A study of human recognition rates for foveola-sized image patches selected from initial and final fixations on calibrated natural images

Ian van der Linde[†], Umesh Rajashekar[‡], Lawrence K. Cormack[*], Alan C. Bovik[‡]
[†] Department of Computing, APU, Chelmsford Essex CM1 1JJ (UK) {i.v.d.linde@apu.ac.uk}
[‡] Department of Electrical & Computer Engineering, University of Texas at Austin (USA)
[*] Department of Psychology, University of Texas at Austin (USA)

ABSTRACT

Recent years have seen a resurgent interest in eye movements during natural scene viewing. Aspects of eye movements that are driven by low-level image properties are of particular interest due to their applicability to biologically motivated artificial vision and surveillance systems. In this paper, we report an experiment in which we recorded observers' eye movements while they viewed calibrated greyscale images of natural scenes. Immediately after viewing each image, observers were shown a test patch and asked to indicate if they thought it was part of the image they had just seen. The test patch was either randomly selected from a different image from the same database or, unbeknownst to the observer, selected from either the first or last location fixated on the image just viewed. We find that several low-level image properties differed significantly relative to the observers' ability to successfully designate each patch. We also find that the differences between patch statistics for first and last fixations are small compared to the differences between hit and miss responses. The goal of the paper was to, in a non-cognitive natural setting, measure the image properties that facilitate visual memory, additionally observing the role that temporal location (first or last fixation) of the test patch played. We propose that a memorability map of a complex natural scene may be constructed to represent the low-level memorability of local regions in a similar fashion to the familiar saliency map, which records bottom-up fixation attractors.

Keywords: eye-tracking, natural image statistics, visual memory.

1. INTRODUCTION

Spatial contrast sensitivity varies considerably across the human field of view, being high only at the fovea and dropping rapidly as the visual periphery is approached. This variation in sensitivity occurs as a result of the changing density, type, and interconnection of retinal cells, and necessitates ballistic eye movements to bring areas of interest into incidence with the fovea sequentially as visual attention is shifted. Saccades are made approximately 3 times every second[15] to re-orient the high-acuity fovea at the central retina onto regions of interest in the field of fixation. During the fixations between saccades, eye orientation is relatively static for a period of 200-300ms, and detailed information from the scene may be garnered. Saccades do not rely on continuous visual feedback, but are programmed during the final 150-175ms of the previous fixation using some selection criteria to determine the most salient target[13, 15]. Internal muscular feedback conveying the eye-globe orientation is the trigger to halt the oculomotor system when the saccade has oriented the visual axis at/near the desired target, with saccades rarely being pre-empted in mid-flight with little or no new visual information being acquired during the saccade, a phenomenon known as saccadic suppression[22].

By analysing image characteristics surrounding fixation coordinates, researchers have postulated some locally conspicuous features that attracted the observer to program a saccade, and have demonstrated that fixation patches bear significantly different properties when compared with patches selected at random from the same image[16], including the observation that the measured conspicuity of these fixation patches may deplete over time as progressively less attractive regions are regarded[13]. A matrix of scalar values representing the conspicuity of each image location, commonly referred to as a saliency map, may be constructed. This may be used to predict where an observer will most likely place fixations on a given input image. This has applications for machine vision, auto-foveated video compression[5], and to improve our ecological understanding of the human visual system[19].

Three primary forms of visual memory have been proposed to exist to support the human visual system (HVS):

1. Iconic Memory;
2. Visual Short Term Memory (VSTM);
3. Visual Long Term Memory (VLTM).

Of these, only VSTM and VLTM are believed to survive eye-movements. VSTM (operating for periods of several seconds) is considered to possess a finite capacity of several objects[7]. VLTM (operating for periods longer than VSTM, and having an enormous capacity) has been proposed to function by coherent representation of objects rather than a conjunction of parts[29], involving some abstraction from the sensory stimulus.

