze span among participants within and across viewing conditions.

2. The average portion of an irregular digit that receives significant gaze may or may not be bigger than in the case of a regular digit depending on viewing condition and individual variations. However, irregular digits yield a higher duration-based to count-based gaze span ratio than regular digits. This is consistent with the intuitive notion that irregular digits have a higher portion of challenging-informative features than regular digits.

Results in Gaze during Re-Identification

In the latter part of the analysis stage, a smaller set of ambiguous digits were identified based on how often participants changed their minds about the numeral they represent. We referred to these as the most *re-identified* digits. For each of these digits, visual fixations preceding a given response were combined into a single fixation heat map (see Fig. A33-39). Analysis of the resulting heat maps yielded the following preliminary results:

1. Gaze duration density strongly correlates with informative features especially when feature characteristics are difficult to determine.

2. When identification is consistent with the presence of a loop, repeated attempts and long gaze can be observed probing for features that complete the loop in a way that is consistent with the verbal response.

3. When identification is consistent with the presence of a curvature, repeated attempts and long gaze can be observed along the edge of the target stroke rather than the stroke itself. Moreover, when the curve resembles a corner, the gaze becomes more

concentrated.

4. When identification is consistent with the presence of a corner, high gaze density can be observed only when the digit is difficult to reconcile with the numeral. Moreover, gaze is more likely to overlap the target stroke than in the case of a curve.

5. When identification is consistent with the presence of a straight stroke or variations in the girth of the stroke, repeated attempts and long gaze can be observed along the stroke itself.

In conclusion, most gaze activity preceding verbal response in re-identified digits seems to target features *consistent* with the verbal response about to be given. Hence, the strategy appears to be predominantly of *confirmation* rather than *exploration*. This suggests that the visual system hypothesizes the identity of the digit prior to fixating particular features. This is consistent with research findings on eye movement during image recognition from Noton and Stark (1971) to Maw and Pomplun (2004).

As far as applicability to Pattern Recognition is concerned, results are consistent with the principle premise upon which our approach rests; namely, that visual fixations not only correlate with the informative-disambiguating value of features fixated but that they also reflect the numeral the visual system is hypothesizing. Therefore, data collected using this approach represents a promising and hitherto untapped potential in the training of handwriting classifiers.

Suggestions for Further Inquiry

In this thesis, the author analysed a small fraction of the visual response data recorded during data collection with much of the analysis left for future research. Below, we make a number of suggestions for such inquiries:

1. During analysis of gaze spatial span, the author used the ratio of the duration-based heat-map bounding box to the count-based heat-map bounding box as a measure of

the challenging-informative portion of handwritten digit regions that received gaze. A possible inquiry aiming to quantify digit ambiguity may explore manipulating the low activity cut-off threshold used during heat map generation to produce better approximation of the challenging portion of a digit and evaluate the accuracy of this measure by correlating it to digit identification rate. The same approach could be used to determine digit regions and features of higher interest.

2. During analysis of gaze spatial span across contrast conditions, the author used full-trial fixations and concluded that luminance contrast has no clear effect on gaze span. An alternative approach may be to use fixations that started during the *bVI* trial period (i.e. preceding the first verbal response) instead.

3. In a study on eye movement during recognition of actor faces, Maw and Pomplun (2004) proposed that the moment of recognition may be indicated by a participant's dilated pupils. A possible inquiry could use the pupil size data in our EDF files along with verbal response time data to evaluate this observation in the context of handwritten digit identification.

4. A possible inquiry into the use of eye movement to improve performance in handwritten digit recognition applications may use our visual fixation database to train a selective attention recognition scheme similar to that proposed by Salah, Alpaydin, and Akarun (2001) to train handwritten digit classifiers using a combined top-down-bottom-up scheme similar to the one used in Peters and Itti (2007).

REFERENCES

- EyeLink data viewer* 2010. Vol. 1.10.1SR Research Ltd. <http://www.sr-research.com/> (accessed 4/16/2011).
- Experiment builder* 2009. Vol. 1.5.58SR Research Ltd. <http://www.sr-research.com/> (accessed 4/16/2011).
- EyeLink II head-mounted user manual* 2009. Vol. 2.14SR Research Ltd. <http://www.sr-research.com/> (accessed 4/16/2011).
- MATLAB* 2009. Vol. 7.9.0.529 (R2009b). Natick, Massachusetts: The MathWorks Inc.
- Barriere, C. and R. Plamondon. 1998. Human identification of letters in mixed-script handwriting: An upper bound on recognition rates. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 28, no. 1: 78-81.
- Brandt, Stephan A. and Lawrence W. Stark. 1997. Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience* 9, no. 1: 27-38.
- Brookes, D. M. 2010. *VOICEBOX: A speech processing toolbox for MATLAB*. <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>. Vol. 2010.
- Brubeck, M., J. Haberman, and D. Mazzone. 2010. *Audacity: Free audio editor and recorder*. Vol. 1.3 Beta. SourceForge.Net. <http://audacity.sourceforge.net/>.
- Brysbaert, Marc. 1995. Arabic number reading: On the nature of the numerical scale and the origin of phonological recoding. *Journal of Experimental Psychology: General* 124, no. 4: 434-452.
- Buswell, G. T. 1937. *How adults read*. Chicago, IL: University of Chicago.

- Caldara, R. and S. Mielliet. 2011. iMap: A novel method for statistical fixation mapping of eye movement data. *Behavior Research Methods*: 1-15.
- Chernyak, D. A. and L. W. Stark. 2001. Top-down guided eye movements. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 31, no. 4: 514-522.
- Côté, M., E. Lecolinet, M. Cheriet, and C. Y. Suen. 1998. Automatic reading of cursive scripts using a reading model and perceptual concepts. *International Journal on Document Analysis and Recognition* 1, no. 1: 3-17.
- Duchowski, Andrew T. and SpringerLink. 2007. *Eye tracking methodology*. 2nd ed. London: Springer.
- Exel, S. and L. Pessoa. 1998. Attentive visual recognition. *Proceedings of the Fourteenth International Conference on Pattern Recognition, 1998* 1, no. 1: 690-692.
- Hacisalihzade, S. S., L. W. Stark, and J. S. Allen. 1992. Visual perception and sequences of eye movement fixations: A stochastic modeling approach. *IEEE Transactions on Systems, Man and Cybernetics* 22, no. 3: 474-481.
- Itti, L., C. Koch, and E. Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, no. 11: 1254-1259.
- Jain, A. K., R. P. W. Duin, and Jianchang Mao. 2000. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, no. 1: 4-37.
- Just, M. A. and P. A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review* 87, no. 4: 329-354.
- Keller, J. G., S. K. Rogers, M. Kabrisky, and M. E. Oxley. 1999. Object recognition based on human saccadic behaviour. *Pattern Analysis & Applications* 2, no. 3: 251-

- 263.
- Kienzle, W., F. A. Wichmann, B. Scholkopf, and M. O. Franz. 2007. A nonparametric approach to bottom-up visual saliency. *Advances in Neural Information Processing Systems* 19, no. 1: 689-696.
- Lauer, F., C. Y. Suen, and G. Bloch. 2007. A trainable feature extractor for handwritten digit recognition. *Pattern Recognition* 40, no. 6: 1816-1824.
- LeCun, Y. and C. Cortes. MNIST handwritten digit database.
<http://yann.lecun.com/exdb/mnist/> (accessed 4/16/2011).
- Legge, G. E., S. J. Ahn, T. S. Klitz, and A. Luebker. 1997. Psychophysics of reading—XVI. the visual span in normal and low vision. *Vision Research* 37, no. 14: 1999-2010.
- Maw, N. N. and M. Pomplun. 2004. Studying human face recognition with the gaze-contingent window technique. In *Proceedings of the twenty-sixth annual meeting of the cognitive science society*, ed. Forbus K., Gentner D., Regier T., 927-932. Chicago, Illinois: Citeseer.
- Meng X. and Z. Wang. 2009. A pre-attentive model of biological vision. *IEEE International Conference on Intelligent Computing and Intelligent Systems, 2009. ICIS 2009* 3, no. 1: 154-158.
- Noton, D. and L. Stark. 1971. Eye movements and visual perception. *Scientific American* 224, no. 6: 35-43.
- Ojanpää, H. 2006. Visual search and eye movements: Studies of perceptual span. University of Helsinki, Faculty of Behavioural Sciences, Department of Psychology and Finnish Institute of Occupational Health. In University of Helsinki.
- Ojanpää, H. and R. Näsänen. 2003. Effects of luminance and colour contrast on the search of information on display devices. *Displays* 24, no. 4-5: 167-178.

- Osberger, W. and A. J. Maeder. 1998. Automatic identification of perceptually important regions in an image. *Proceedings of the Fourteenth International Conference on Pattern Recognition, 1998* 1, no. 1: 701-704.
- Paulson, E. J. and K. S. Goodman. 1999. Influential studies in eye-movement research. International Reading Association, Inc.
<http://www.readingonline.org/research/eyemove.html> (accessed 4/16/2011).
- Peters, R. J. and L. Itti. 2007. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07.* 1, no. 1: 1-8.
- Plannerer, B. 2005. The speech signal. Chap. 1, In *An introduction to speech recognition*. Vol. 1.1, 3. Munich, Germany: speech-recognition.de.
- Privitera, C. M. and L. W. Stark. 2000. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, no. 9: 970-982.
- Rao, R., G. Zelinsky, M. Hayhoe, and D. Ballard. 1996. Modeling saccadic targeting in visual search. *Advances in Neural Information Processing Systems*: 836-842.
- Rao, R. P., G. J. Zelinsky, M. M. Hayhoe, and D. H. Ballard. 2002. Eye movements in iconic visual search. *Vision Research* 42, no. 11: 1447-1463.
- Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124, no. 3: 372-422.
- Reichle, E. D., K. Rayner, and A. Pollatsek. 2003. The E-Z reader model of eye-movement control in reading: Comparisons to other models. *The Behavioral and Brain Sciences* 26, no. 4: 445-76; discussion 477-526.
- Reilly, R. G. and J. K. O'Regan. 1998. Eye movement control during reading: A simulation of some word-targeting strategies. *Vision Research* 38, no. 2: 303-317.

- Rybak, I. A., V. I. Gusakova, A. V. Golovan, L. N. Podladchikova, and N. A. Shevtsova. 1998. A model of attention-guided visual perception and recognition. *Vision Research* 38, no. 15-16: 2387-2400.
- Salah, A. A., E. Alpaydin, and L. Akarun. 2001. A selective attention based method for visual pattern recognition. *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society (Online)*: 881-886.
- Salah, A. A., E. Alpaydin, and L. Akarun. 2002. A selective attention based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, no. 3: 420-425.
- Schomaker, L. and E. Segers. 1999. Finding features used in the human reading of cursive handwriting. *International Journal on Document Analysis and Recognition* 2, no. 1: 13-18.
- Sirotenko, M. 2009. *MNIST-import script for matlab*.
<http://sites.google.com/site/mihailsirotenko/projects/convolutional-neural-network-class> (accessed 4/16/2011).
- Stark, L. W. and C. Privitera. 1997. Top-down and bottom-up image processing. *International Conference on Neural Networks, 1997* 4, no. 1: 2294-2299.
- Stark, L. W. and Y. S. Choi. 1996. Experimental metaphysics: The scanpath as an epistemological mechanism. In *Advances in psychology*, ed. H. S. Stiehl and C. Freksa W.H. Zangemeister. Vol. Volume 116, 3-69North-Holland.
- Suen, C. Y., J. Kim, K. Kim, Q. Xu, and L. Lam. 2000. Handwriting recognition-the last frontiers. *Proceedings of the 15th International Conference on Pattern Recognition, 2000* 4, no. 1: 1-10.
- Suen, C. Y. and J. Tan. 2005. Analysis of errors of handwritten digits made by a

- multitude of classifiers. *Pattern Recognition Letters* 26, no. 3: 369-379.
- Tappert, C. C., C. Y. Suen, and T. Wakahara. 1990. The state of the art in online handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12, no. 8: 787-808.
- Tinker, M. A. 1936. Reliability and validity of eye-movement measures of reading. *Journal of Experimental Psychology* 19, no. 6: 732-746.
- Tolonen, T. and M. Karjalainen. 2000. A computationally efficient multipitch analysis model. *IEEE Transactions on Speech and Audio Processing* 8, no. 6: 708-716.
- Watanabe, S. 1985. *Pattern recognition: Human and mechanical*. New York: Wiley.
- Watanabe, Y., J. Gyoba, and K. Maruyama. 1983. Reaction time and eye movements in the recognition task of hand-written katakana-letters: An experimental verification of the discriminant analysis of letter recognition by Hayashi's quantification. *Shinrigaku Kenkyu : The Japanese Journal of Psychology* 54, no. 1: 58-61.
- Woodford, O. 2007. *SC - powerful image rendering*. Vol. 2010The MathWorks Inc.
- Yagi, T., K. Gouhara, and Y. Uchikawa. 1993. An algorithm of eye movement in selective fixation. *IEEE International Conference on Neural Networks, 1993* 2, no. 1: 761-765.
- Yarbus, A. L. 1967. *Eye movements and vision*. New York: Plenum press.
- Zhang, W., Y. Hyejin, S. Dimitris, and G. Zelinsky. 2006. A computational model of eye movements during object class detection. In *Advances in neural information processing systems 18*, ed. Y. Weiss, B. Schölkopf, and J. Platt, 1609-1616. Cambridge, MA: MIT Press.

APPENDIX

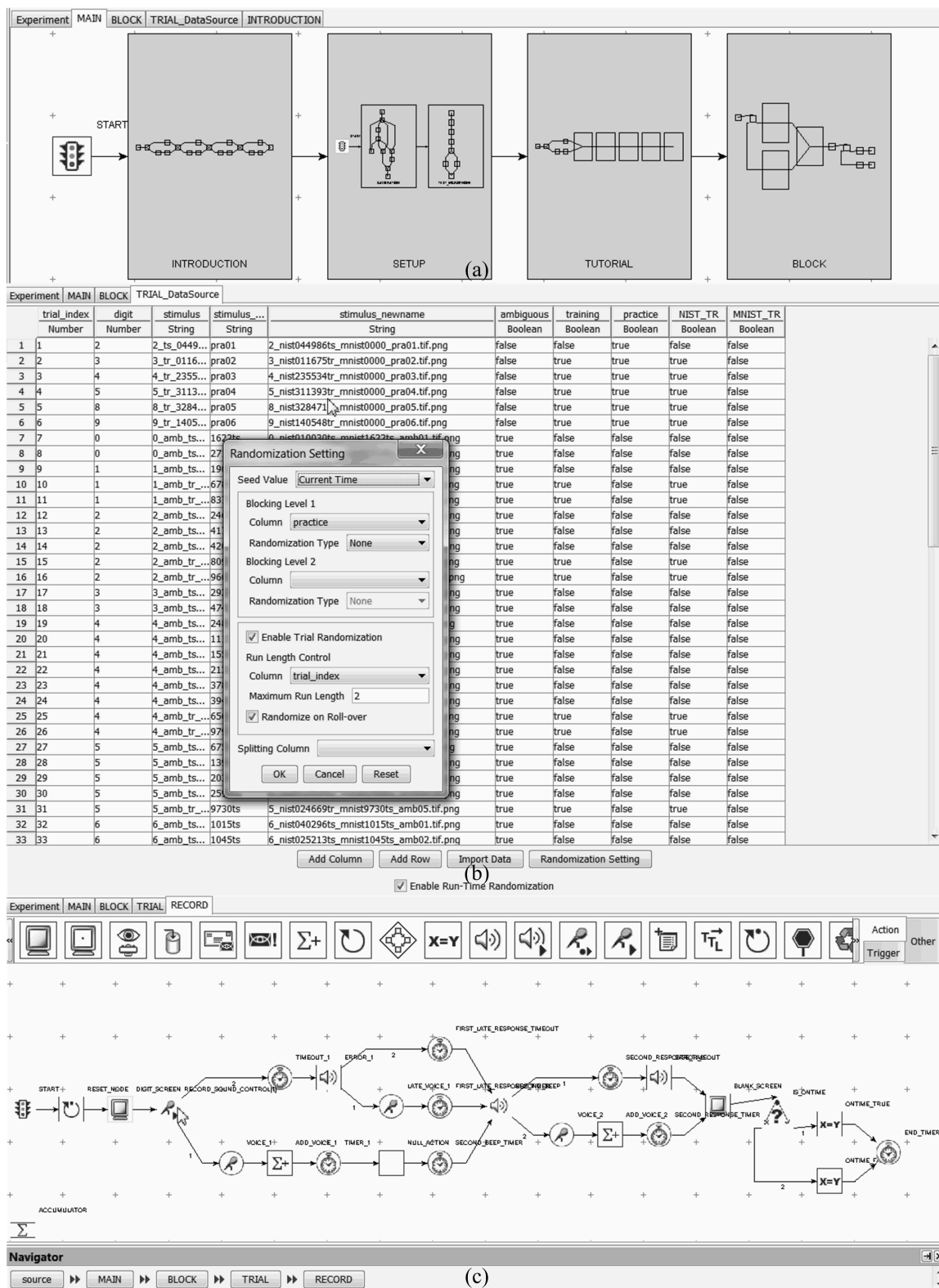


Fig. A1. Verbal task storyboard in Experiment Builder: (a) Main sequence, (b) trial data source and randomization and (c) trial record sequence.

The image displays two screenshots of the SR Research Experiment Builder v 1.10.1 software interface.

Top Screenshot: Task Introduction Slide

The main window shows a slide with the following text:

In this task:

In this task, a *handwritten-number* from **0** to **9** will appear in a **very small** area at the center of the screen for **ten** seconds.

During this time you are asked to *identify* the displayed number by **pressing** the *number-key* that **best corresponds** to it using your *number-pad*.

A brief *sound* will then play indicating whether the answer you provided was **correct** or **incorrect**.

When ready, press 0 on your number-pad to practice.

The interface includes a menu bar (File, Edit, View, Experiment, Help), a toolbar, an Overview window, a Structure window showing a hierarchy of nodes (KEY_PRESS_TASK, INTRODUCTION, WELCOME_SCREEN, KEYBOARD, BLANK_SCREEN), a Properties window with a table of properties, and a Navigator window.

Property	Value
Label	WELCOME_SCREEN[3]
Type	DisplayScreen
Node Path	KEY_PRESS_TASK.IN...
Time	
Start Time	
Clear Input Queues	<input checked="" type="checkbox"/>
Prepare Time	
Width	1024
Height	768
Background Color	
Bits Per Pixel	32
Auto Generate Sync ...	<input type="checkbox"/>
Resource Count	
Grid Rows	2

Bottom Screenshot: Trial Storyboard

The main window shows a storyboard diagram on a grid. The diagram illustrates the flow of a trial sequence:

- START** leads to **INF_TRIAL_VARIABLES_SEQUENCE**.
- INF_TRIAL_VARIABLES_SEQUENCE** leads to **TRIAL_EVENT**.
- TRIAL_EVENT** leads to **NULL_ACTIONS**.
- NULL_ACTIONS** leads to **IS_PRACTICE**.
- IS_PRACTICE** leads to **IS_CORRECT**.
- IS_CORRECT** leads to **CORRECT_SCREEN**.
- CORRECT_SCREEN** leads to **TIMER**.
- TIMER** leads to **RECYCLE_DATA_WEDGES_SCREEN**.
- RECYCLE_DATA_WEDGES_SCREEN** leads to **TIMER_SOUND**.
- TIMER_SOUND** leads to **BLANK_SCREEN**.

The Properties window for the TRIAL node is visible, showing:

Property	Value
Label	TRIAL
Type	Sequence
Node Path	KEY_PRESS_TASK.BLO...
Time	
Is Real Time	<input type="checkbox"/>
Iteration	
Iteration Count	91
Split by	[17, 74]
Data Source	Co[91]s: 10 / Rows: 91
Freeze Display Until Fir...	<input checked="" type="checkbox"/>
Prompt for Dataset File	<input type="checkbox"/>

The Navigator window shows the sequence: source >> KEY_PRESS_TASK >> BLOCK >> TRIAL.

Fig. A2. Manual task storyboard in Experiment Builder. *Top*, Task introduction slide; *bottom*, trial sequence with mouse cursor pointing to total number of trials and how they are *split* in practice and test blocks.

Name	Date modified	Date created
76_7_amb_tr_32184...	09/03/2010 2:35 PM	08/11/2010 10:34 AM
77_6_amb_tr_08072...	09/03/2010 2:35 PM	08/11/2010 10:34 AM
78_4_amb_ts_05530...	09/03/2010 2:36 PM	08/11/2010 10:34 AM
79_2_amb_tr_06732...	09/03/2010 2:36 PM	08/11/2010 10:34 AM
80_9_amb_ts_02542...	09/03/2010 2:36 PM	08/11/2010 10:34 AM
81_6_amb_ts_00262...	09/03/2010 2:36 PM	08/11/2010 10:34 AM
82_9_ts_028771.wav	09/03/2010 2:36 PM	08/11/2010 10:34 AM
ab_mar09.asc	06/10/2010 6:57 PM	08/11/2010 10:34 AM
ab_mar09.edf	09/03/2010 2:37 PM	08/11/2010 10:34 AM
actual_MAIN_DataS...	09/03/2010 2:12 PM	08/11/2010 10:34 AM
actual_TRIAL_DataS...	09/03/2010 2:36 PM	08/11/2010 10:34 AM
actual_TUTOR_AM...	09/03/2010 2:14 PM	08/11/2010 10:34 AM
notes.txt	09/03/2010 3:56 PM	08/11/2010 10:34 AM
warning.log	09/03/2010 2:36 PM	08/11/2010 10:34 AM

(a)

Name
actual_TRIAL_DataSource_EyeSpy_Exp2_Full_KEY_PRESS_TASK_BLOCKTRIAL.dat
RESULTS_FILE.txt
warning.log

(b)

Fig. A3. Session data folder and files for the same participant under (a) Verbal task and (b) Manual task.