Experiments have shown that both fixated and non-fixated objects may be recalled in memory tasks[11], but irrespective of the memory model employed, researchers have found that information at the point of fixation is preferentially retained compared with image regions that have only been sensed non-foveally[6], despite our ability to attend to one area of the visual field while fixating at another (so-called covert attention[25]). Existing work has measured the ability of observers to memorise objects in complex scenes, to distinguish differently configured gratings, or their ability to memorise symbols or objects in controlled synthetic environments[8, 9, 11]. This paper adds to existing work by investigating visual memory performance using natural scenes containing few semantically interesting features. It is theorised that the HVS is specifically optimised for processing natural images rather than synthetically generated stimuli[17]. The set of natural images is a tiny subset of all possible images exhibiting some interesting properties, such as their scale invariance, preponderance of horizontal and vertical edges, and smooth harmonics. Natural images are preferable for the study of pre-attentive and low-level/bottom-up visual phenomena, having been the causal stimuli driving the evolution of the HVS, and since experimentation with natural stimuli can confirm or refute results obtained with synthetic stimuli. This paper presents an investigation of visual memory that combines eye tracking with a simple visual memory task, and applies a set of metrics to sets of natural images patches that observers were/were not able to recall.

Existing studies have proposed which image dimensions (such as orientation, spatial frequency, and hue) are retained by visual memory (for a review see Magnussen, 2000 [28]), but used experimental techniques relying on synthetic stimuli. These psychophysical methods typically involve asking the participant to make a forced choice between a reference stimulus and a previously displayed stimulus or stimulus sequence. In this paper we indicate which image dimensions assisted visual memory when viewing static natural images, finding that our results agree with earlier work that observed the limited ability of human observers to robustly encode contrast [9,28]. We also suggest additional dimensions that appeared to be robustly encoded, without the use of cognitive stimuli that have been represented abstractly in VLTM, using image patches with negligible holistic structure or semantic interest.

2. METHOD

2.1 Natural Image Stimuli

101 static diurnal images were manually selected from a calibrated greyscale natural image database[21], finding and omitting images containing man-made structures and features such as animals, faces, and other items of high-level semantic interest that would have instinctively attracted attention[3, 24]. Images whose luminance statistics suggested saturation of the capture device, and thus exhibited non-linearity, were also omitted. Typical images contained natural habitats of trees, grasses, and water. Examples (with overlaid fixations for a single observer) are shown in Fig. 1. The images from the database have dimensions 1536×1024. A central region of 1024×768 was cropped from these images for the experiment, avoiding the need to degrade the images through down-sampling to make them fit within the desired screen configuration of 1 pixel per arc minute. Image linearity was maintained, but brightness increased such that the brightest point in the image corresponded to the brightest output level of the monitor. The 101 images satisfying the selection criteria were verified by a second reviewer prior to use to ensure the complete absence of cognitive/man-made features, with replacements for rejected images obtained from the same database where necessary.

2.2 Instrumentation & Configuration

An SRI/Fourward Generation V Dual Purkinje eye tracker (Fourward Technologies Inc., Buena Vista, VA) was used to gather observer's fixation coordinates while they viewed the stimuli, configured at sampling rate of 200Hz with a

National Instruments data acquisition unit (National Instruments Corp., Austin, TX). Bite-bars to keep the maxilla stationary were created for each human observer using a standard dental compound adhered to a mandible sized aluminium frame. Two forehead rests were employed to provide tactile feedback to discourage fore-aft head movements. Monocular eye tracking was used to reduce calibration time, and the observers wore an occluder on the unmeasured eye. Observers viewed the images on an Image Systems 21" greyscale monitor (Image Systems Corp. Minnetonka, MN). Monitor output accuracy was measured with a photometer and stimuli appropriately gamma corrected. The ambient illumination in the experiment room was kept constant for all observers, with a minimum of 5 minutes dark adaptation provided while the eye-tracker was calibrated. Images were displayed using a Matrox Parahelia graphics card (Matrox Graphics Inc., Dorval, Québec, Canada) at a screen resolution of 1024×768 pixels, providing a DPI of 64 horizontally and vertically, and a refresh rate of 60Hz. The screen was placed 134cm from the observer and subtended a visual angle of $16^{o}\times12^{o}$, giving approximately 1 minute of arc per screen pixel. Image patches harvested for subsequent use in the visual memory task were set to 64×64 pixels, slightly over 1º. The patch size used corresponds to the size of the high-acuity cone-dominated foveola at the centre of the human fovea.