CONSENT FORM

This is to state that I agree to participate in a program of research being conducted by Prof. Michael von Grünau of the department of Psychology, Prof. Ching Y. Suen and Prof. Adam Krzyzak of the department of Computer Science at Concordia University. This experiment will be conducted by Masters Candidate Nabil Khoury.

A. PURPOSE

I have been informed that the purpose of the research is to find out how computers can be taught to recognize handwritten script in a way similar to humans. The results of this study will be used towards the creation of a computer model of human visual recognition.

B. PROCEDURES

Your participation in this study will involve participating in a lab experiment where you will be asked to identify handwritten symbols while a mounted camera device, called an Eye-tracker, collects various eye-related data such as focus and pupil size. The experiment will be conducted at the Concordia Vision Laboratory located in the Science Pavilion at Concordia University's Loyola Campus. The duration of your participation will be 30 minutes. The results of this study may be published but your name and identity will not be revealed and your record will remain confidential; data collected during the experiment will be stored anonymously and used to teach computers how to recognize images in a way similar to ours. Your data will not be used as a measure of individual aptitude and no score will be assigned to you on the basis of this experiment.

C. RISKS AND BENEFITS

The risks to you as a participant are minimal. You will be exposed to very safe levels of infrared light designed specifically for eye tracking research. Participation in this study will be compensated with 1 participation credit (*½ mark towards final grade in a pre-approved course.*) Identifying images may be tiresome for some but you will be free to discontinue the experiment at any time you feel any discomfort without negative consequences. As such, if you withdraw from participating you will still receive the same number of credits. Your participation will also benefit others by promoting the understanding of Man's vast recognition abilities and to contribute to the development of computer programs that can echo these abilities hence allowing us to further that understanding.

D. CONDITIONS OF PARTICIPATION

- I understand that I am free to withdraw my consent and discontinue my participation at anytime without negative consequences.
- I understand that my participation in this study is:
CONFIDENTIAL (i.e., the researcher will know, but will not disclose my identity)
- I understand that the data from this study may be published.

I HAVE CAREFULLY STUDIED THE ABOVE AND UNDERSTAND THIS AGREEMENT. I

FREELY CONSENT AND VOLUNTARILY AGREE TO PARTICIPATE IN THIS STUDY.

NAME (please print) _____

SIGNATURE _____

If at any time you have questions about your rights as a research participant, please contact Adela Reid, Research Ethics and Compliance Officer, Concordia University, at (514) 848-2424 x7481 or by email at areid@alcor.concordia.ca.

Fig. A4. Participant consent form.

DEBRIEFING SHEET
HANDWRITING IDENTIFICATION STUDY

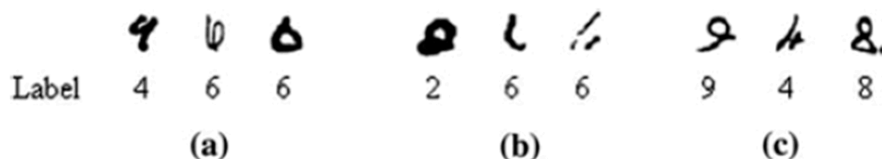
Thank you for participating in the Handwriting Identification Study conducted by Masters Candidate Nabil Khoury under supervision of Prof. Michael von Grünau of the department of Psychology, Prof. Ching Y. Suen and Prof. Adam Krzyzak of the department of Computer Science at Concordia University.

A. CONTACTS

- Research Assistant Nabil Khoury. Phone: (514) 848-2424 ext. 2212. E-mail: n_khoury@cse.concordia.ca
- Dr. Michael von Grünau. Phone: (514) 848-2424 ext. 2190. E-mail: michael.vongrunau@concordia.ca
- Research Compliance Officer at the Concordia University Office of Research: 848-2424, ext.4888
- Dr. Virginia Penhune. Associate Professor and Chair, Psychology Department Ethics Committee. Tel: 514-848-2424 ext. 7535. E-mail:Virginia.Penhune@Concordia.ca

B. STUDY OBJECTIVES

In handwritten digit recognition, computers are found to make three categories of errors.



1. Category 1, accounting for around a quarter of recognition errors, is of digit images that are easily confused with other numerals because they resemble more than one numeral. Images in this category usually belong to these confusing pairs: 4–9, 0–6, and 3–5 (Figure a).
2. Category 2, accounting for around an eighth of all errors, is of digits that are difficult to recognize by classifiers and humans because of extraneous factors like poor scanners to peculiar writing habits (Figure b).
3. Category 3, accounting for 62.70% of recognition errors, is of digits that humans can recognize without any ambiguity (Figure c.) but are nevertheless misrecognized by computers because they somewhat differ from the samples used to train the computer.

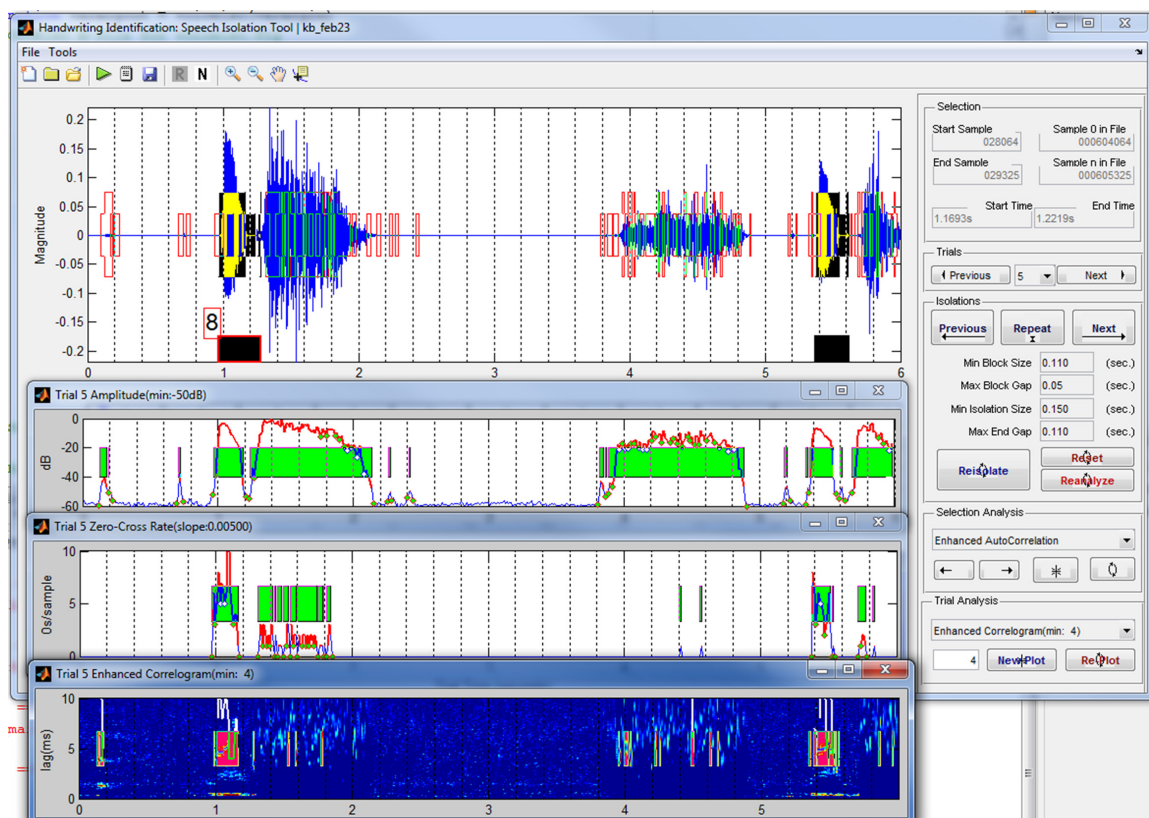
Our hypothesis is that since humans are superior to computers in identifying category 3 digits and since human eye-movement often attends to the most distinctive features of an image, we can significantly improve computer recognition of these digits by using participant eye-movement recorded during identification of these digits.

A secondary purpose of the research is to find out whether our participants will be better at identifying category 1 and 2 digits than the best computers and whether we can use the above scheme to improve computer identification of these two categories.

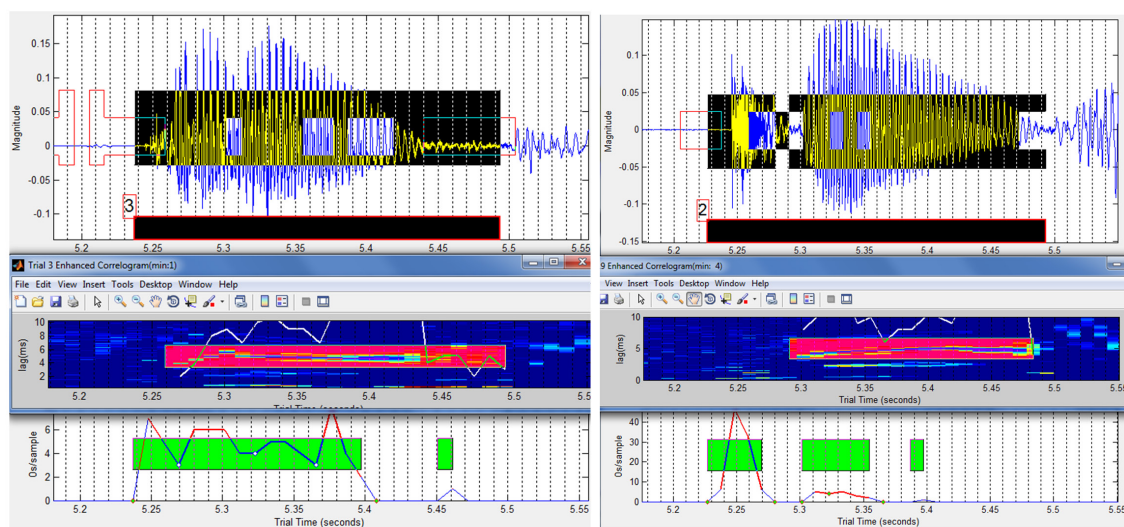
C. SUGGESTED ARTICLES

- C.Y. Suen, J. Tan, Analysis of errors of handwritten digits made by a multitude of classifiers, *Pattern Recognition Letters Volume 26*, 2005, pp.369–379.
- C.M. Privitera, and L.W. Stark, "Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations", *IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 22, Issue 9, September 2000*, pp. 970 - 982.

Fig. A5. Participant debriefing sheet.



(a)



(b)

(c)

Fig. A6. Overview of verbal response isolation tool. (a) Shows trial audio with excessive breathing noise and the three main heuristics used to isolate verbal responses; *top to bottom*: Audio-recording view, amplitude view, zero-crossing rate view and enhanced correlagram (ESACF) view; (b) and (c) show how trial periods selected using ESACF complements those selected using zero-crossing rate allowing accurate determination of verbal responses as shown in isolation labels 3 and 2 respectively.

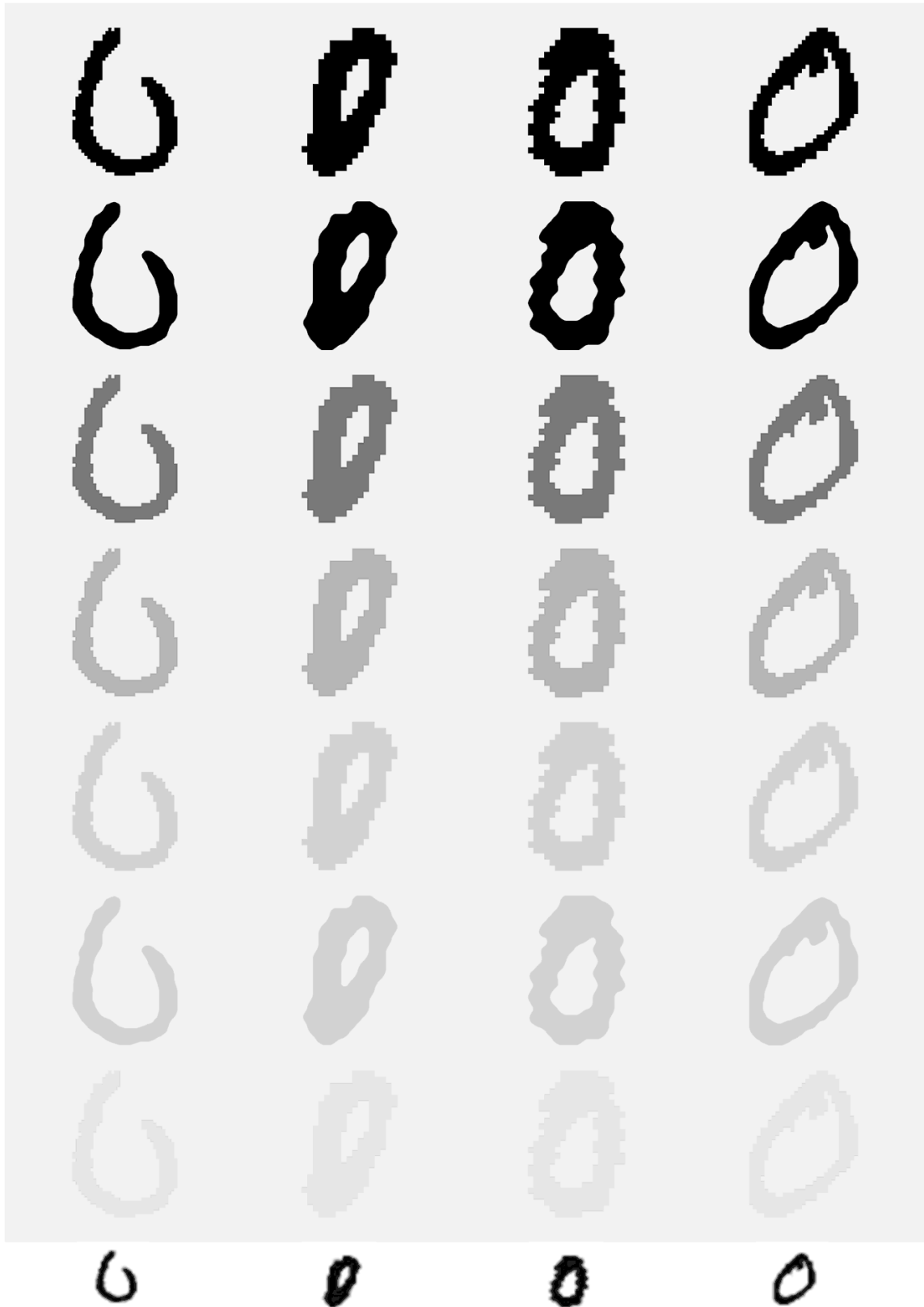


Fig. A7. Verbal and Manual task zeros. *Top to bottom*: Unsmoothed-FG0, Smoothed-FG0, Unsmoothed-FG120, Unsmoothed-FG180, Unsmoothed-FG210, Smoothed-FG210, Unsmoothed-FG228 and Manual.



Fig. A8. Verbal and Manual task ones. *Top to bottom*: Unsmoothed-FG0, Smoothed-FG0, Unsmoothed-FG120, Unsmoothed-FG180, Unsmoothed-FG210, Smoothed-FG210, Unsmoothed-FG228 and Manual.

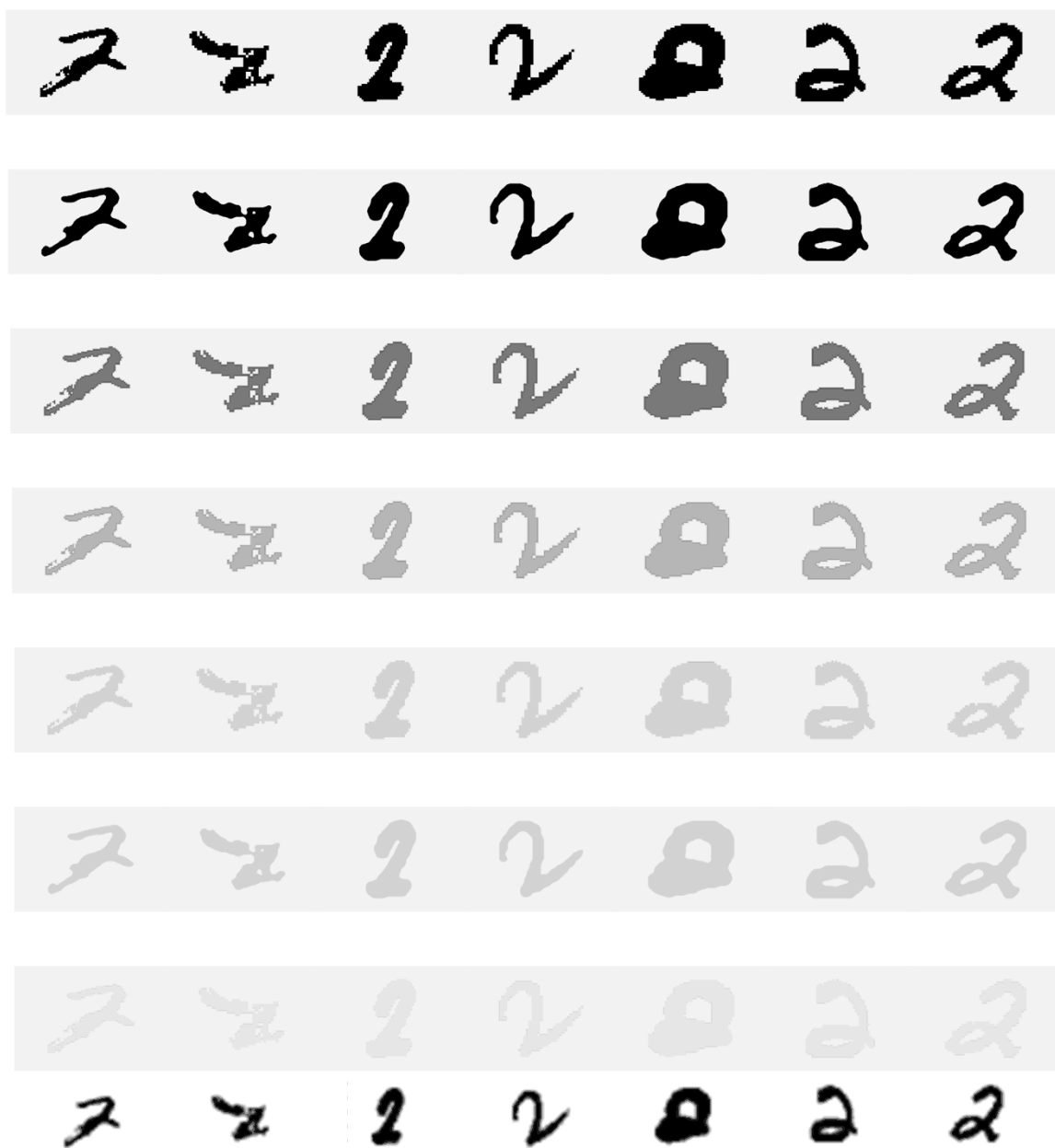


Fig. A9. Verbal and Manual task twos. *Top to bottom*: Unsmoothed-FG0, Smoothed-FG0, Unsmoothed-FG120, Unsmoothed-FG180, Unsmoothed-FG210, Smoothed-FG210, Unsmoothed-FG228 and Manual.

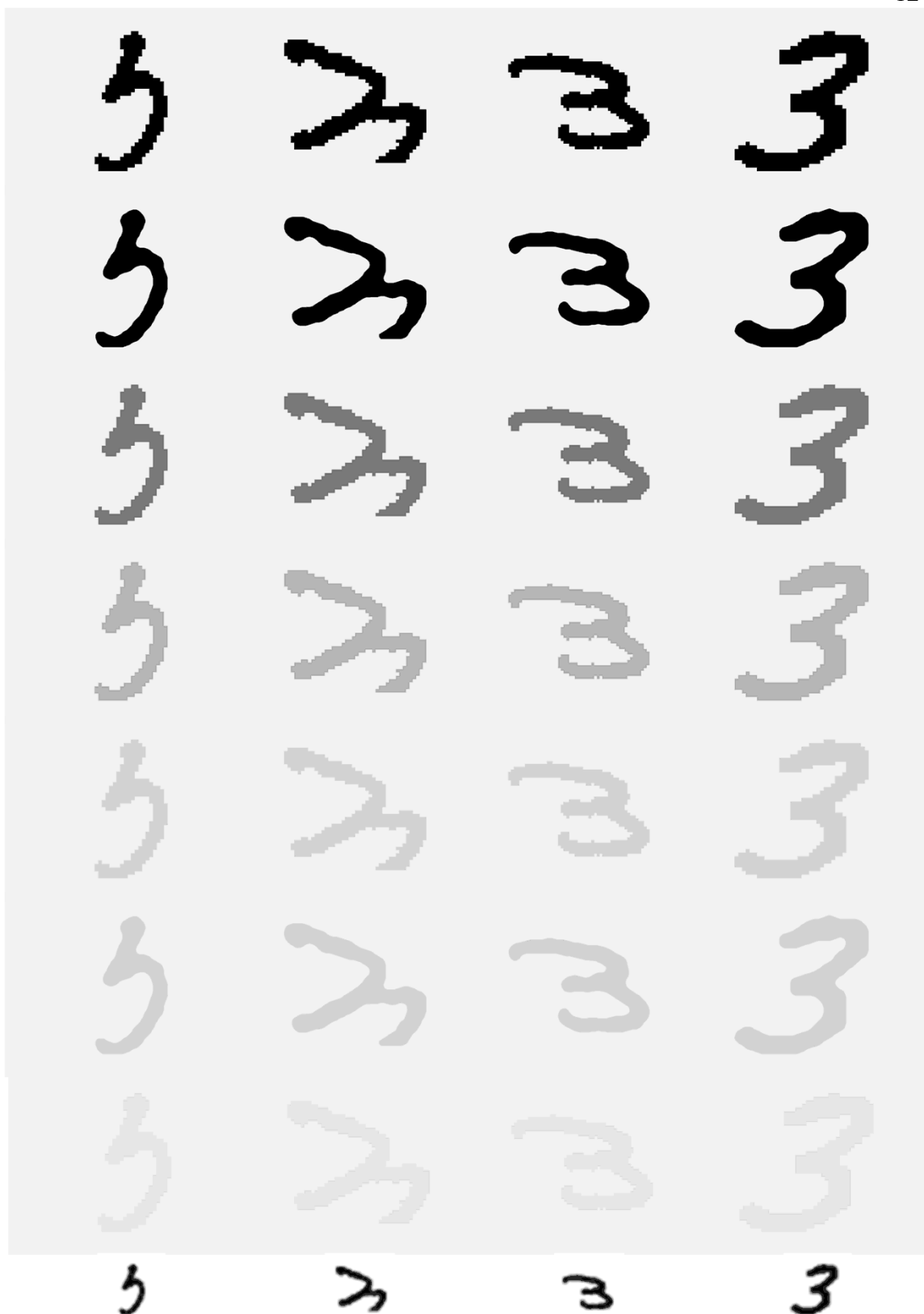


Fig. A10. Verbal and Manual task three. *Top to bottom*: Unsmoothed-FG0, Smoothed-FG0, Unsmoothed-FG120, Unsmoothed-FG180, Unsmoothed-FG210, Smoothed-FG210, Unsmoothed-FG228 and Manual.

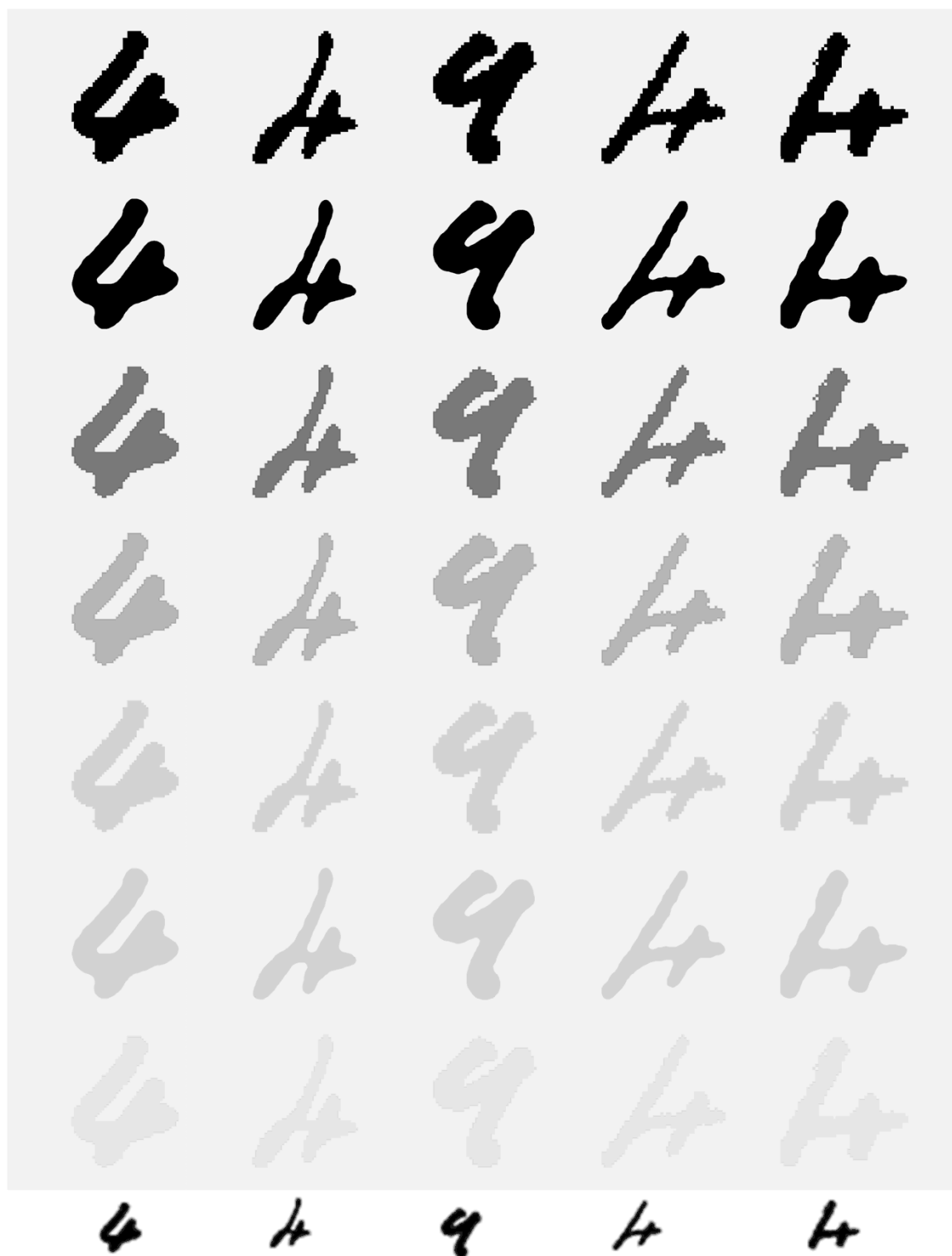


Fig. A11. Verbal and Manual task fours I. *Top to bottom*: Unsmoothed-FG0, Smoothed-FG0, Unsmoothed-FG120, Unsmoothed-FG180, Unsmoothed-FG210, Smoothed-FG210, Unsmoothed-FG228 and Manual.

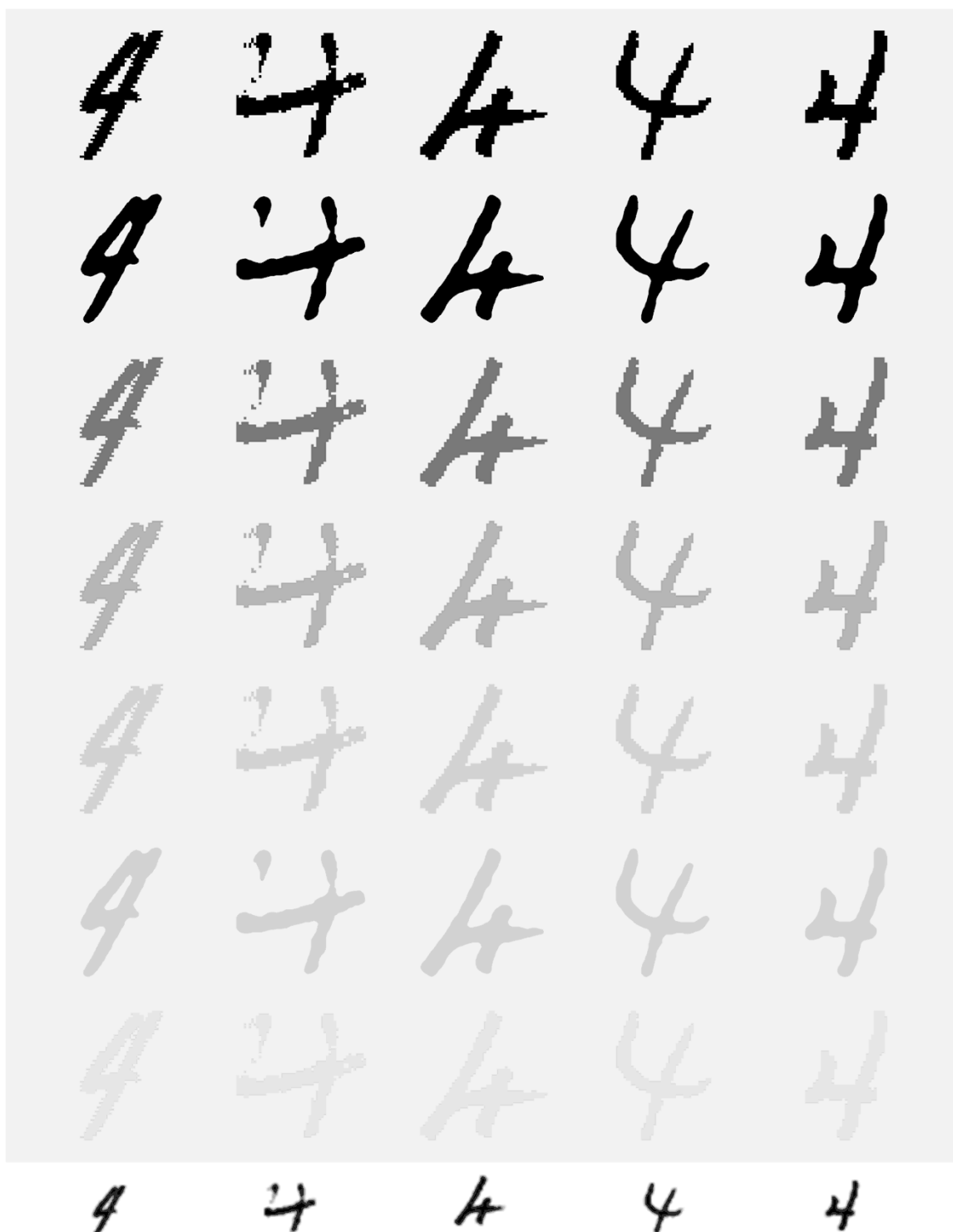


Fig. A12. Verbal and Manual task fours II. *Top to bottom*: Unsmoothed-FG0, Smoothed-FG0, Unsmoothed-FG120, Unsmoothed-FG180, Unsmoothed-FG210, Smoothed-FG210, Unsmoothed-FG228 and Manual.

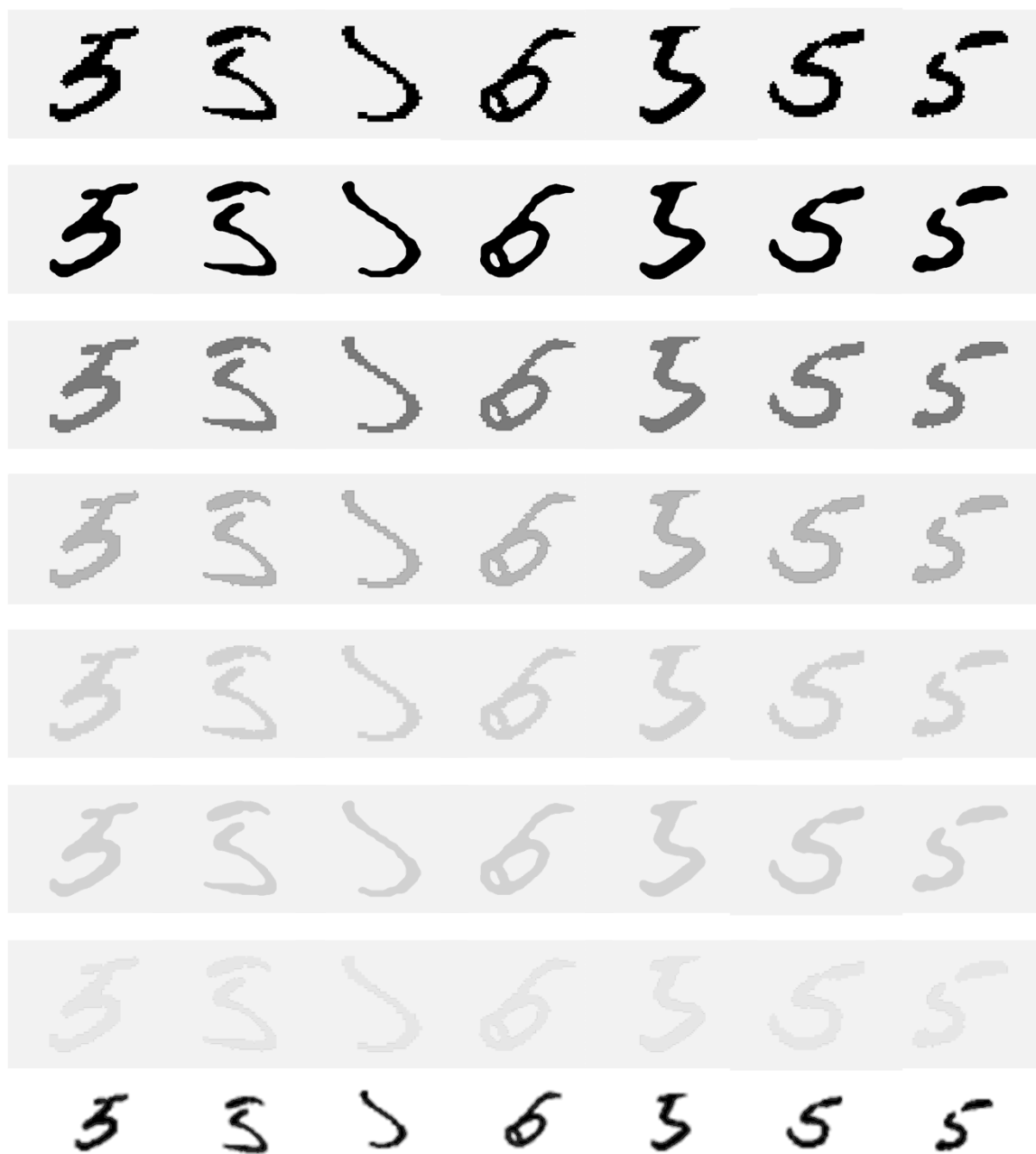


Fig. A13. Verbal and Manual task fives. *Top to bottom*: Unsmoothed-FG0, Smoothed-FG0, Unsmoothed-FG120, Unsmoothed-FG180, Unsmoothed-FG210, Smoothed-FG210, Unsmoothed-FG228 and Manual.



Fig. A14. Verbal and Manual task sixes I. *Top to bottom*: Unsmoothed-FG0, Smoothed-FG0, Unsmoothed-FG120, Unsmoothed-FG180, Unsmoothed-FG210, Smoothed-FG210, Unsmoothed-FG228 and Manual.



Fig. A15. Verbal and Manual task sixes II. *Top to bottom*: Unsmoothed-FG0, Smoothed-FG0, Unsmoothed-FG120, Unsmoothed-FG180, Unsmoothed-FG210, Smoothed-FG210, Unsmoothed-FG228 and Manual.

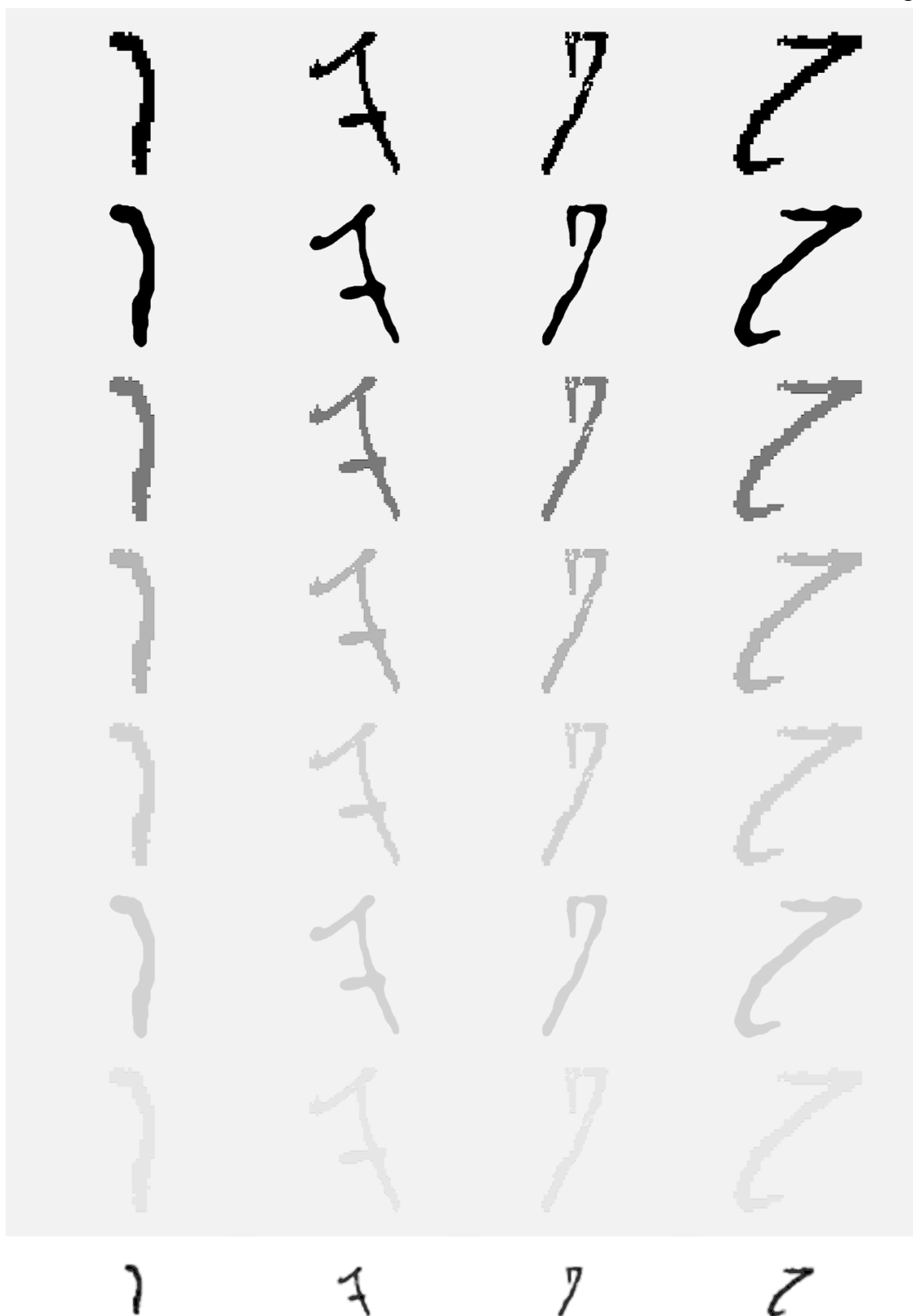


Fig. A16. Verbal and Manual task sevens I. *Top to bottom:* Unsmoothed-FG0, Smoothed-FG0, Unsmoothed-FG120, Unsmoothed-FG180, Unsmoothed-FG210, Smoothed-FG210, Unsmoothed-FG228 and Manual.



Fig. A17. Verbal and Manual task sevens II. *Top to bottom:* Unsmoothed-FG0, Smoothed-FG0, Unsmoothed-FG120, Unsmoothed-FG180, Unsmoothed-FG210, Smoothed-FG210, Unsmoothed-FG228 and Manual.



Fig. A18. Verbal and Manual task eights I. *Top to bottom*: Unsmoothed-FG0, Smoothed-FG0, Unsmoothed-FG120, Unsmoothed-FG180, Unsmoothed-FG210, Smoothed-FG210, Unsmoothed-FG228 and Manual.