2.3 Visual Memory Task

The main experiment, consisting of 101 trials per observer, was preceded by a dry-run session of 10 trials to ensure that the observer became familiar with the handheld control box, dark adapted, and comfortable in the experimental environment prior to data collection. Images for the dry-run session were selected from the same database as the images used for the experiment proper. Stimuli were presented using software written in Matlab 6.5 (Mathworks Inc., Nantick, MA) using the Psychophysics Toolbox extensions[2, 14]. Images were shown in a fixed order for all observers. The task required the observer to view a small (64×64, foveola-sized) image patch after each full-screen 5 *s* stimulus image and respond via the control box to indicate whether they believed the patch was present in the image they had just viewed. Observers were advised that they should free-view the images as they desired (typical scan-paths are shown in Fig. 1).

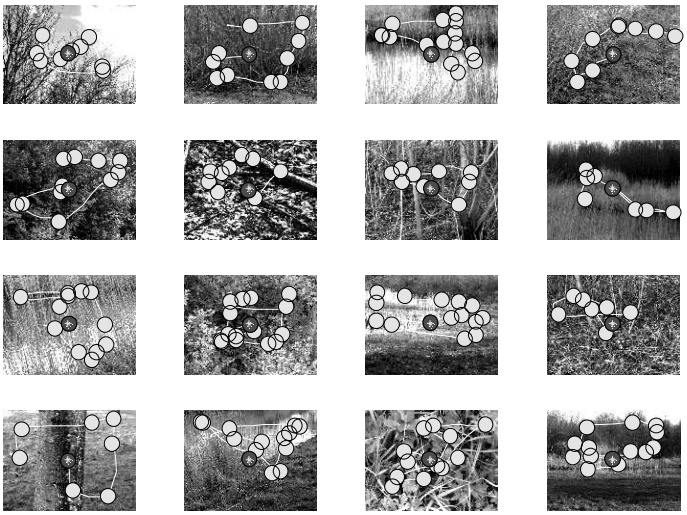

Fig. 1. Example scan-paths. Light circles are fixations, dark circles are the disregarded central start positions, and crosses inside the dark circles are the true screen centres.

With the experimental setup, a display time of 5 *s* only allowed the observers to make a small number of fixations (on average, 10 per observer per image). The probability of the patch having been part of the stimulus image was 50%, and immediate auditory feedback (*via* a sampled voice) was provided to indicate a correct or incorrect response. Further to this, if the patch was part of the stimulus image, it had a 50% probability of being taken directly from the observer's first fixation and 50% probability of being the observer's final fixation, calculated from the sample data on the fly using an

adaptation of the ASL 3-boundary fixation detection algorithm[1]. Since patches had a 50% probability of being real or decoy, the possibility of learned probability matching was removed. To remove the possibility of luminance matching, the patches were brightness jittered by a random amount up or down. In order to preserve structure, the jitter magnitude was randomly selected from a range limited such that it did not cause the binning of values from the extremes of the intensity histogram. The observers were made aware of the brightness jittering during the dry run session. Additionally, before each full stimulus image a Gaussian noise image was displayed to help remove after-images corresponding to the previous image that may otherwise have attracted fixations.