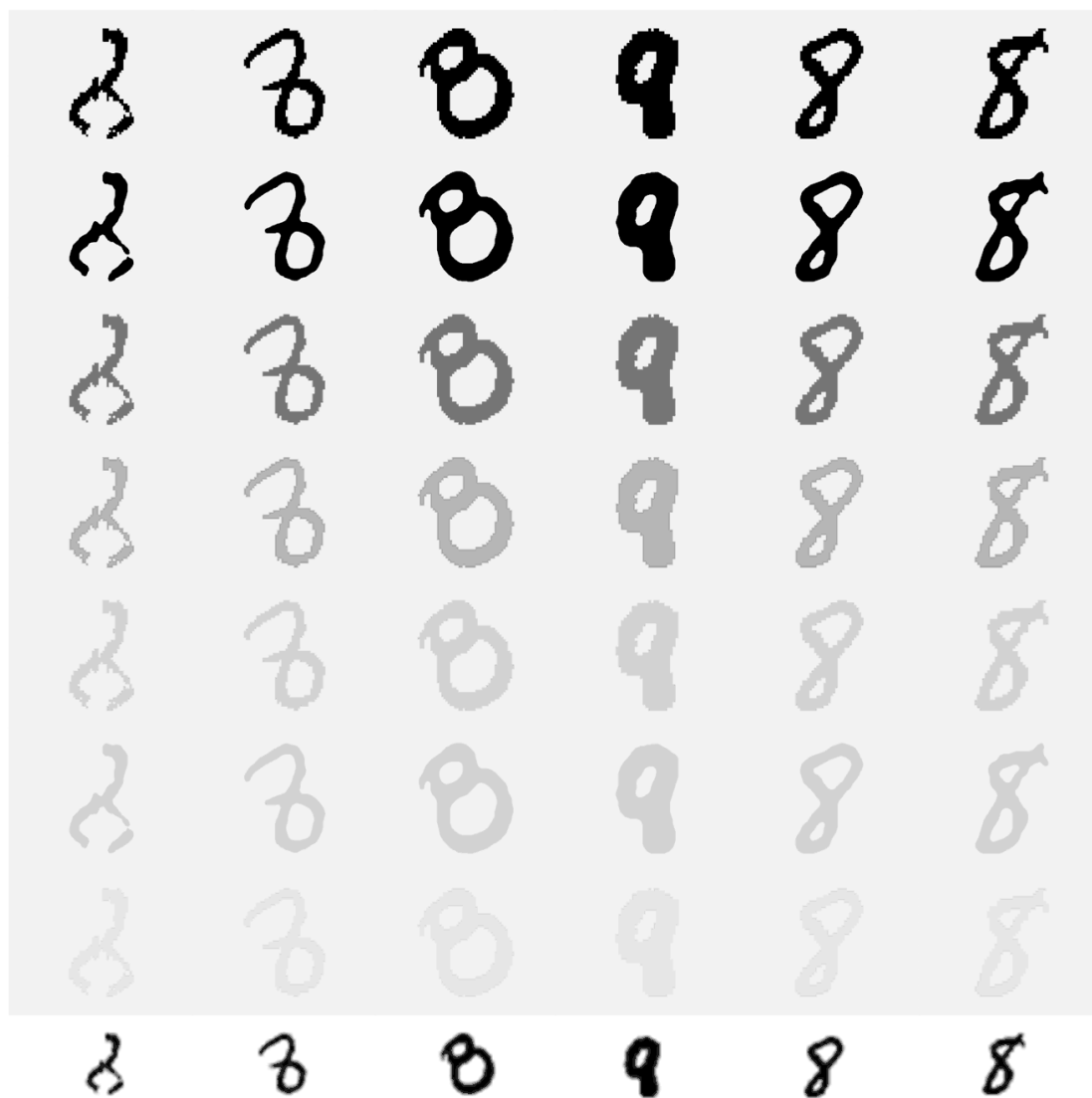


Fig. A19. Verbal and Manual task eights II. *Top to bottom*: Unsmoothed-FG0, Smoothed-FG0, Unsmoothed-FG120, Unsmoothed-FG180, Unsmoothed-FG210, Smoothed-FG210, Unsmoothed-FG228 and Manual.

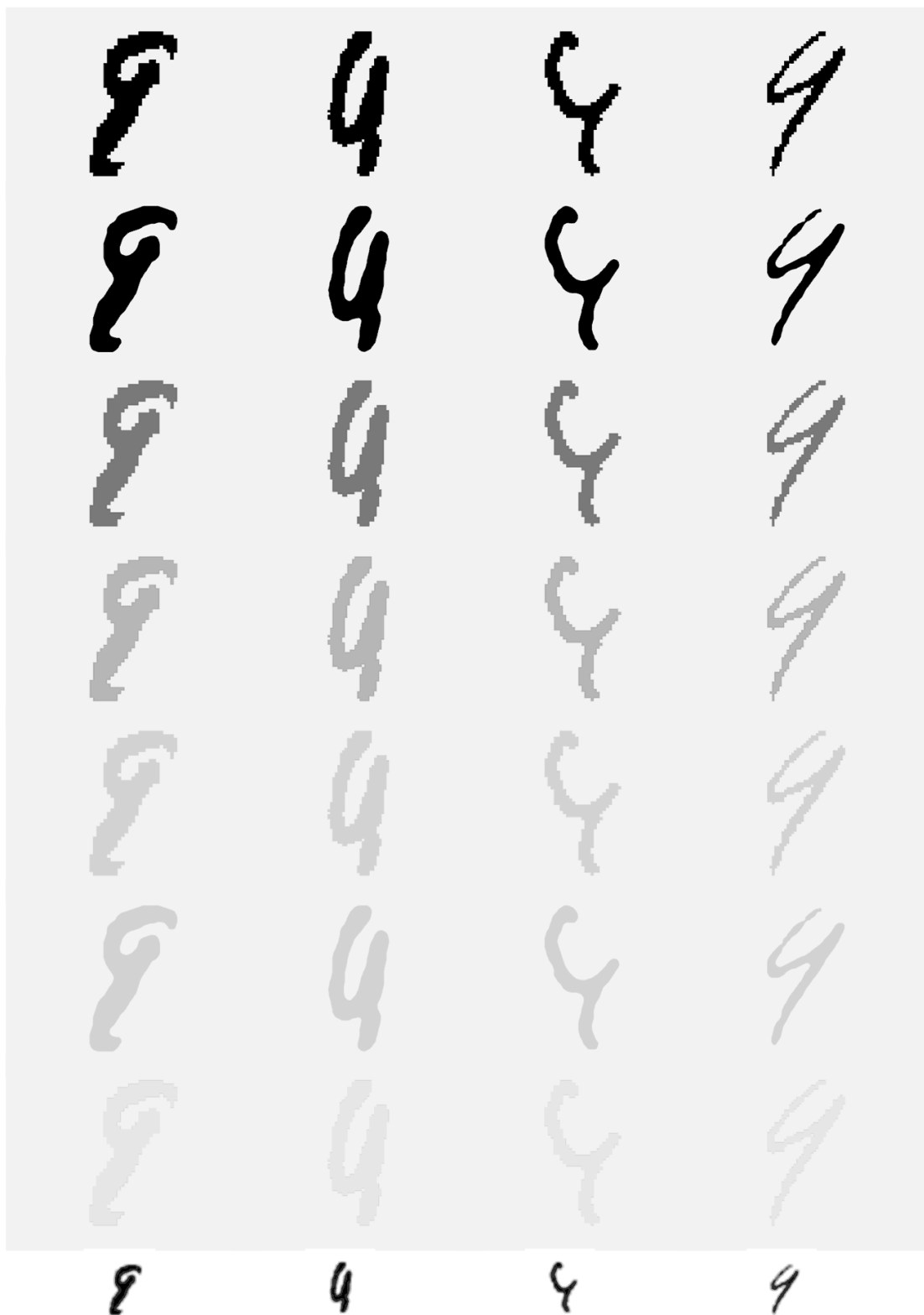


Fig. A20. Verbal and Manual task nines I. Top to bottom: Unsmoothed-FG0, Smoothed-FG0, Unsmoothed-FG120, Unsmoothed-FG180, Unsmoothed-FG210, Smoothed-FG210, Unsmoothed-FG228 and Manual.

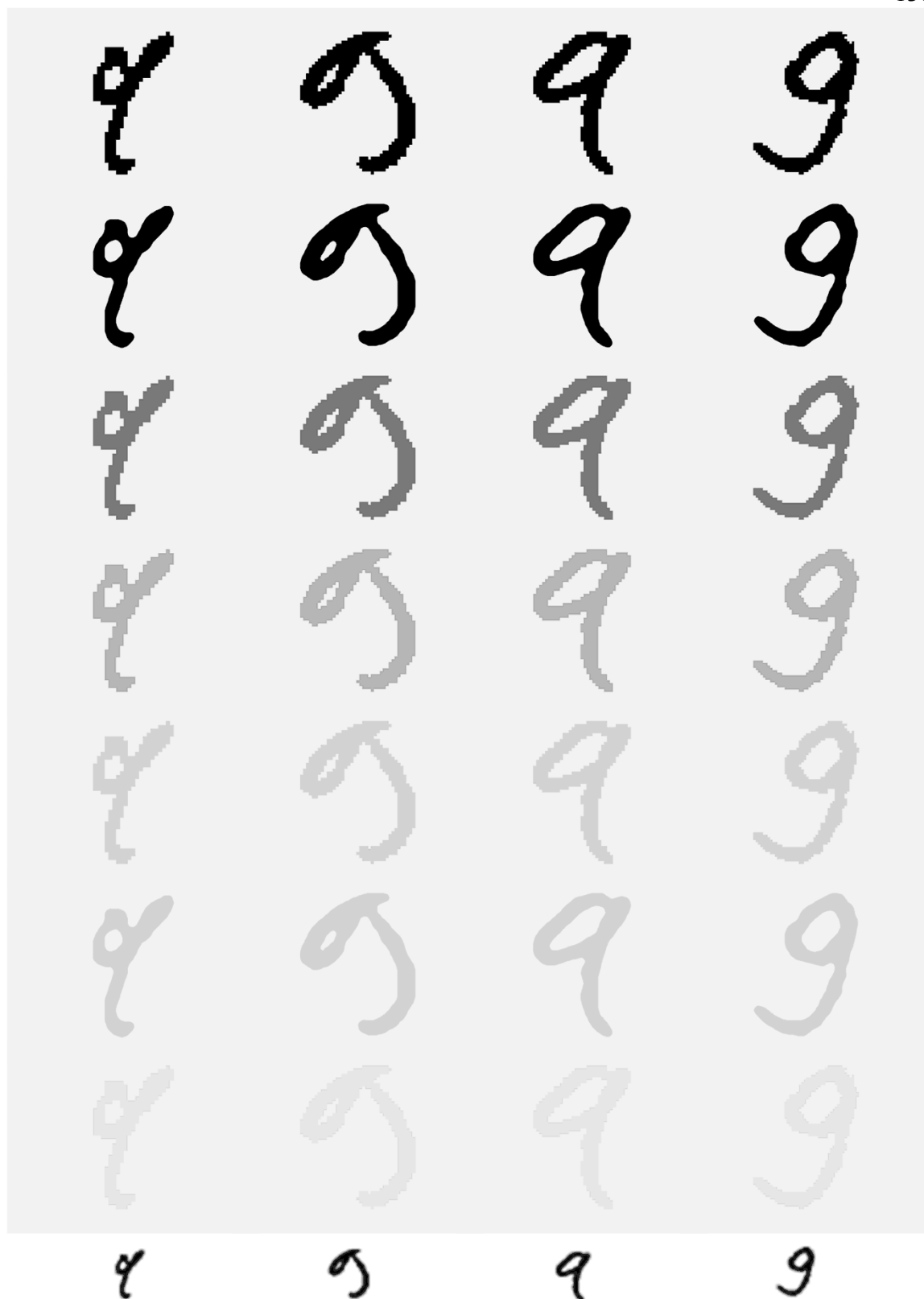


Fig. A21. Verbal and Manual task nines II. *Top to bottom*: Unsmoothed-FG0, Smoothed-FG0, Unsmoothed-FG120, Unsmoothed-FG180, Unsmoothed-FG210, Smoothed-FG210, Unsmoothed-FG228 and Manual.

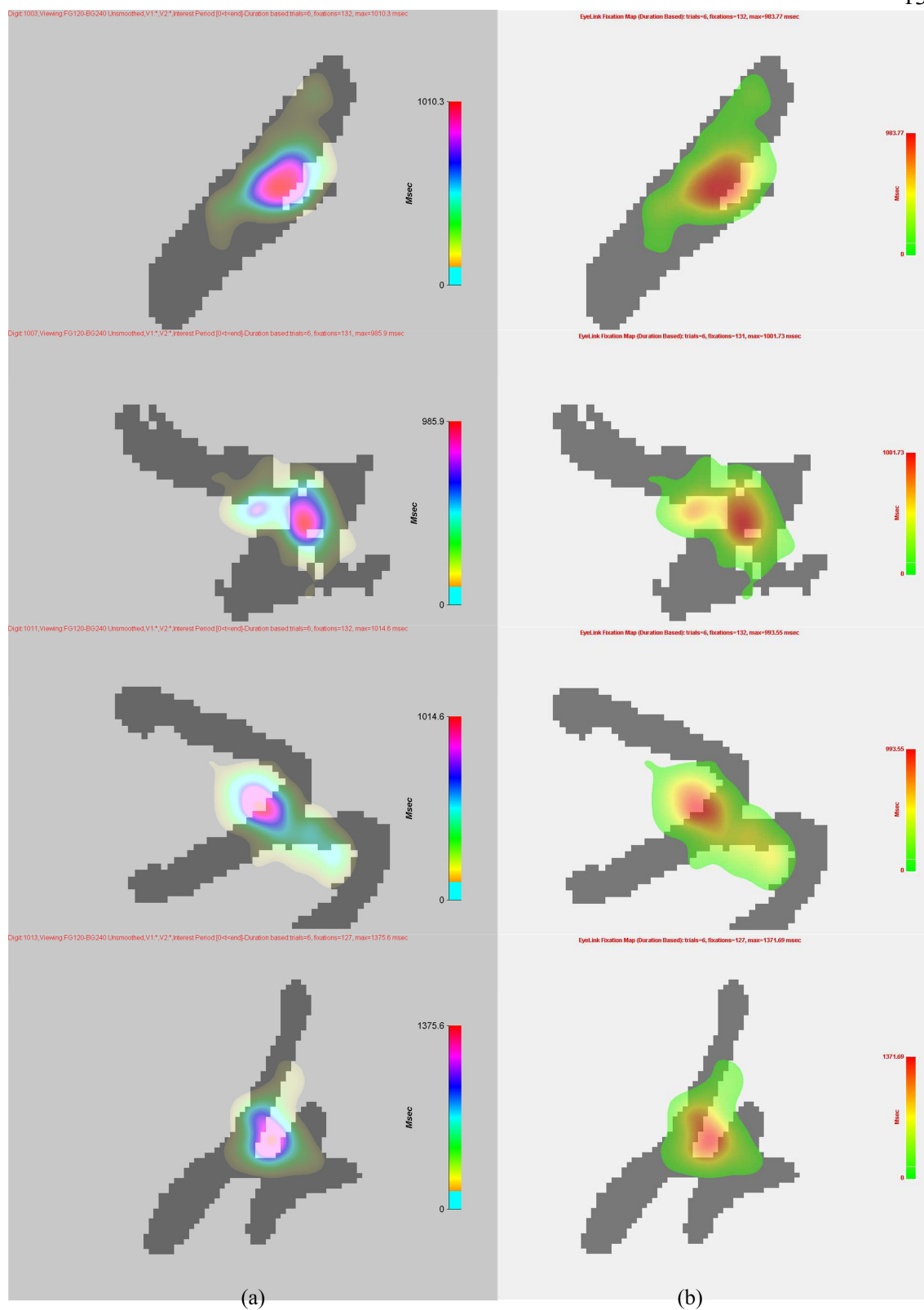


Fig. A22. Side-by-side comparison of duration-based heat maps generated using: (a) Gaussian convolution and (b) SR Research Data Viewer on the same set of digit stimuli and visual fixations taken from all participants in Unsmoothed-FG120.

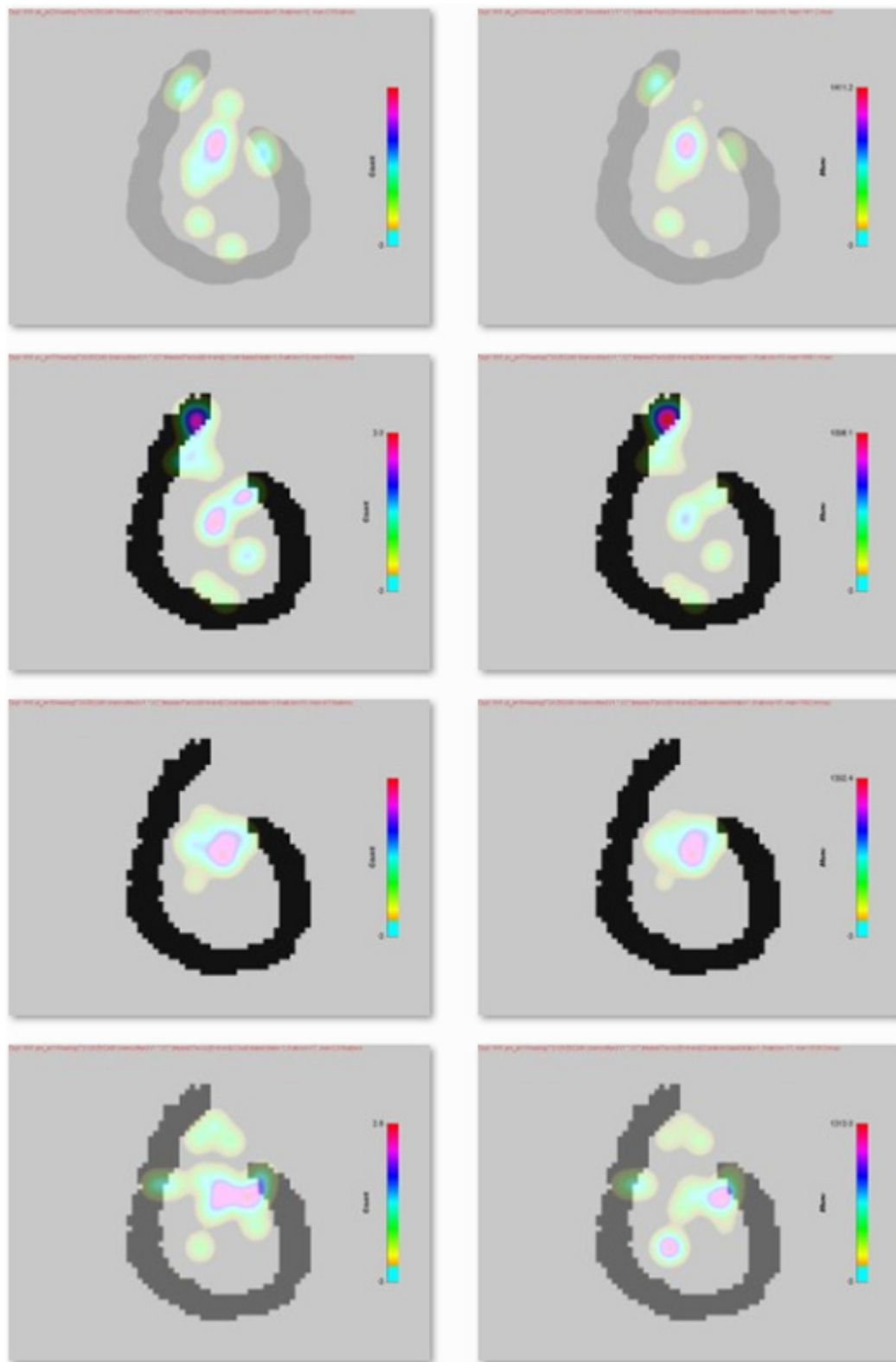


Fig. A23. Side-by-side comparison between count-based (*left*) and corresponding duration-based (*right*) heatmaps each representing gaze during a full-trial period in Verbal task.

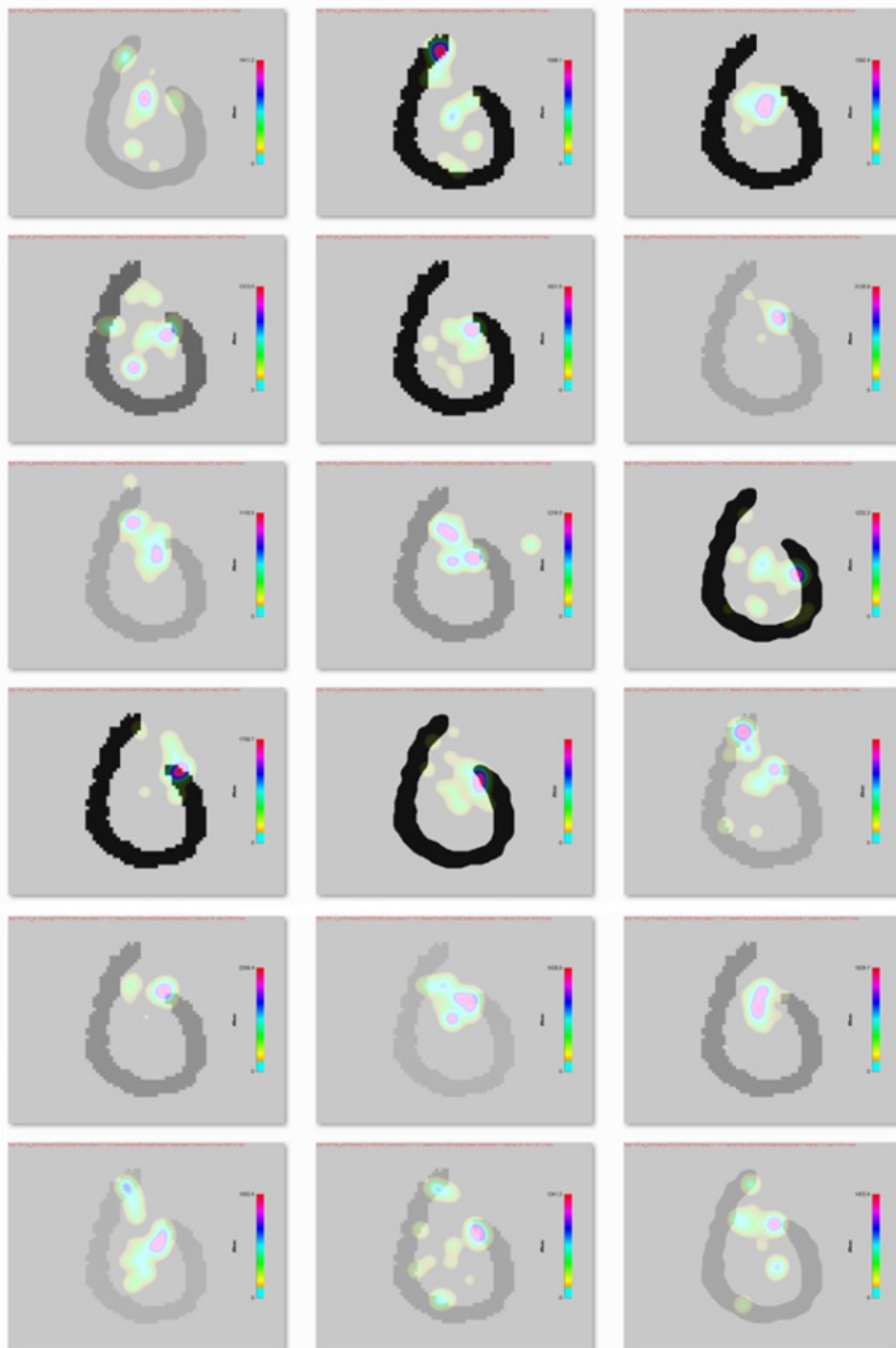


Fig. A24. A random assortment of duration-based heat maps each representing gaze during full-trial period within and across viewing conditions of Verbal task.

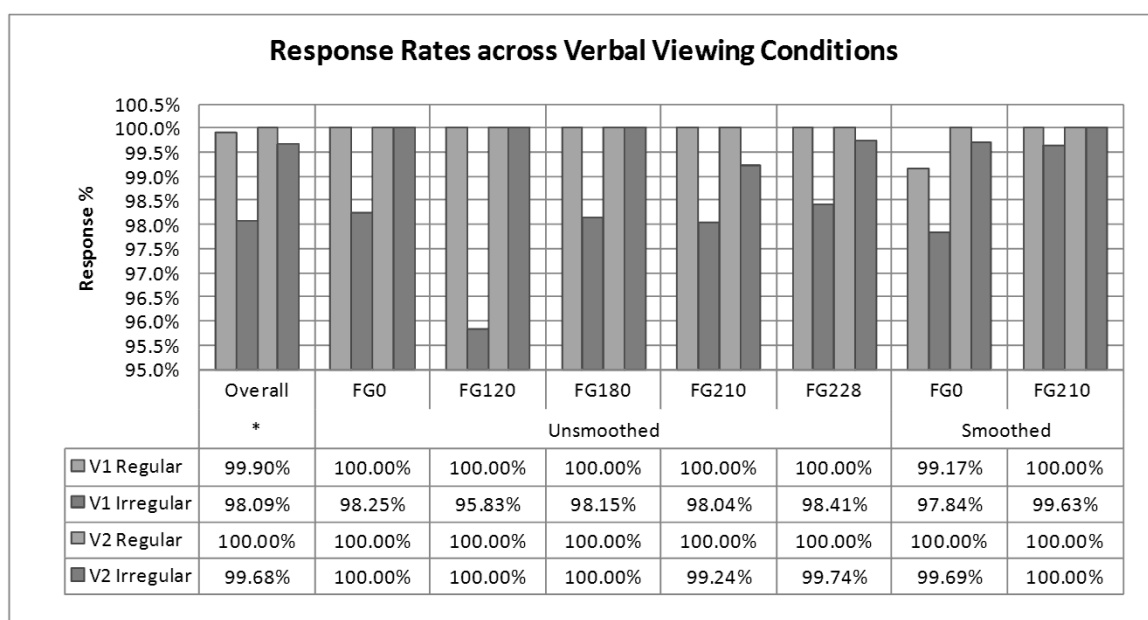
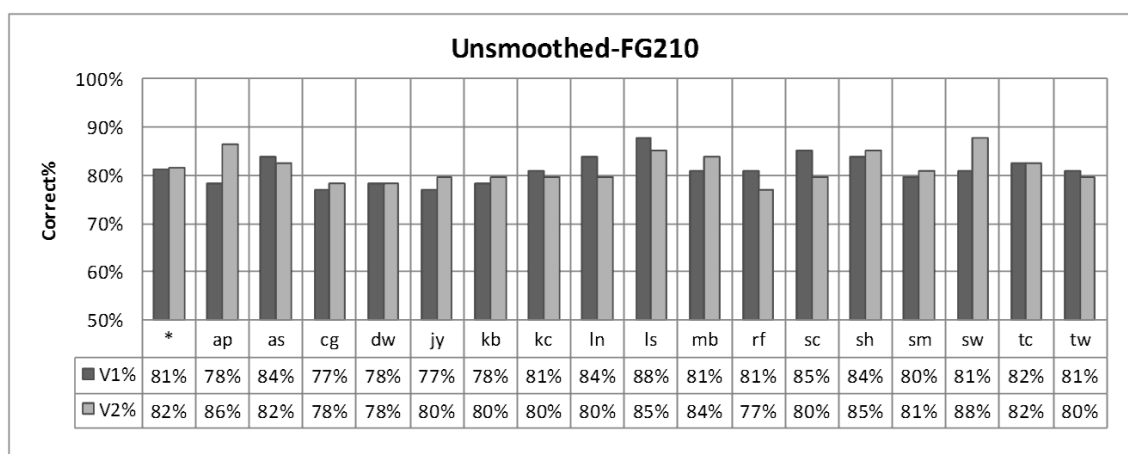
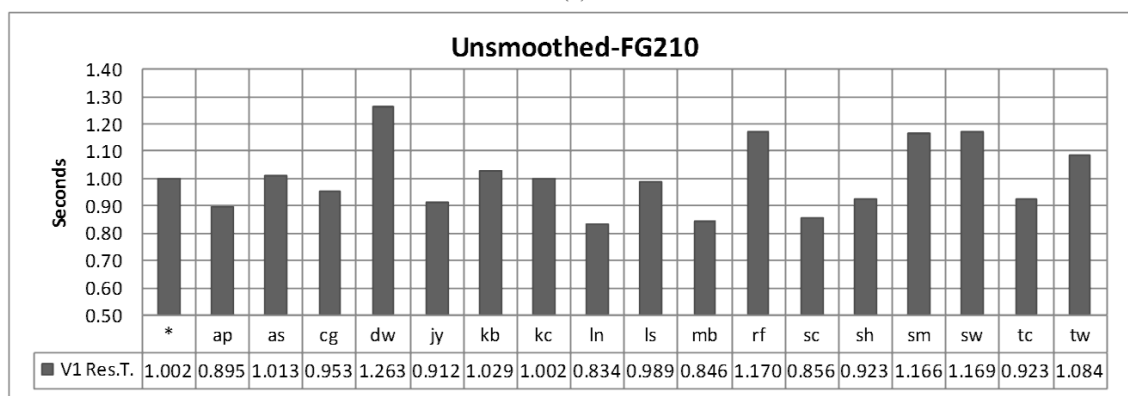


Fig. A25. Response rate across viewing conditions in Verbal task



(a)



(b)

Fig. A26. Individual variations among 17 participants under Unsmoothed-FG210 in (a) average correct identification for irregular digits and (b) average response time for irregular digits.

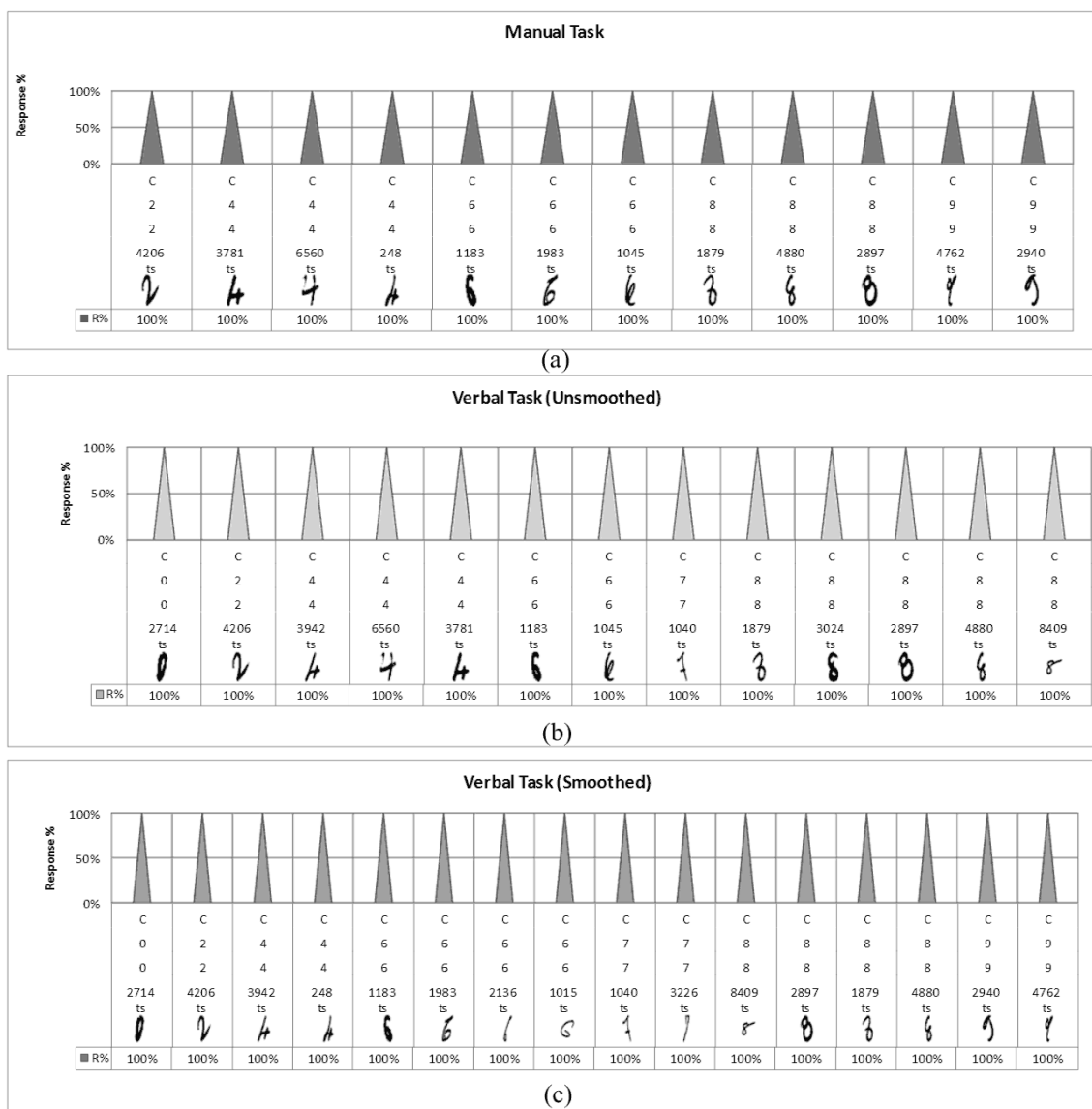


Fig. A27. Most correctly identified irregulars and how they were identified in (a) Manual, (b) unsmoothed conditions of Verbal and (c) smoothed conditions of Verbal. Chart labels read from *bottom to top*: (1) Percentage of a response, (2) digit index in MNIST, (3) correct numeral, (4) numeral in response and (5) whether the response is correct (C). *Top left*, In Manual task, digit 4206ts (testing database) is a 4 and was identified as such in all responses. Data based on: 61 participants in Manual, 42 in unsmoothed and 11 in smoothed.

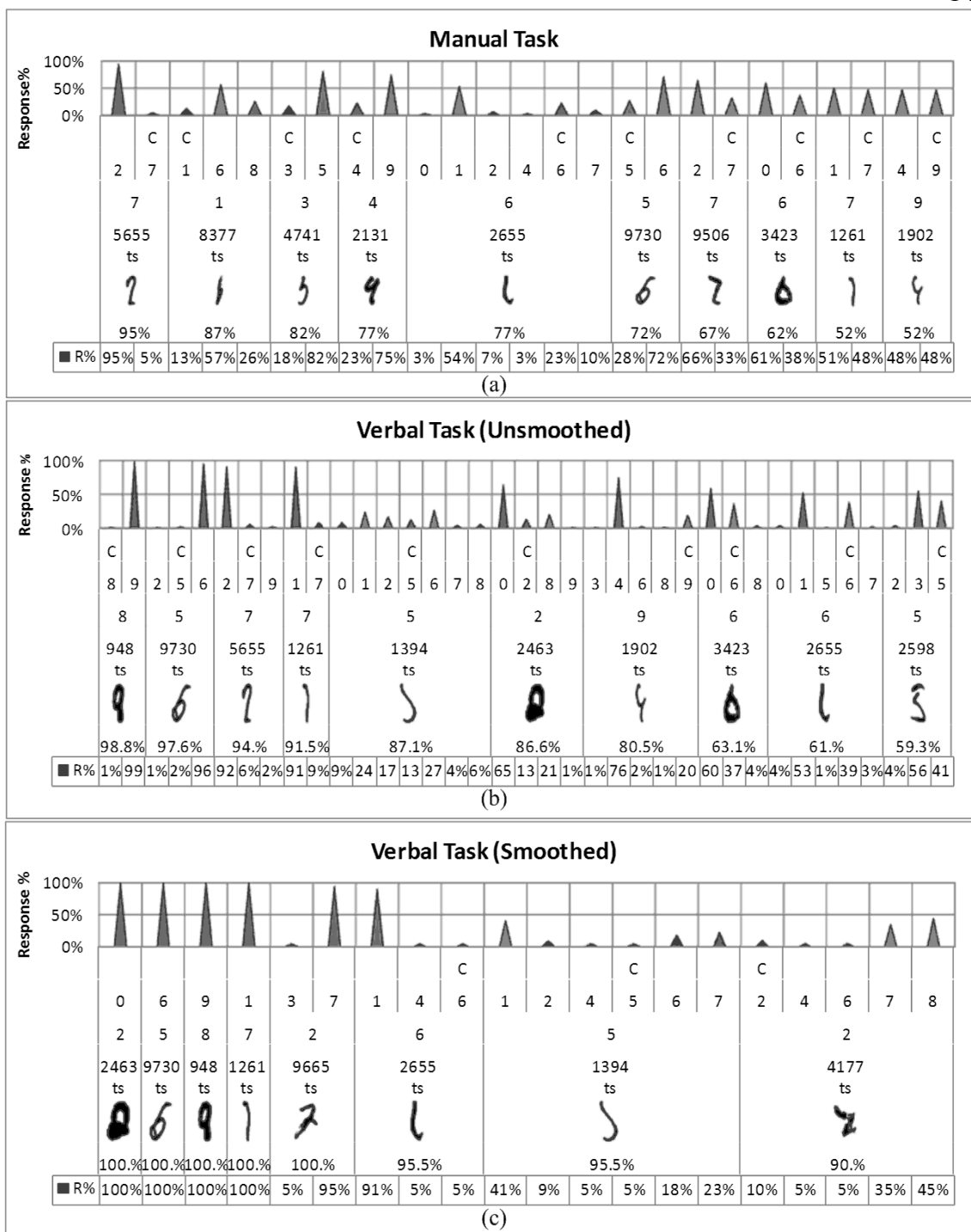


Fig. A28. Most incorrectly identified irregulars and how they were identified in (a) Manual, (b) unsmoothed conditions of Verbal and (c) smoothed conditions of Verbal. Chart labels read from bottom to top: (1) Percentage of each response, (2) percentage of incorrect responses, (3) digit index in MNIST, (4) correct numeral, (5) numeral in each response, and (6) whether a response is correct (C). Top left, In Manual task, digit 5655ts represents a 7 but was mistaken for a 2 in 95% of responses. Responses occurring only once are hidden. Empty responses ignored. Based on: 61 participants (Manual), 42 (unsmoothed) and 11 smoothed.

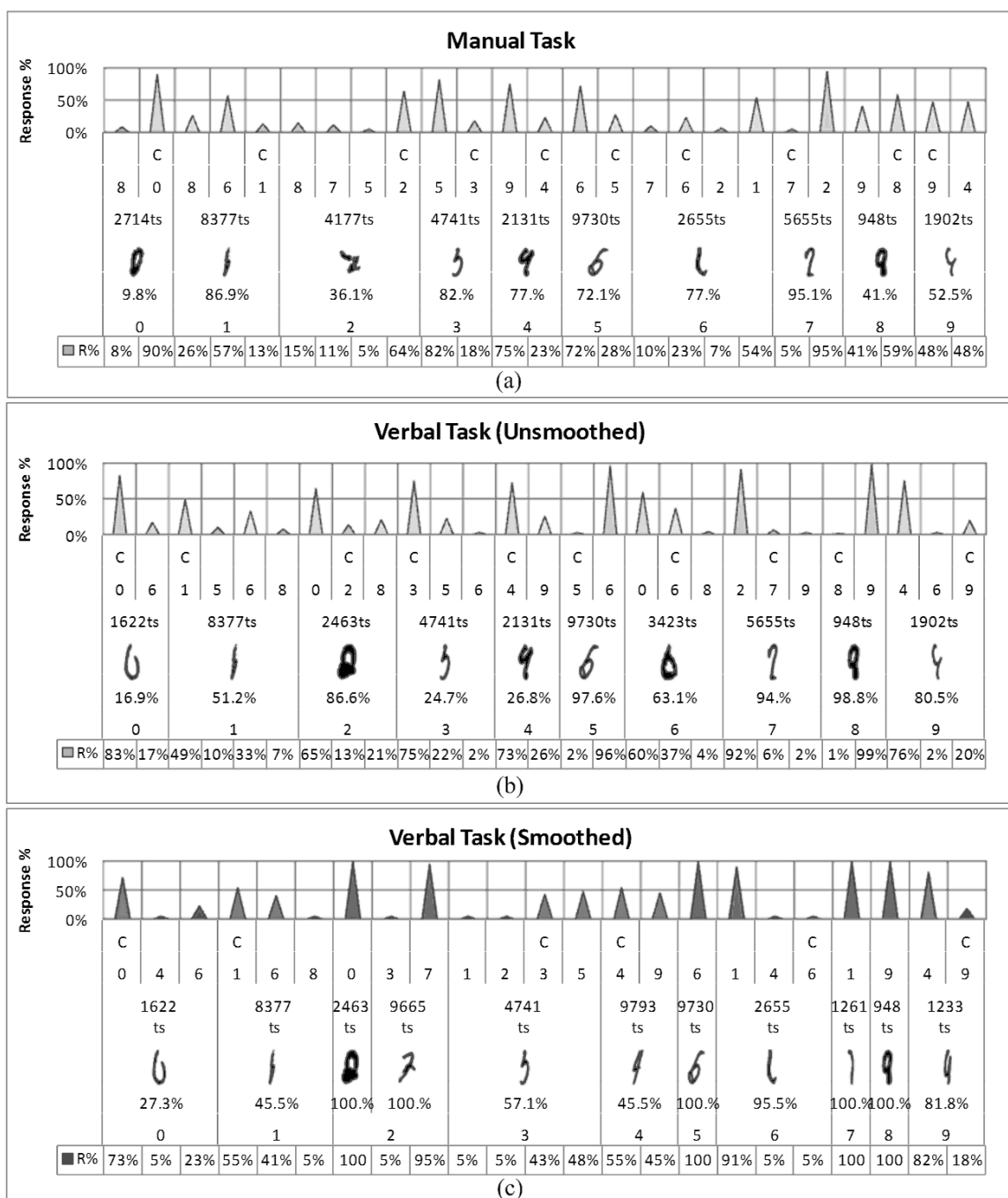


Fig. A29. Most incorrectly identified irregulars by numeral in (a) Manual, (b) unsmoothed conditions of Verbal and (c) smoothed conditions of Verbal. Chart labels read from *bottom to top*: (1) Percentage of each response, (2) correct numeral, (3) percentage of incorrect responses, (4) digit index in MNIST, (5) numeral in each response, and (6) whether a response is correct (C). Responses occurring only once are hidden. Empty responses ignored. *Top left*, In Manual task, digit 2714ts is a 0 and was identified otherwise in 9.8% of responses and as an 8 in 8% of responses. Data based on: 61 participants in Manual, 42 in unsmoothed and 11 in smoothed.