Each session was preceded with a 9-point calibration routine, in which voltages were recorded and interpolated to screen coordinates on a 3×3 grid. This calibration routine was repeated compulsorily every 10 images, and a calibration test run after every image. This was achieved by introducing the requirement that the observer fixate for 500ms within a 5 *s* time limit on a central dot prior to progressing to the next image in the stimulus collection. If the observer was unable to satisfy this test, the full calibration procedure was re-run. The average number of calibrations per observer for the 101 images was 16.5, i.e. between 6 and 7 images were typically viewed before the calibration test was failed. Average calibration error for passed calibration tests was 5.48 pixels horizontally and vertically, thus the maximal displacement of patches harvested for analysis was typically under ±10% of the total patch size in each dimension. The requirement for a central fixation prior to displaying the next image also ensured that all observers commenced viewing the image stimuli from the same location, facilitating a concurrent study of fixation clustering (documented in a forthcoming paper). The average duration for the experiment was approximately 1 hour, including calibrations, display of images and task completion. Observers who became uncomfortable during the experiment were allowed to disengage from the eye tracker for a break of any duration they desired. Post-experimental feedback collection revealed that most observers rated the eye-tracker as only mildly uncomfortable. Plotting the mean performance of the observers over time does not suggest a prevailing fatigue factor, with performance sustained up until the last trial (Fig. 2).

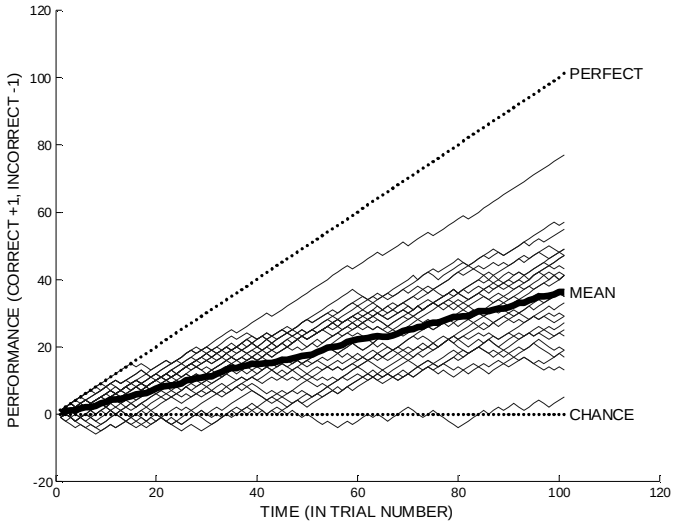

Fig. 2. Graph of observer performance over time. Each correct response gained +1 and each incorrect response –1.

A total of 29 human volunteers participated in the study; 19 male and 11 female, with an age range between 16 and 54 and a mean age of 27. All observers reported emmetropic vision or were myopic or hyperopic only, which were appropriately corrected, a selection criterion introduced to minimise calibration difficulties resulting from complex visual impairments. Correction was verified with a Snellen eye chart displayed on the eye-tracking screen used to display the stimulus images. Observers consisted of members of the public, undergraduates, graduates students, research fellows, alumni and faculty from the University of Texas at Austin from a range of academic disciplines. Each observer visited for a single session, only 2 had seen the image stimuli previously, and 24 were naïve as to the purpose of the experiment. All participants were unpaid. Data showed no significant differences for the performance of naïve *vs.* non-naïve participants. All participants provided written consent in accordance with University regulations, and the experiment was conducted in accordance with University ethical regulations.

A post-experiment evaluation and question session revealed that none of the naïve observers were cognisant of the fact that the patches they were asked to designate were their own first or last fixations, removing the possibility of cheating by consciously electing to keep the eyes still. This was verified manually by plotting and reviewing observer's scan-paths.

3. RESULTS

Results are shown in Table 1. Columns for hit (real patch, yes response), miss (real patch, no response), correct rejection (decoy patch, no response), and false alarm (decoy patch, yes response) were prepared. Additionally, hit and miss responses were split into those relating to patches originating from first fixations, and those originating from last fixations. A representation of average values was prepared in the form of statistics for patches centred at all fixations (on average, 10 per image per observer), and for a set of foveola-sized non-overlapping tiles (NOTs) across each of the stimulus images. Since each image had a resolution of 1024×768, 192 NOTs at 64×64 pixels were harvested for analysis. Each set of patches was subjected to 8 image statistics. It should be noted that the image statistics were calculated for the original patches and not the brightness-jittered patches the observers were actually exposed to.