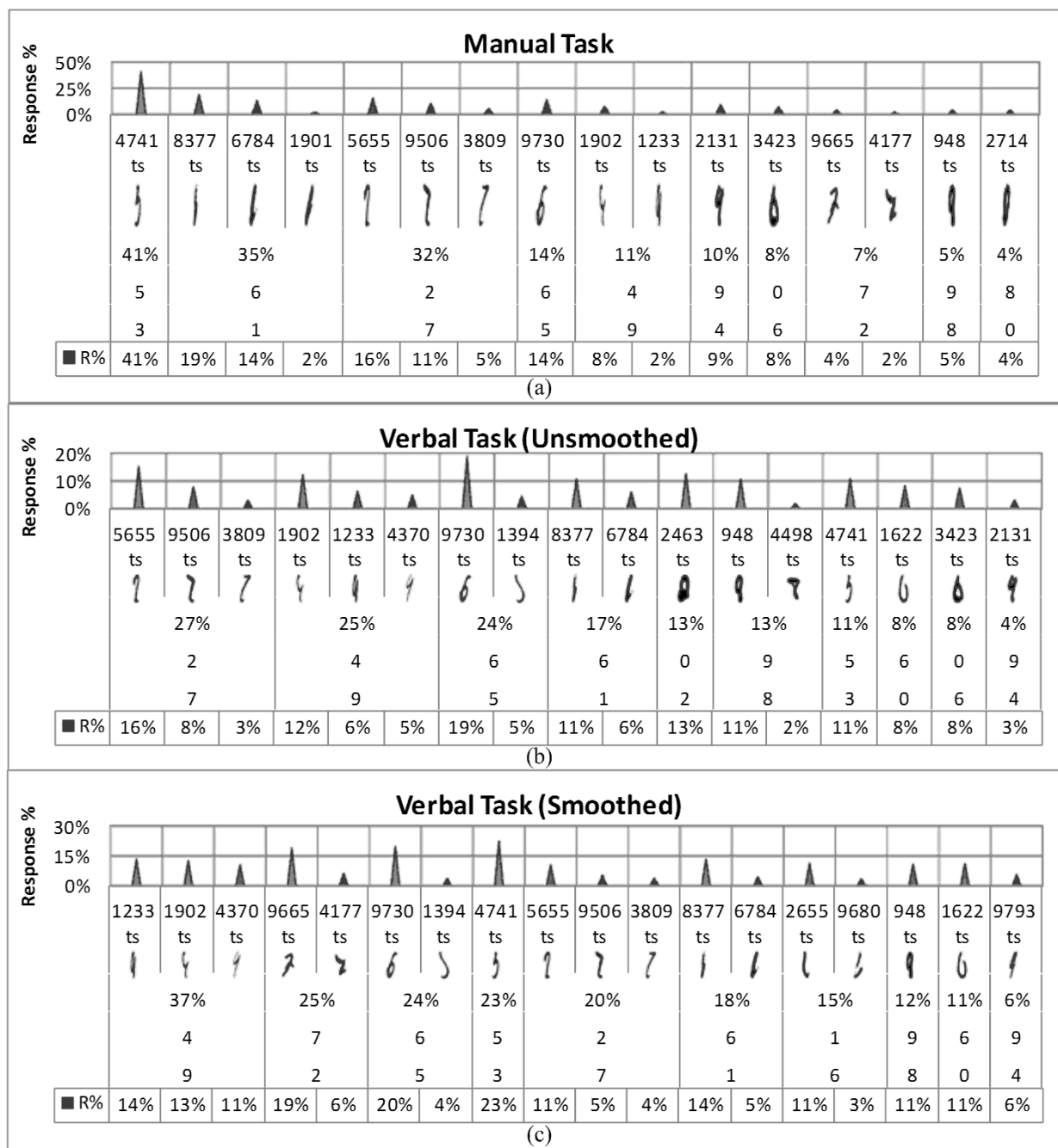


Fig. A30. Most common confusion for each numeral among irregulars in (a) Manual, (b) unsmoothed conditions of Verbal and (c) smoothed conditions of Verbal. Chart labels read from bottom to top: (1) Percentage of each digit-response pair in numeral, (2) correct numeral, (3) most common error in numeral, (4) percentage of most common error in numeral, and (5) digit index in MNIST. Digits accounting for <1% of responses in a given numeral are hidden. Empty responses are ignored. Top left, In Manual task, irregular 3s were most often confused with 5. 3-5 confusion represents 41% of responses in irregular 3s all of which occurred on digit 4741ts. Data based on: 61 participants in Manual, 42 in unsmoothed and 11 in smoothed.

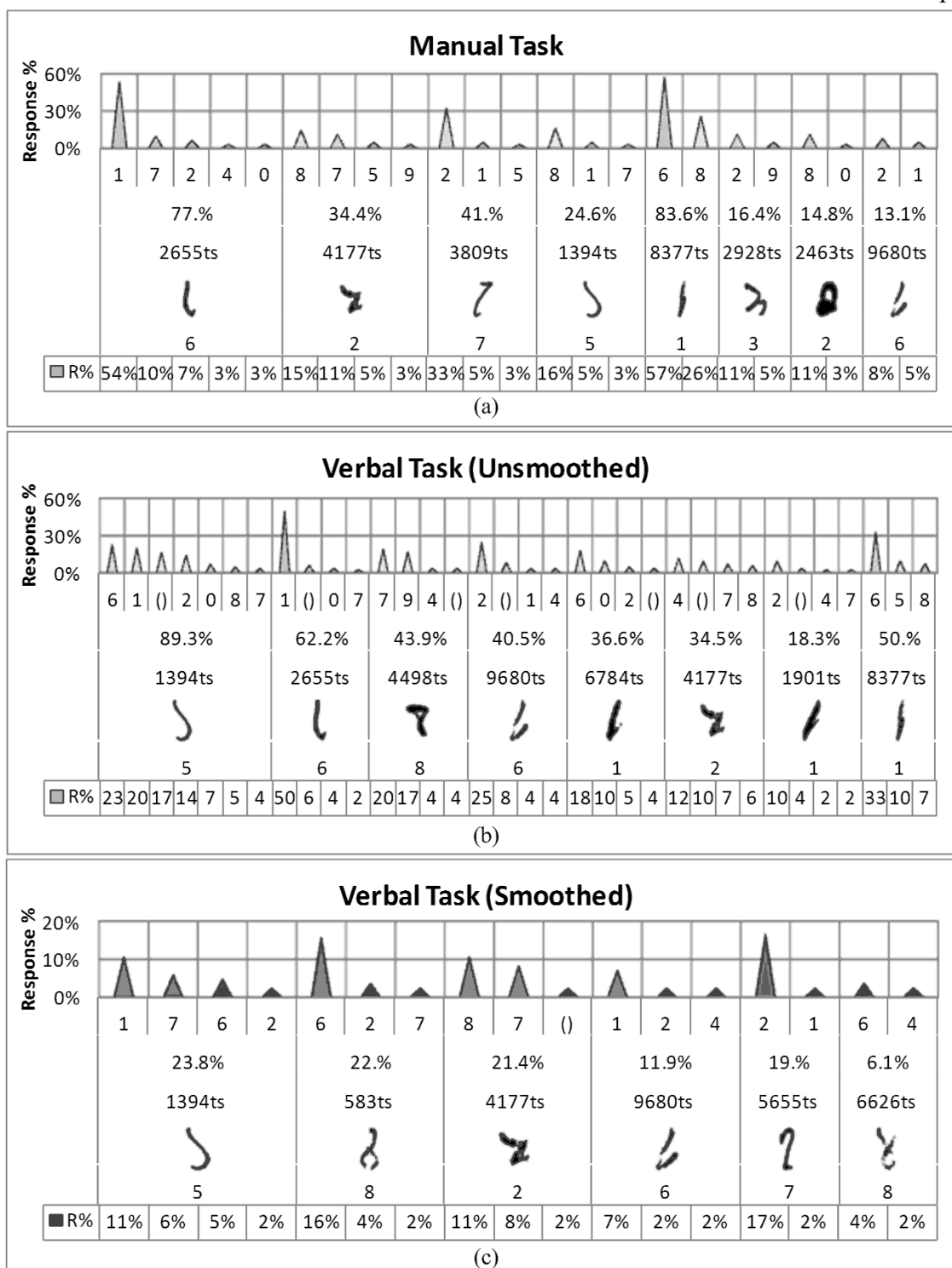


Fig. A31. Irregulars that are confused with the most numerals in (a) Manual, (b) unsmoothed conditions of Verbal and (c) smoothed conditions of Verbal. Chart labels read from *bottom to top*: (1) Percentage of each response, (2) correct numeral, (3) digit index in MNIST, (4) percentage of incorrect responses, and (5) numeral in each response. Responses occurring only once ignored. Empty responses considered. *Top left*, In Manual task, digit 2655ts is confused with five numerals: 1, 7, 2, 4 and 0; the digit is confused in 77% of all responses and 1 represents 54% of responses on this digit. Data based on: 61 participants in Manual, 42 in unsmoothed and 11 in smoothed.

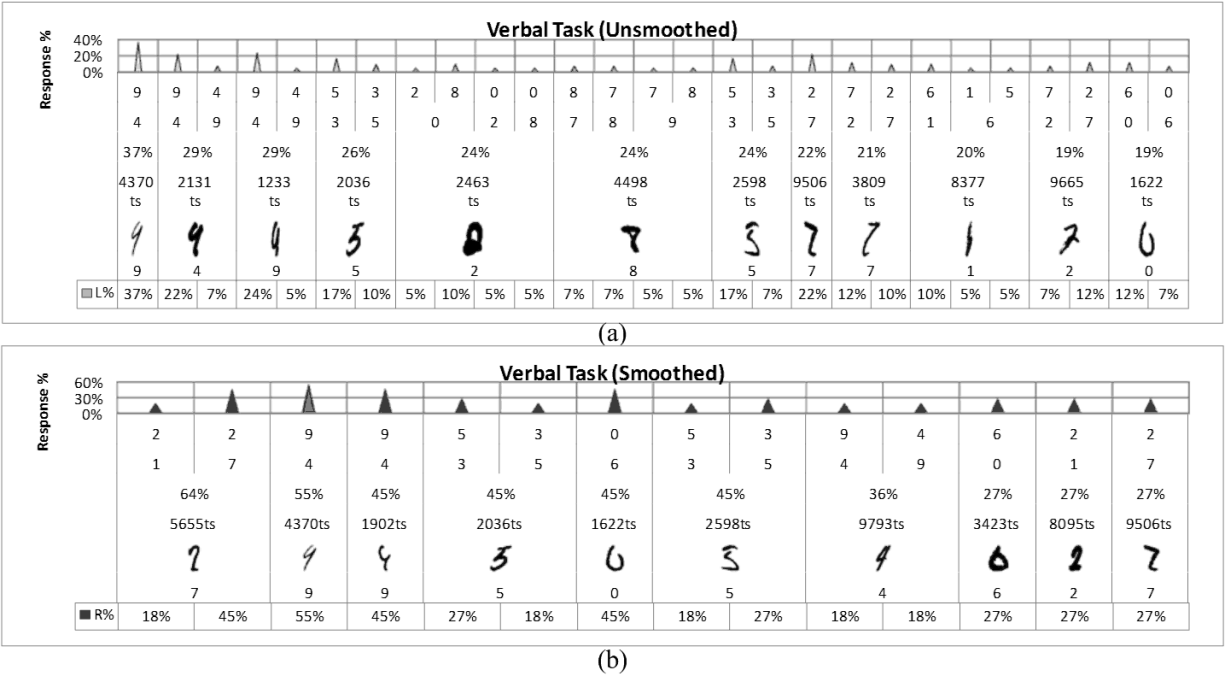


Fig. A32. Irregulars that were re-identified the most in (a) unsmoothed and (b) smoothed conditions of Verbal. Chart labels read from *bottom to top*: (1) Percentage of each response pair, (2) correct numeral, (3) digit index in MNIST, (4) percentage of response pairs, and (5) response pairs. Trials with an empty response are ignored. *Top left*, in unsmoothed conditions of Verbal, digit 4370ts is a 9 and is identified as a 4 then re-identified as 9 in 37% of trials. Data based on: 42 participants in unsmoothed and 11 in smoothed.

Table A1. Irregular digits in ascending order of identifiability in Manual task.

Manual		Digit Identifiability							
MNIST no.									
Numeral		Response							
Total	Incorrect%	Response%							
2	#5655 7 95.1%	2 95%	c 7 5%						
1	#8377 1 86.9%	1 13%	5 2%	6 57%	7 2%	8 26%			
5	#4741 3 82.0%	3 18%	c 5 82%						
6	#2655 6 77.0%	0 3%	1 54%	2 7%	4 3%	c 6 23%	7 10%		
4	#2131 4 77.0%	3 2%	4 23%	9 75%					
8	#9730 5 72.1%	5 28%	c 6 72%						
7	#9506 7 67.2%	2 66%	4 2%	c 7 33%					
6	#3423 6 62.3%	0 61%	1 2%	c 6 38%					
7	#1261 7 52.5%	1 51%	4 2%	c 7 48%					
4	#1902 9 52.5%	3 2%	4 48%	6 2%	7 2%	c 9 48%			
1	#6784 1 47.5%	0 2%	1 52%	2 2%	5 2%	6 41%	9 2%		
7	#3809 7 41.0%	1 5%	2 33%	5 3%	c 7 59%				
9	#948 8 41.0%	8 59%	c 9 41%						
2	#4177 2 36.1%	0 2%	2 64%	5 5%	7 11%	8 15%	9 3%		
5	#1394 5 26.2%	1 5%	2 2%	5 74%	7 3%	c 8 16%			

Table A1-Continued

Manual MNIST no. Numeral	Digit	Identifiability			
Total Incorrect%	Response	Response	Response	Response	Response
	Response%	Response%	Response%	Response%	Response%
7	#9665	c			
	2	2	7		
	21.3%	79%	21%		
3	#2928	c			
	3	2	3	5	9
	18.0%	11%	82%	2%	5%
0	#2463	c			
	2	0	1	2	8
	16.4%	3%	2%	84%	11%
4	#1233	c			
	9	4	9		
	14.8%	15%	85%		
6	#9680	c			
	6	1	2	6	7
	14.8%	5%	8%	85%	2%
5	#2036	c			
	5	3	4	5	
	11.5%	10%	2%	89%	
5	#2598	c			
	5	2	3	5	
	9.8%	2%	8%	90%	
0	#2714	c			
	0	0	2	8	
	9.8%	90%	2%	8%	
1	#1901	c			
	1	0	1	6	7
	9.8%	2%	90%	7%	2%
8	#9531	c			
	9	8	9		
	8.2%	8%	92%		
6	#1015	c			
	6	0	5	6	
	4.9%	2%	3%	95%	
4	#9793	c			
	4	4	7	9	
	4.9%	95%	2%	3%	
8	#3024	c			
	8	1	8	9	
	3.3%	2%	97%	2%	
4	#1550	c			
	4	1	4	6	
	3.3%	2%	97%	2%	
2	#4498	c			
	8	4	7	8	
	3.3%	2%	2%	97%	

Table A1-Continued

Manual MNIST no. Numeral		Digit Identifiability			
Total	Incorrect%	Response	Response%		
3	#675 5 3.3%	3	c 5		
4	#4370 9 3.3%	4	c 9		
1	#1040 7 3.3%	4	c 7	8	2%
7	#3226 7 1.6%	1	c 7		
6	#1622 0 1.6%	0	c 6		
2	#8095 2 1.6%	2	c 5		
8	#6626 8 1.6%	2	c 8		
6	#2136 6 1.6%	4	c 6		
3	#583 8 1.6%	6	c 8		
8	#8409 8 1.6%	5	c 8		
4	#1113 4 1.6%	4	c 6		
4	#3942 4 1.6%	1	c 4		
8	#2897 8 0.0%	8	c		100%
4	#3781 4 0.0%	4	c		100%
6	#1983 6 0.0%	6	c		100%

Table A1-Continued

Manual		Digit Identifiability	
MNIST no.			
Numeral		Response	
Total Incorrect%		Response%	
4	#248 4 0.0%	c 4 100%	
8	#4880 8 0.0%	c 8 100%	
6	#1045 6 0.0%	c 6 100%	
3	#1879 8 0.0%	c 8 100%	
9	#2940 9 0.0%	c 9 100%	
4	#6560 4 0.0%	c 4 100%	
2	#4206 2 0.0%	c 2 100%	
9	#4762 9 0.0%	c 9 100%	
6	#1183 6 0.0%	c 6 100%	

Notes: Data based on 61 participants.

Table A2. Irregular digits in ascending order of identifiability in unsmoothed conditions of Verbal task.

Unsmoothed Verbal		Digit Identifiability							
MNIST no.									
Numeral		Response							
Total	Incorrect%	Response%							
9	#948	c							
	8	9	8						
	98.8%	99%	1%						
6	#9730	c							
	5	6	5	2					
	97.6%	96%	2%	1%					
2	#5655	c							
	7	2	7	9					
	94.8%	92%	6%	2%					
7	#1261	c							
	7	1	7						
	91.5%	91%	9%						
5	#1394	c							
	5	6	1	()	2	5	0	8	
	89.3%	23%	20%	17%	14%	11%	7%	5%	
0	#2463	c							
	2	0	8	2	9				
	86.6%	65%	21%	13%	1%				
4	#1902	c							
	9	4	9	()	6	8	3		
	81.8%	74%	19%	2%	2%	1%	1%		
6	#2655	c							
	6	1	6	()	0	7	5		
	63.4%	50%	37%	6%	4%	2%	1%		
0	#3423	c							
	6	0	6	8					
	63.1%	60%	37%	4%					
3	#2598	c							
	5	3	5	2	()				
	59.8%	55%	40%	4%	1%				
1	#8377	c							
	1	1	6	5	8	0			
	51.2%	49%	33%	10%	7%	1%			
2	#9506	c							
	7	7	2						
	48.8%	51%	49%						
8	#4498	c							
	8	8	7	9	()	4	0		
	45.1%	55%	20%	17%	4%	4%	1%		
6	#9680	c							
	6	6	2	()	1	4	0	7	
	42.9%	57%	25%	8%	4%	4%	1%	1%	
2	#9665	c							
	2	2	7	6					
	40.5%	60%	39%	1%					

Table A2-Continued

Unsmoothed Verbal		Digit Identifiability							
MNIST no.		Response							
Numeral		Response%							
Total	Incorrect%								
4	#1233 9 38.1%	c 9 62%	4 38%						
5	#2036 5 36.9%	c 5 63%	3 35%	6 2%					
6	#6784 1 36.6%	c 1 63%	6 18%	0 10%	2 5%	() 4%			
7	#4177 2 35.7%	c 2 64%	4 12%	() 10%	7 7%	8 6%	1 1%		
8	#583 8 31.7%	c 8 68%	6 20%	2 7%	() 4%	3 1%			
9	#4370 9 30.5%	c 9 70%	4 30%						
4	#2131 4 26.8%	c 4 73%	9 26%	2 1%					
5	#4741 3 25.6%	c 3 74%	5 22%	6 2%	() 1%				
5	#675 5 20.7%	c 5 79%	3 21%						
1	#1901 1 20.7%	c 1 79%	2 10%	() 4%	7 2%	4 2%	6 1%	0 1%	
7	#3809 7 20.2%	c 7 80%	2 18%	1 2%					
6	#1622 0 17.9%	c 0 82%	6 17%	() 1%					
3	#2928 3 7.3%	c 3 93%	() 2%	2 2%	7 1%	4 1%			
9	#4762 9 6.0%	c 9 94%	4 6%						
2	#8095 2 2.4%	c 2 98%	9 1%	1 1%					

Table A2-Continued

Unsmoothed Verbal		Digit Identifiability	
MNIST no.		Response	
Numeral	Total Incorrect%	Response	Response%
4	#9793	c	
	4	4	9
	4.8%	95%	5%
6	#1983	c	
	6	6	2
	2.4%	98%	2%
5	#1015	c	
	6	6	3
	1.2%	99%	1%
4	#248	c	
	4	4	2
	1.2%	99%	1%
4	#1550	c	
	4	4	()
	1.2%	99%	1%
8	#6626	c	
	8	8	6
	1.2%	99%	1%
7	#3226	c	
	7	7	1
	1.2%	99%	1%
4	#1113	c	
	4	4	6
	1.2%	99%	1%
9	#2940	c	
	9	9	2
	1.2%	99%	1%
6	#2136	c	
	6	6	1
	1.2%	99%	1%
8	#9531	c	
	9	9	8
	1.2%	99%	1%
6	#1045	c	
	6	6	
	0.0%	100%	
8	#4880	c	
	8	8	
	0.0%	100%	
7	#1040	c	
	7	7	
	0.0%	100%	
4	#3781	c	
	4	4	
	0.0%	100%	

Table A2-Continued

Unsmoothed Verbal		Digit Identifiability	
MNIST no.		Response	
Numeral		Response%	
Total	Incorrect%		
8	#8409	c	
	8	8	
	0.0%	100%	
0	#2714	c	
	0	0	
	0.0%	100%	
2	#4206	c	
	2	2	
	0.0%	100%	
8	#1879	c	
	8	8	
	0.0%	100%	
6	#1183	c	
	6	6	
	0.0%	100%	
8	#3024	c	
	8	8	
	0.0%	100%	
8	#2897	c	
	8	8	
	0.0%	100%	
4	#3942	c	
	4	4	
	0.0%	100%	
4	#6560	c	
	4	4	
	0.0%	100%	

Notes: Data based on 42 participants. *Top left*, digit #8409 is an 8 and was identified as such (C) in all responses.

Table A3. Irregular digits in ascending order of identifiability in smoothed conditions of Verbal task.

Smoothed Verbal MNIST no.		Digit Identifiability						
Numeral		Response						
Total	Incorrect%	Response%						
	#9730							
6	5	6						
	100.0%	100%						
	#2463							
0	2	0						
	100.0%	100%						
	#9665							
7	2	7	3					
	100.0%	95%	5%					
	#1261							
7	7	1						
	100.0%	100%						
	#948							
8	8	9						
	100.0%	100%						
	#1394							
5	5	1	7	6	2	5	4	
	95.5%	41%	23%	18%	9%	5%	5%	
	#2655							
6	6	1	6	4				
	95.5%	91%	5%	5%				
	#4177							
2	2	8	7	()	2	4	6	
	90.9%	41%	32%	9%	9%	5%	5%	
	#583							
2	8	6	2	8	7	()		
	86.4%	59%	14%	14%	9%	5%		
	#1233							
4	9	4	9					
	81.8%	82%	18%					
	#1902							
4	9	4	9					
	77.3%	77%	23%					
	#5655							
2	7	2	7	1				
	72.7%	64%	27%	9%				
	#4370							
4	9	4	9					
	63.6%	64%	36%					
	#4741							
5	3	5	3	1	()	2		
	59.1%	45%	41%	5%	5%	5%		
	#3423							
6	6	0	6					
	54.5%	55%	45%					

Table A3-Continued

Smoothed Verbal		Digit Identifiability				
MNIST no.						
Numeral		Response				
Total Incorrect%		Response%				
6	#9680	c				
	6	6	1	4	2	3
	50.0%	50%	27%	9%	9%	5%
1	#8377	c				
	1	1	6	8		
	45.5%	55%	41%	5%		
4	#9793	c				
	4	4	9			
	45.5%	55%	45%			
7	#9506	c				
	7	7	2	8		
	36.4%	64%	32%	5%		
8	#4498	c				
	8	8	7	9		
	31.8%	68%	27%	5%		
8	#6626	c				
	8	8	6	4	()	9
	31.8%	68%	14%	9%	5%	5%
5	#2036	c				
	5	5	3			
	31.8%	68%	32%			
5	#2598	c				
	5	5	3			
	31.8%	68%	32%			
0	#1622	c				
	0	0	6	4		
	27.3%	73%	23%	5%		
3	#2928	c				
	3	3	()	8	7	
	27.3%	73%	14%	9%	5%	
7	#3809	c				
	7	7	2			
	22.7%	77%	23%			
2	#8095	c				
	2	2	1			
	18.2%	82%	18%			
9	#9531	c				
	9	9	8			
	13.6%	86%	14%			
1	#6784	c				
	1	1	6			
	13.6%	86%	14%			
5	#675	c				
	5	5	3			
	13.6%	86%	14%			

Table A3-Continued

Smoothed Verbal		Digit Identifiability	
MNIST no.		Response	
Numeral	Total Incorrect%	Response	Response%
1	#1901	c	
	1	1	4
	9.1%	91%	9%
6	#1045	c	
	6	6	()
	4.5%	95%	5%
4	#1550	c	
	4	4	8
	4.5%	95%	5%
9	#2131	c	
	4	4	9
	4.5%	95%	5%
4	#1113	c	
	4	4	0
	4.5%	95%	5%
8	#3024	c	
	8	8	7
	4.5%	95%	5%
4	#6560	c	
	4	4	7
	4.5%	95%	5%
4	#3781	c	
	4	4	5
	4.5%	95%	5%
0	#2897	c	
	8	8	
	0.0%	100%	
6	#1183	c	
	6	6	
	0.0%	100%	
7	#3226	c	
	7	7	
	0.0%	100%	
8	#8409	c	
	8	8	
	0.0%	100%	
8	#1879	c	
	8	8	
	0.0%	100%	
6	#1983	c	
	6	6	
	0.0%	100%	
4	#248	c	
	4	4	
	0.0%	100%	

Table A3-Continued

Smoothed Verbal MNIST no.	Digit	Identifiability
Numeral	Response	
Total Incorrect%	Response%	
6	#2136	c
	6	6
	0.0%	100%
8	#4880	c
	8	8
	0.0%	100%
9	#2940	c
	9	9
	0.0%	100%
0	#2714	c
	0	0
	0.0%	100%
4	#3942	c
	4	4
	0.0%	100%
6	#1015	c
	6	6
	0.0%	100%
9	#4762	c
	9	9
	0.0%	100%
7	#1040	c
	7	7
	0.0%	100%
2	#4206	c
	2	2
	0.0%	100%

Notes: Data based on 11 participants. *Top left*, digit #2136 is a 6 and was correctly identified in all responses.

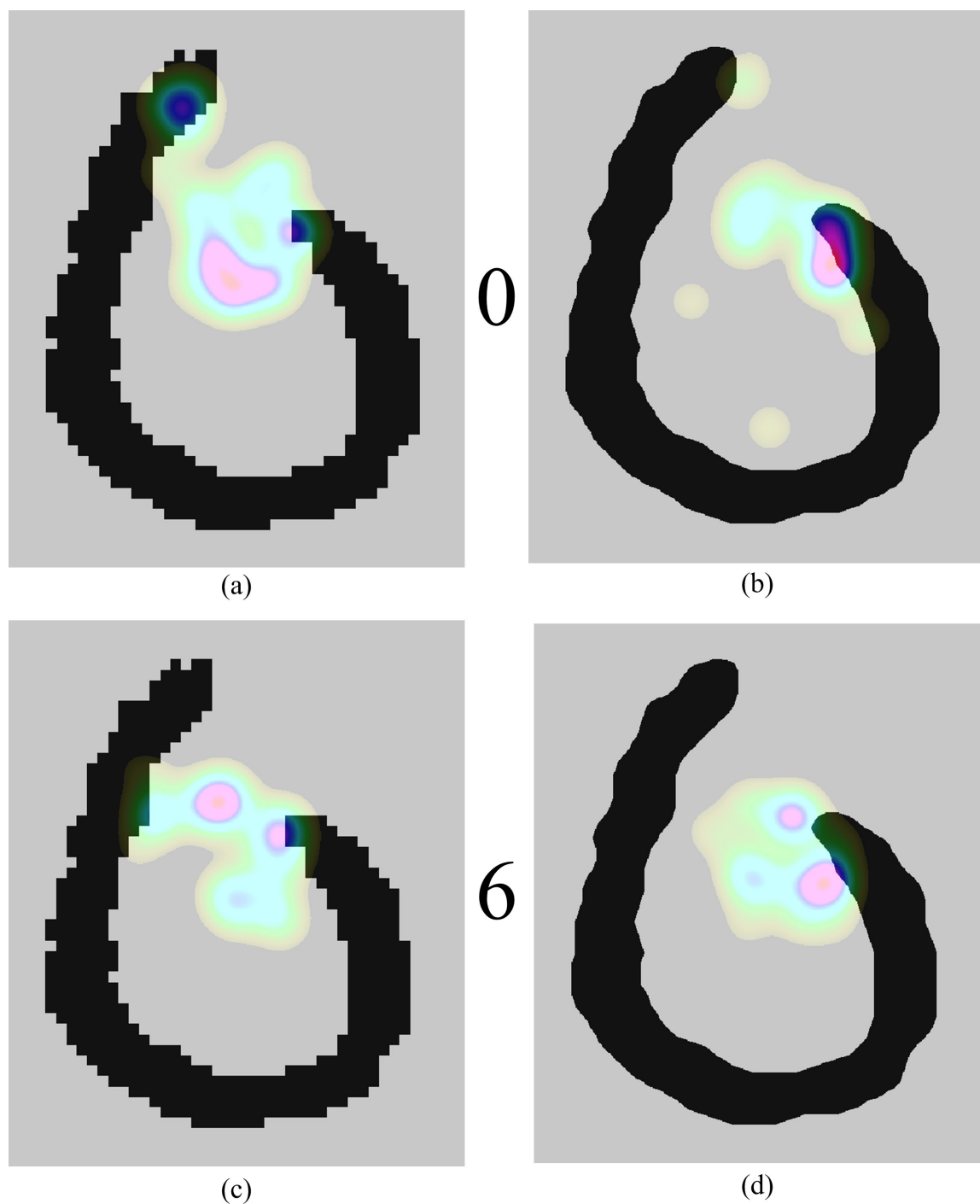


Fig. A33. Four duration-based heat maps showing gaze density during 1-second period preceding verbal identification of NIST image 0-#010030 (MNIST no. 1622) as 0(*top*) and 6(*bottom*). Heat maps were generated based on: (a) 17 fixations in 6 trials, (b) 16 fixations in 5 trials, (c) 21 fixations in 8 trials, and (d) 16 fixations in 5 trials.

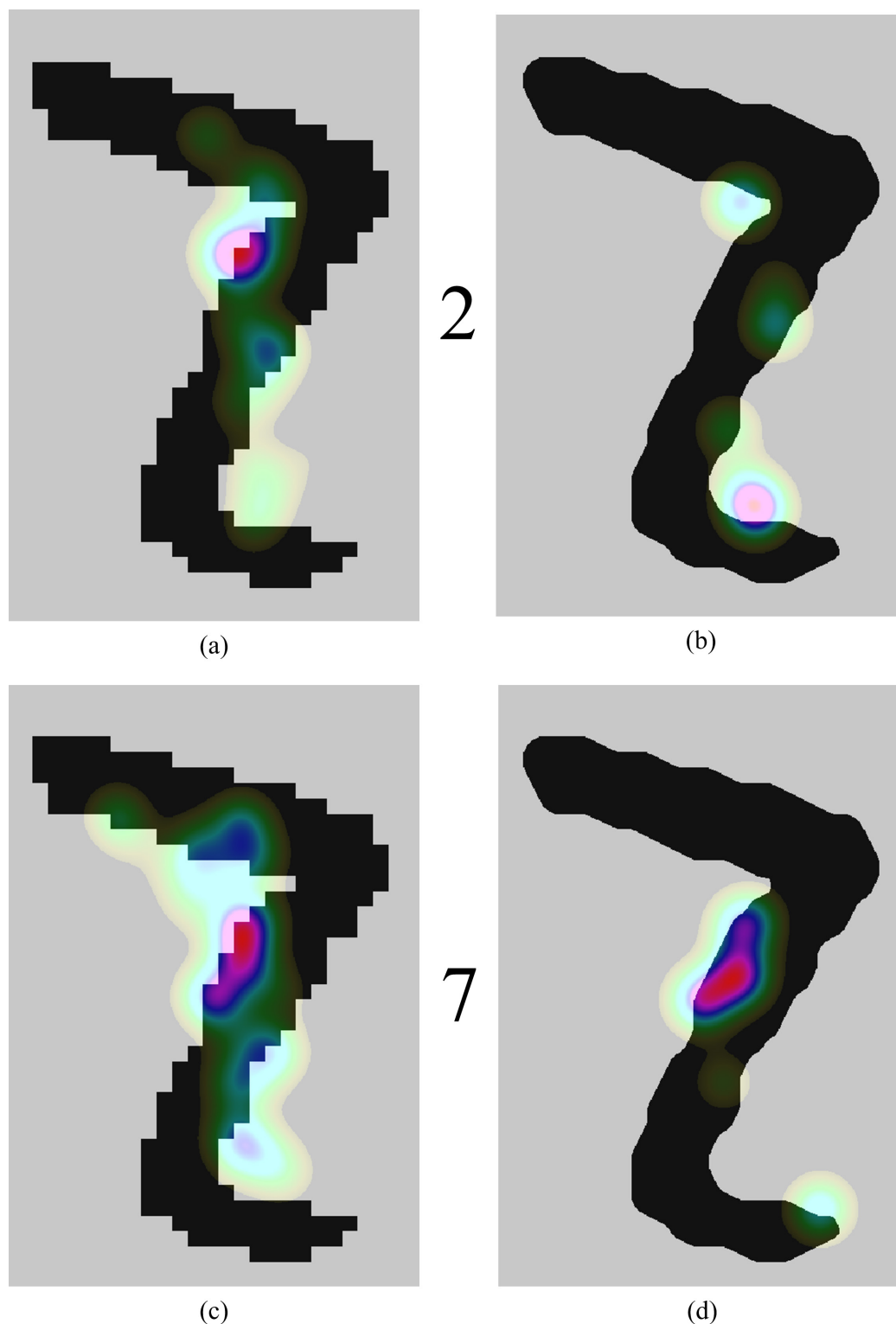


Fig. A34. Four duration-based heat maps showing gaze density during 1-second period preceding verbal identification of NIST image 7-#320036 (MNIST #9506) as 2(*top*) and 7(*bottom*). Heat maps were generated based on: (a) 25 fixations in 10 trials, (b) 6 fixations in 3 trials, (c) 26 fixations in 10 trials, and (d) 7 fixations in 3 trials.

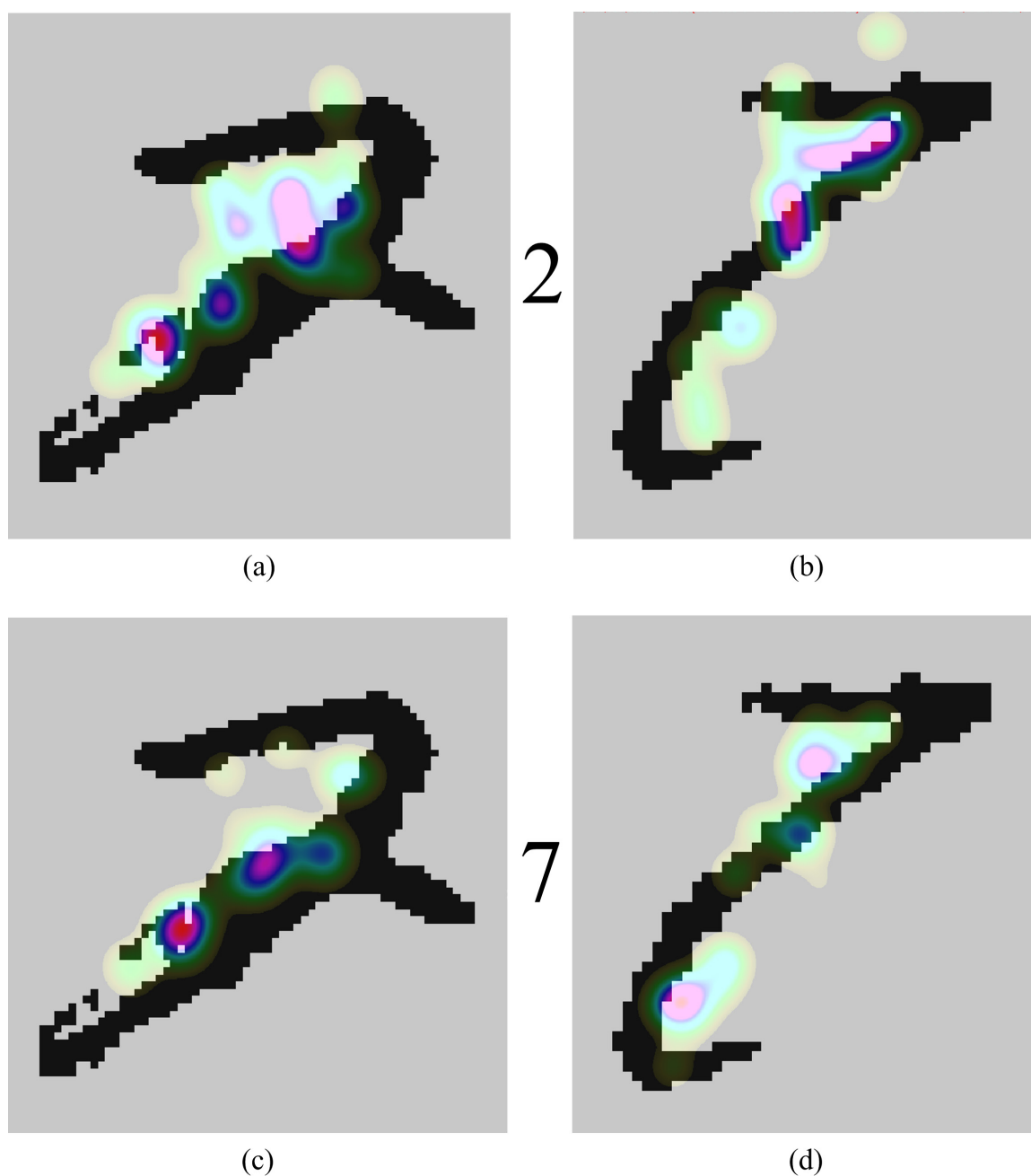


Fig. A35. Four duration-based heat maps showing gaze density during 1-second period preceding verbal identification of NIST images (a) 2-#067322 and (b) 7-#026696 (MNIST #9665 and #3809) as 2(*top*) and 7(*bottom*). Heat maps were generated based on: (a) 26 fixations in 8 trials, (b) 22 fixations in 9 trials, (c) 23 fixations in 8 trials, and (d) 23 fixations in 9 trials.

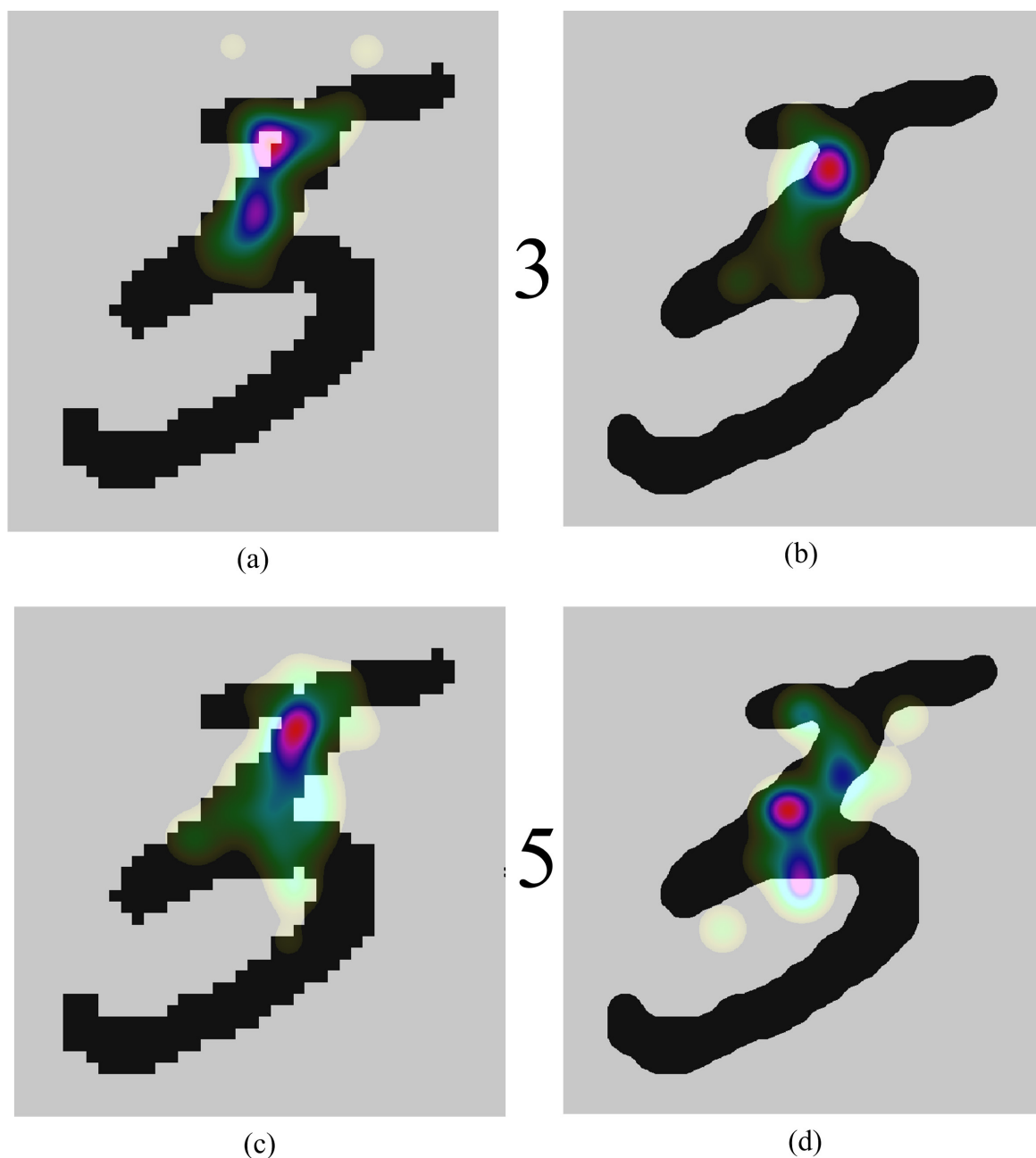


Fig. A36. Four duration-based heat maps showing gaze density during 1-second period preceding verbal identification of NIST image 5-055771 (MNIST #2036) as 3(*top*) and 5(*bottom*). Heat maps were generated based on: (a) 27 fixations in 11 trials, (b) 14 fixations in 5 trials, (c) 32 fixations in 11 trials, and (d) 15 fixations in 5 trials.