		PATCH SET TYPE								
		FIX		HIT		MISS			CR	FA
		All F1..Fn	Real NOT	F_Co	L_Co	F_In	L_In	Rand NOT	R_Co	R_In
PATCH STATISTIC	Fixations within 1 Foveola	5.8578		9.7579	7.774	9.3621	7.0393			
	Entropy (Log energy)	34792	33945	35427	35288	33199	33079	28639	28858	29889
	Michelson Contrast	0.7547	0.7241	0.7658	0.751	0.7641	0.7542	0.8556	0.8879	0.8779
	RMS Contrast	0.3517	0.3285	0.3637	0.3573	0.3624	0.3385	0.552	0.6053	0.5559
	Mean Luminance	74.2267	75.7063	79.5364	78.6839	62.1284	60.8265	47.475	43.8077	48.0478
	Mean Square Error	7435.8	7698.7	8400.7	8293.8	5446	5052.9	4167.1	3582.1	3870.7
	Number of Edges	593.9648	556.2342	620.685	636.95	562.852	508.073	478.5589	478.516	524.446
	Unique Colours	85.1956	81.917	92.5569	88.9748	75.6099	75.9805	104.8797	110.422	108.123
	Standard Deviation	24.3919	23.2541	26.6882	25.6314	21.1696	19.7811	23.6902	24.7897	24.1179
	No. of Patches in Set	30810	19392	501	436	223	205	7680	914	448

Table 1. Task results and patch statistics relating to each patch set type. Non-Overlapping Tile [NOT], First Fixation Patch Correct [F_Co], First Fixation Patch Incorrect [F_In], Last Fixation Patch Correct [L_Co], Last Fixation Patch Incorrect [L_In], Random Patch Correct [R_Co], Random Patch Incorrect [R_In].

Overall numbers of correct and incorrect responses are represented in Fig.3, also showing the number of responses falling into each of the 4 categories (hit, miss, FA, CR). Results show a 68% correct response rate composed almost equally of hits and CRs. Incorrect responses are also almost equally composed of misses and FAs.

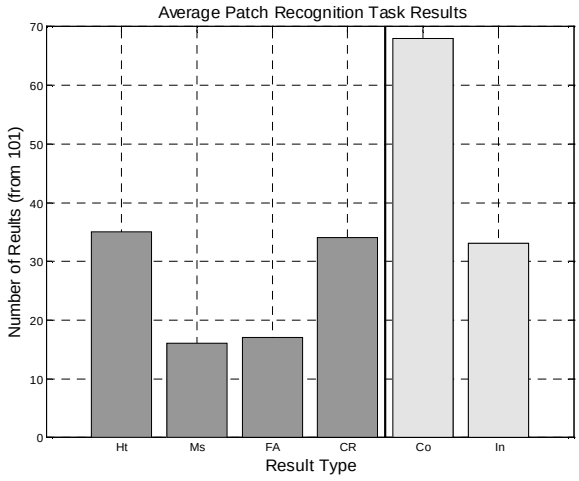

Fig. 3. Bar chart summary of all observers showing Hit [Ht], Miss [Ms], False Alarm [FA], Correct Rejection [CR], and total Correct [Co] and Incorrect [In] results.

Fig. 4 illustrates how correct and incorrect responses were divided up into first fixation, last fixation, and random patches. Observers were close in their ability to respond to trials using patches from the first and last fixations, with a slightly higher success rate for first fixations of around 10%.

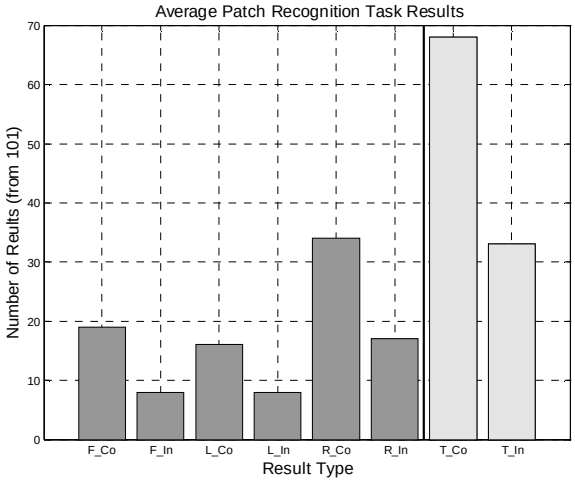

Fig. 4. Bar chart summary of all observers showing First Fixation Patch Correct [F_Co], First Fixation Patch Incorrect [F_In], Last Fixation Patch Correct [L_Co], Last Fixation Patch Incorrect [L_In], Random Patch Correct [R_Co], Random Patch Incorrect [R_In], and Total Correct [T_Co] and Incorrect [T_In] results.

It was noted that the dwell-time for correct responses (121.88ms) was nearly 10% longer than the average incorrect response dwell-time (109.04ms), but that this was the case for only 17 of the 29 observers (59%) and is not considered to be a primary factor in performance.

It was further noted that the average number of fixations for correct and incorrect trials showed minor difference. As one might expect, trials where more fixations were made were more prone to memory task error than those with fewer fixations since the number of patches the observer was exposed to would be larger, though this effect was small: for correct responses an average of 11.98 fixations per observer per image was noted, and for incorrect responses 12.49 fixations per observer per image. This trend was seen for 79% of observers (23 of 29). Individual results giving dwell-time and number of fixations for correct and incorrect trials are provided in Table 2.

To verify that task performance was not linked to how interesting or varied each of the images were, six human volunteers were requested post-experiment to designate the 101 images from the calibrated image database as interesting or not interesting, corresponding to the number and complexity of the different features they contained. Two groups of images were assembled corresponding to low interest and high interest. The low interest group was populated with images having 0/6 votes, and high interest group with images having ≥4/6 votes. Observers were allowed to designate as interesting as many images from the collection as they wished. As a result, there were 28 images in the high interest group and 36 images in the low interest group. Analyses confirm that images belonging to the two sets did not yield significantly different results in the patch recognition task, with identical miss, FA and CR percentages and only an 8% higher percentage of hits for the interesting set. With 317 fixations per image for the non-interesting set and 318 for the interesting set and identical dwell-time averages, the eye movement characteristics were also not considered to play a significant role in the memory task success rate.

		DWELLTIME (s)		NUMBER OF FIXATIONS		HIT RATE	
		CORRECT	INCORRECT	CORRECT	INCORRECT	CORRECT	INCORRECT
SUBJECT	ABT	92.19	79.75	13.19	13.75	74	26
	ACB	97.96	130.08	11.61	11.96	60	40
	AV	112.24	91.62	13.70	13.85	72	28
	BSQ	83.67	103.33	15.49	16.17	75	25
	BW	142.55	134.47	12.00	11.68	62	38
	CE	104.11	110.29	10.94	9.48	71	29
	CLM	105.85	117.89	13.25	13.63	68	32
	CMG	266.41	187.31	8.98	10.46	78	22
	CNG	94.54	105.83	12.89	12.83	71	29
	CS	112.96	98.94	11.68	13.50	68	32
	FMT	111.83	82.79	12.53	12.93	71	29
	GK	112.26	75.32	11.76	12.63	66	34
	HRS	75.32	94.67	11.48	11.93	53	47
	IVDL	97.76	87.50	11.35	13.00	69	31
	JLS	107.95	111.58	10.48	11.16	65	35
	JM	141.19	121.78	9.58	9.61	57	43
	JNK	122.24	112.26	11.24	11.91	74	26
	JSS	82.18	77.29	13.92	14.76	75	25
	KAC	186.24	89.00	10.02	11.50	79	21
	KW	101.72	74.73	15.26	16.27	60	40
	LKC	118.54	130.33	9.85	12.33	89	11
	LW	116.42	106.84	14.27	14.28	64	36
	MPF	91.20	72.38	11.98	12.15	74	26
	NSS	313.48	224.00	6.88	6.88	67	33
	RGR	101.48	115.67	13.39	13.27	59	41
	SCP	87.38	91.47	14.41	14.80	63	37
	TT	121.04	135.42	11.96	12.08	65	35
	UR	99.23	109.88	12.80	12.29	68	32
	YL	134.73	89.74	10.49	11.05	71	29
AVERAGE		121.89	109.04	11.98	12.49	68.55	31.45

Table 2: Results for individual observers, showing differences between correct and incorrect responses for dwell-time and fixations/trial. Dwell-time entries marked in grey are those for which dwell-time for correct responses was on average higher than for incorrect responses. The number of fixations average shown in grey are those for which the number of fixations associated with incorrect response was on average higher than for correct responses.