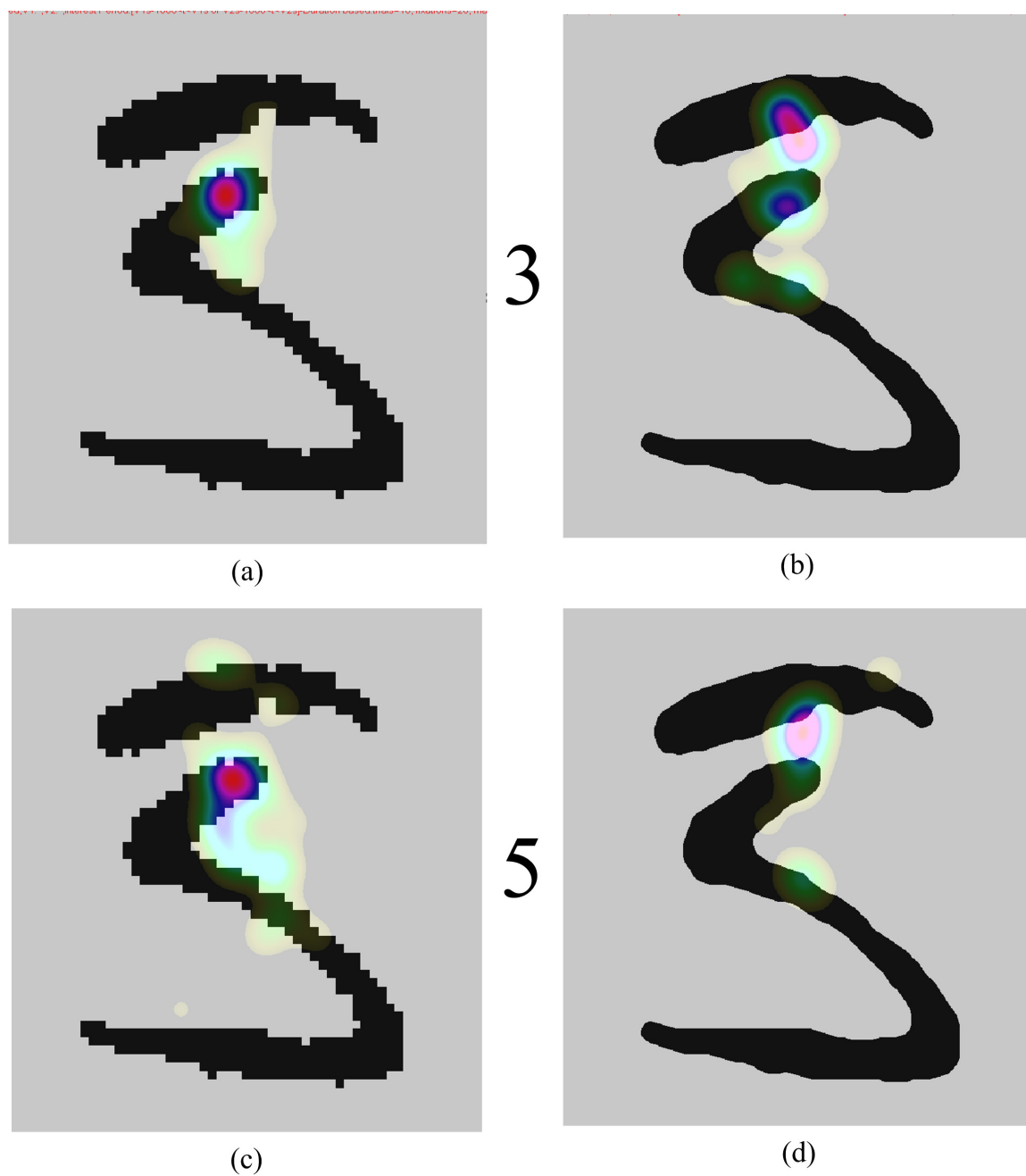


Fig. A37. Four duration-based heat maps showing gaze density during 1-second period preceding verbal identification of NIST image 5-#032654 (MNIST #2598) as 3(*top*) and 5(*bottom*). Heat maps were generated based on: (a) 26 fixations in 10 trials, (b) 11 fixations in 5 trials, (c) 25 fixations in 10 trials, and (d) 13 fixations in 5 trials.

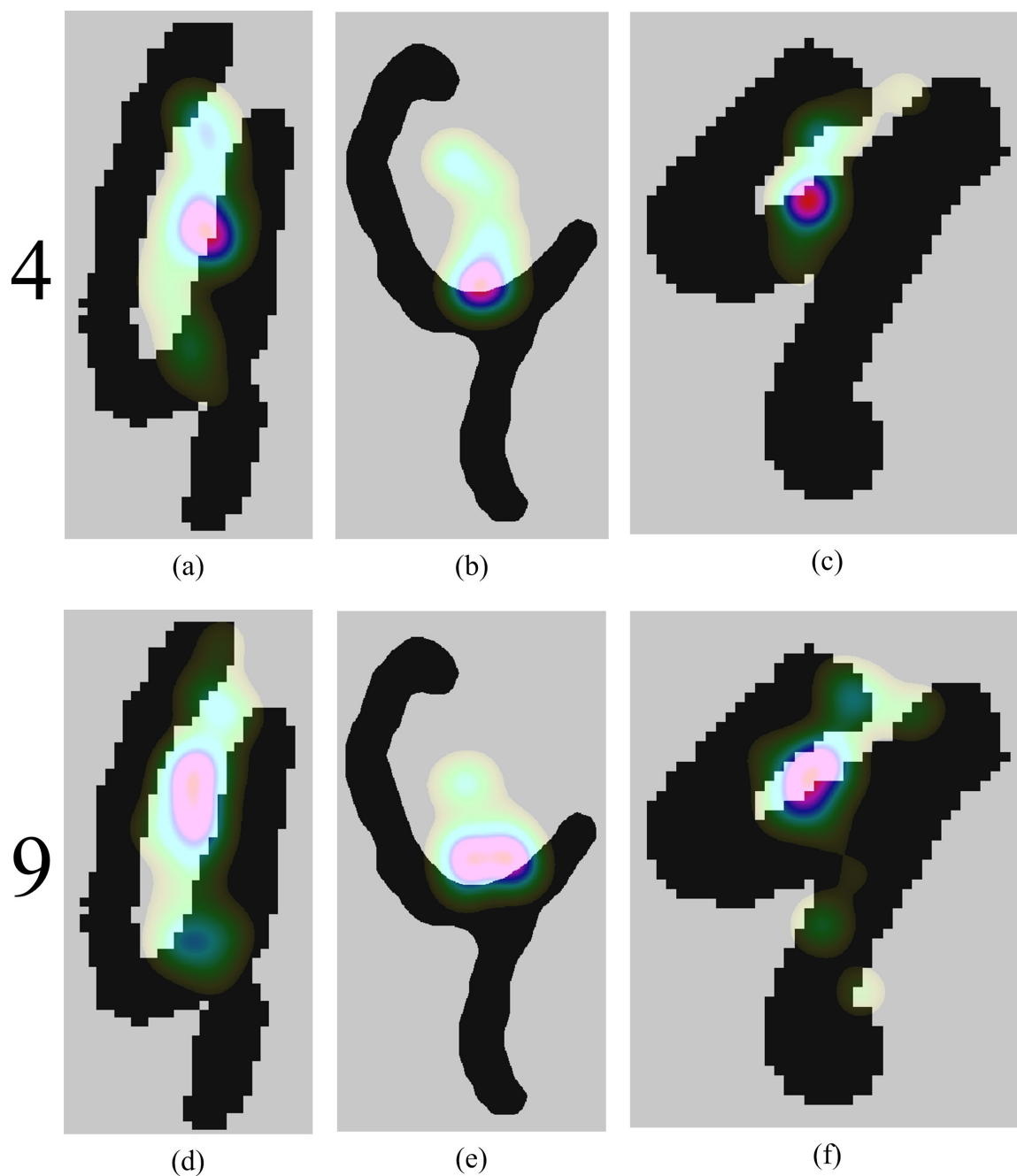


Fig. A38. Six duration-based heat maps showing gaze density during 1-second period preceding verbal identification of NIST images 9-#025422, 9-#021018 and 4-#035302 (MNIST #1233, #1902 and #2131) as 4 (top) and 9 (bottom). Heat maps were generated based on: (a) 26 fixations in 12 trials, (b) 10 fixations in 5 trials, (c) 26 fixations in 12 trials, (d) 27 fixations in 12 trials, (e) 9 fixations in 5 trials, and (f) 36 fixations in 12 trials.

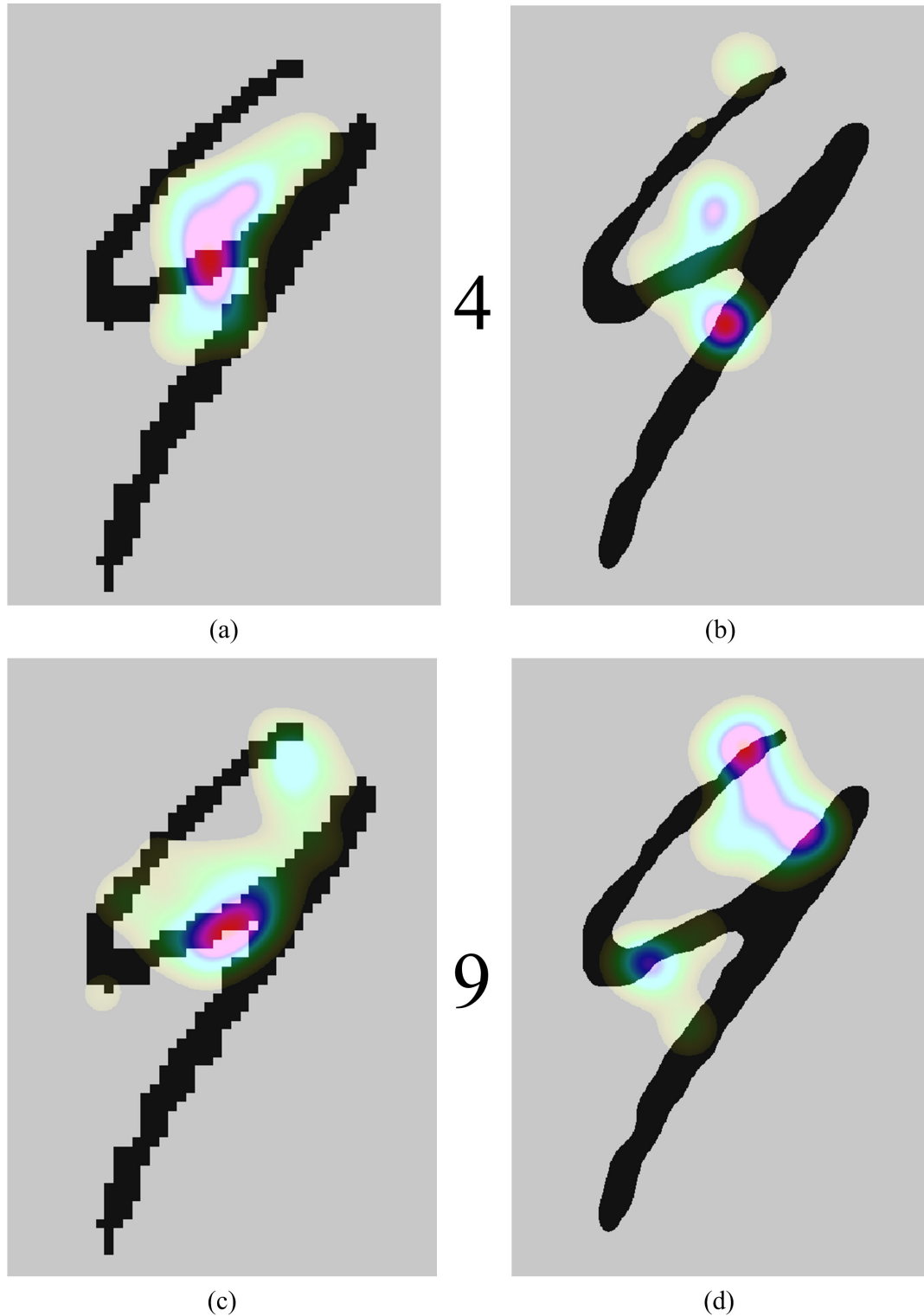


Fig. A39. Four duration-based heat maps showing gaze density during 1-second period preceding verbal identification of NIST image 9-#014734 (MNIST #4370) as 4 (*top*) and 9 (*bottom*). Heat maps were generated based on: (a) 40 fixations in 15 trials, (b) 16 fixations in 6 trials, (c) 41 fixations in 15 trials, and (d) 14 fixations in 6 trials.

GLOSSARY

aliasing. See spatial aliasing.

answer. See verbal answer.

anti-aliasing. Refers to the process of minimizing the perceived effect of distortions and artefacts associated with aliasing. This can be achieved by using a low-pass filter to remove image components with spatial frequencies that are too high to be sampled accurately.

ASC. Refers to a file extension used to store eye movement data after its conversion into ASCII-plain text format.

BG. Background (of an image).

binary image. A type of image where pixel values are set to black or white.

bottom-up or BU. Refers to bottom-up approaches in image analysis and pattern recognition whereby low-level features like colour, contrast and orientation guide the analysis or task execution.

bottom-up visual saliency. Refers to a number of image processing or mathematical techniques used to calculate an importance map of a visual medium based on low-level features such as colour, contrast and orientation.

bounding box. The smallest rectangular enclosure within which all points of a given image element can fit.

bV1. Occurring before first verbal response.

bV2. Occurring before second verbal response.

C or c. Correct numeral; response with correct numeral.

confusion pair. Refers to an ordered pair of numerals: the first, represented by a set of handwritten digits, and the second being the most common identification response error on those digits by a group of participants.

correct numeral or correct numeral label. The Arabic numeral that a handwritten digit

actually represents; refers to the label of a handwritten digit as specified in the MNIST database.

Data Viewer. An SR Research program to visualize and analyse EDF files.

digit no. Refers to the corresponding MNIST sequence number of the handwritten digit in question regardless of whether the digit was obtained from the MNIST database or the NIST database.

double-prompt identification. Refers to prompting the participant to call out the displayed handwritten digit twice during a Verbal task trial.

EDF. Binary eye movement data file generated by experiments created using Experiment Builder.

empty or missing response. Refers to failure of participant to provide a response within an expected period. In Verbal task, an empty response occurs every time a participant fails to identify a handwritten digit after each prompt in a given trial.

ESACF. Enhanced summary autocorrelation function.

Experiment Builder. A graphical environment made by SR Research to create experiments that run on the eye-tracking hardware used in this study.

eye movement. Voluntary or involuntary movement of the eyes to acquire, fixate and track visual stimuli or parts thereof. It includes a wide range of parameters relating to changes in eye gaze and orientation and is typically made up of an alternating sequence of saccades and visual fixations.

eye tracker. An eye-tracking device.

eye tracking. The process of measuring eye gaze and orientation typically using the pupil's infra-red reflection.

FG. Foreground (of an image)

FG0. Verbal task viewing condition where the foreground-handwritten digit colour is RGB(0,0,0) and the background colour is set to RGB(240,240,240).

FG120. Verbal task viewing condition where the foreground-handwritten digit colour is RGB(120,120,120) and the background colour is set to RGB(240,240,240).

FG180. Verbal task viewing condition where the foreground-handwritten digit colour is RGB(180,180,180) and the background colour is set to RGB(240,240,240).

FG210. Verbal task viewing condition where the foreground-handwritten digit colour is

RGB(210,210,210) and the background colour is set to RGB(240,240,240).

FG228. Verbal task viewing condition where the foreground-handwritten digit colour is RGB(228,228,228) and the background colour is set to RGB(240,240,240).

fixate. To carry out one or more visual fixations.

fixation. See visual fixation.

foveate. To maintain visual gaze on an object or feature such that its image is projected onto the foveal region – the region with the highest visual acuity – of the retina.

gaze span or gaze spatial span. Pixels of a visual stimulus spanned by a set of fixations; may also refer to an approximation of this given by the size of the bounding box of the fixations heat map relative to the size of the bounding box of the displayed handwritten digit as follows:

$$\text{Gaze Spatial Span} = \frac{\text{Heatmap Boundingbox Area}}{\text{Digit Boundingbox Area}}$$

greyscale image. A type of image where pixel values are set to a shade of grey

handwritten digit. An instance of an Arabic numeral written by hand.

heat map, fixation map or fixation heat map. Colour-coded probability density map representing the average share of gaze that various parts of a given image receive; may also refer to this overlaid on top of the corresponding visual stimulus.

Identifiability. Refers to the degree to which a handwritten digit is identifiable usually expressed in terms of percentage of correct identification.

identifiable or most identifiable. Refers to a handwritten digit that was correctly identified by participants during an identification condition or task.

identification rate or correct identification rate. Refers to the ratio of count of correct responses to the total count of all responses by a given set of participants on a given set of handwritten digits in the Manual task or the Verbal task.

identification response. Refers to participant response (manual or verbal) specifying the numeral that a handwritten digit represents in the Manual task or the Verbal task.

jaggies. Aliasing artefacts occurring in raster images whereby smooth lines look rough, pixelated or stair-like.

low-level feature. See bottom-up.

luminance contrast. See Michelson contrast ratio.

macro. A scripting language-based program typically used to automate a set of operations inside another software program that would otherwise involve a significant amount of manual work.

main experiment. Refers to experiment design used for main data collection following a smaller-scale pilot experiment.

Manual *or* Manual task. Refers to our digit identification experiment whereby participants are instructed to identify an MNIST handwritten digit using the numeral keys of a keyboard.

Michelson contrast ratio. The ratio of the spread to the sum of two luminance values of interest like the luminance of text foreground and background:

$$C = (Lum_{BG} - Lum_{FG}) / (Lum_{BG} + Lum_{FG})$$

misidentification. Refers to the incorrect identification of a stimulus in an identification task.

MNIST. Modified NIST.

MNIST database. A database of handwritten digit images derived from the larger NIST database.

NIST. National Institute of Standards and Technology.

NIST database. Refers to a large set of handwritten digit images found in one of the databases maintained by the National Institute of Standards and Technology.

non-empty response. See empty response.

numeral. One of ten Arabic numerals: 0, 1, 2, 3, 4, 5, 6, 7, 8 and 9.

Optimal Viewing *or* Optimal Viewing guidelines. Here refers to guidelines that, when used to display a class of visual stimuli, yield eye movement data that are most suitable for pattern recognition applications.

pattern classifier. A machine learning agent or software that can assign a classification or label to a given input value.

pilot, pilot experiment *or* pilot testing. A smaller-scale preliminary experiment intended to uncover design flows and refine experiment parameters prior to the main experiment.

PNG. Portable Network Graphics; an image file format.

raster image. A graphical representation using a rectangular grid of pixels and pixel values to store and display images; bitmap image.

re-identification. Refers to the state – or an instance thereof – of a participant identifying a handwritten digit as one numeral then identifying it as a different numeral during the same trial in the Verbal task.

response rate. Refers to the average number of responses per prompt in Verbal task where a participant is prompted for a response twice in each trial.

Results File. A plain text output file generated by an executable created using SR Research Experiment Builder. It may contain select experiment data like key presses and trial-specific parameters.

ROI. Region of interest (of gaze).

saccade. A fast shift in eye gaze position.

smoothed. Refers to images used in some viewing conditions of the Verbal task whereby a special image processing filter is used to eliminate jaggies and spatial aliasing along the contours of the original *unsmoothed* NIST handwritten digit image.

spatial aliasing. Refers to distortions or artefacts that result when a visual signal reconstructed from discrete samples is different from the continuous signal originally sampled.

string editing. Refers to the use of character strings in information theory to represent and compare two sets of data.

subsampling *or* downsampling. Refers to the process of reducing the number of samples in an image producing a modified image with less details and lower resolution.

task session. Refers to a single participation in one of the two identification tasks.

top-down (TD). Refers to top-down approaches in image analysis and pattern recognition whereby high-level functions like context awareness and task familiarity guide the analysis or task execution.

top-down selective attention. Refers to the use of high-level functions like context awareness and task familiarity to determine image parts to receive closer analysis.

trial *or* experiment trial. A single experiment unit such as the identification of a single handwritten digit in an experiment using seventy-four such digits.

true numeral label. See correct numeral.

tr. Training database; a suffix that refers to handwritten digits from NIST or MNIST training databases.

ts. Testing database; a suffix that refers to handwritten digits from the NIST or MNIST testing databases.

Unsmoothed. See smoothed

V1. First verbal response or occurring during first verbal response.

V2. Second verbal response or occurring during second verbal response.

V2+. Occurring after second verbal response.

VAD. Voice activity detection.

VBA. Visual Basic for Applications. A scripting language to program MS Excel macros

Verbal or Verbal task. Refers to our eye-tracking experiment whereby participants are instructed to identify a NIST handwritten digit verbally into a microphone while their visual fixations are being recorded.

verbal answer. Refers to the actual numeral given in a verbal response of Verbal task.

Verbal task viewing condition. Refers to one of seven display conditions in the Verbal task whereby an image of a handwritten digit covers the entire display and its luminance contrast and aliasing may be manipulated.

Vision Science. The study of visual perception and the visual system from perspectives ranging from cognitive psychology and neuroscience to computer science and psychophysics.

visual fixation. Refers to the maintaining of visual gaze – or an instance thereof – on a single location such that its image is projected onto the foveal region – the region with the highest visual acuity – of the retina.

XLS. Standard file extension for Microsoft Excel spreadsheet documents.

XLSM. File extension for Microsoft Excel macro-enabled workbooks.