4. DISCUSSION

With the 29 observers tested, there were approximately 300 fixations per image, or around 10 fixations per person per image. If these fixations were uniformly distributed across the image (rather than being clustered at visually salient locations), for a foveola sized window of 64×64 pixel we would expect to observe 1.6 fixations in the same window as any randomly nominated fixation. Since fixations are clustered towards the centre[13,26], and at regions of interest, the actual average number of other fixations within the same window as an average recorded fixation was 5.9. Furthermore, since the first fixation is considered to be attracted primarily to regions of significant bottom-up saliency[13] we observed a larger than average number of fixations surrounding the average first fixation, of 9.76. Since this first fixation is attracted to a highly salient region, we might reasonably expect a patch centred at this location to be more easily recognised than the last fixation patch, which was shown by our data (501 correct first fixations *vs.* 436 correct last fixations), though the effect is smaller than one might expect at around 10% higher probability of recognising the first patch (see Fig. 4), because of the dual issues of decreasing saliency over time coupled with decreasing latency between exposure and the memory test. This is in contrast to results for synthetic object memory tasks that showed more recently attended information have a higher probability of being recalled[8], possibly because the issue of fixation order was not measured. Despite research indicating that early fixations are likely to be directed to regions of bottom-up interest, this experiment using natural images with low-cognitive interest did not yield significantly different levels of saliency for first fixation compared with last (approximately 10th) fixation overall for the 29 observers and 101 images mined (Fig. 7) with the 8 image statistics applied.

On average, hit case results show higher entropy, mean luminance, MSE, SD, number of edges, and unique colours than miss cases (Fig. 5). These image dimensions correspond to the information content of the patch, and theories of eye movements conjecture that regions exhibiting higher information are likely to attract fixations to elucidate. Since these image dimensions may have contributed to regional saliency, it may be postulated that more salient regions may be more

memorable, with the caveat to this postulation being that higher contrast (measured either by RMS of Michelson's technique) which is considered by many to contribute to making image regions highly salient[15, 18], did not appear to increase the memorability of the patch. This is in agreement with existing work that observed the less robust encoding of contrast in visual memory using contrast gratings [9,28], though it is interesting that a precursor of the contrast calculation, SD, did show difference.

The data also indicate that for the FA and CR cases, the image statistics did not assist with the designation of the patch as belonging to the image. Counter-intuitively, it appears that for a number of the image metrics, that greater information was more likely to cause the observer to misclassify the patch (Fig. 6). It should be stated however that this effect was weaker than the differences in statistics associated with the Hit and Miss patches.

5. CONCLUSIONS

Using an ensemble of 101 stimuli selected from a calibrated natural image database, a task-driven eye-tracking experiment was undertaken to study the visual memory/recognition performance of 29 human volunteers. For each trial, experimental observers were presented with a full-screen natural image for 5 seconds, and then a foveola-sized image patch that originated either from the full-screen image they were just shown, or from another random image from the same database. Furthermore, if the patch originated from the image just shown, it was selected from either the observer's own first or last fixation coordinate rather than from a random location. Observers were instructed that patch structure rather than the brightness should be used to designate the patch, since the brightness was randomly shifted to remove the possibility of results being predominated by luminance-based matching. Recognition rates were compared across observers for first and last fixations, and the statistical properties of recognised and unrecognised patches analysed using a variety of metrics to investigate the relationship between patches statistical properties and their recognition rate.

For the eight image statistics applied to the patches we demonstrate that those that were recognised scored more highly in the several image dimensions than unrecognised patches, with the exception of the two contrast metrics which showed little variation. The data also indicates that for the FA and CR cases, the image statistics did not assist with the designation of the patch as belonging to the image. Counter intuitively, it appears that for decoy patches a number of image metrics were more likely to cause the observer to misclassify the patch.

For a number of image statistics applied, it can be seen that those that were recognised scored more highly in measurements of visual saliency than unrecognised patches. In agreement with existing work measuring visual memory with synthetic stimuli[9], one metric commonly applied to generate bottom-up maps of visual saliency did not show as dramatic an effect as the other metrics deployed: contrast calculated by both RMS and Michelson was seen to change little between recognised and unrecognised patches.

Although contrast is often considered to be a significant local statistic in the prediction of fixations, it was shown to have a weaker effect on the memorability of an image patch than other image dimensions measured. This may be because the HVS has to satisfy the dual goals of selecting salient regions for immediate attention (eye movement targets) and more gradually also learn about the environment[11]. This study suggests that a bottom-up memorability map could be constructed in a similar fashion to the saliency map but perhaps using differently weighted image properties, to provide an estimated recall rate of different image regions on natural scenes: the topic of a forthcoming study.

ACKNOWLEDGEMENTS

This research was funded by NSF grant ECS-0225451.

REFERENCES

1. Applied Science Laboratories (1998) Eye Tracking System Instruction Manual. Ver. 1.2.
2. Brainard, D.H. (1997) The Psychophysics Toolbox, Spatial Vision 10:433-436.
3. Buswell, G.T. (1935) How People Look at Pictures, Chicago: Univ. Chicago Press.
4. Choi, S-B., Ban S-W., Lee, M. (2004) Biologically motivated visual attention system using bottom-up saliency map and top-down inhibition, Neural Information Processing – Letters and Reviews Vol. 2 No. 1:19-25.
5. Dhavale, N. & Itti, L. (2003) Saliency based multi-foveated MPEG, Proceedings of IEEE Seventh International Symposium on Signal Processing and its Application.
6. Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. Journal of Experimental Psychology: Human Perception and Performance, 28, 113-136.
7. Hollingworth, A. (2004) Constructing Visual Representations of Natural Scenes: The Roles o f Short- and Long-term Visual Memory, Journal of Experimental Psychology: Human Perception and Performance. Vol. 30 No. 3, 519-537.
8. Irwin, D. E., Zelinksky, G. J. (2002) Eye movements and scene perception: Memory for things observed. Perception and Psychophysics Vol. 64 No. 6, 882-895.
9. Lee, B., & Harris, J. (1996). Contrast transfer characteristics of visual short-term memory. Vision Research, 36(14), 2159-2166.
10. McCarley, J. S., Wang, R. F., Kramer, A. F., Irwin, D. E., Petersen, M. S. (2003) How much memory does oculomotor search have? Psychological Science Vol. 14 No. 5 422-426.
11. Melcher, D. Kowler, E. (2001) Visual Scene Memory and the Guidance of Saccadic Eye Movements, Vision Research 41, 3597-3611.
12. Navalpakkam, V. & Itti, L. (2002) A Goal Oriented Visual Guidance Model, Lecture Notes in Computer Science, Vol. 2525: 453-461.
13. Parkhurt, D., Law K., Niebur, E. (2002) Modeling the role of salience in the allocations of overt visual attention, Vision Research 42:107-123.
14. Pelli, D.G. (1997) The VideoToolbox software for visual psychophysics: Transforming numbers into movies, Spatial Vision 10:437-442.
15. Privitera, C. M. & Stark, L. W. (2000) Algorithms for Defining Visual Regions-of-Interest: Comparisons with Eye Fixations, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 9. 970-982.
16. Rajashekar, U., Cormack, L.K., Bovik, A.C. (2002) Point of gaze analysis reveals visual search strategies, Human Vision and Electronic Imaging IX, Proc. of SPIE, Vol. 5292, San Jose, CA, January 18-22, 2004
17. Reinhard, E., Shirley, P., Troscianko, T. (2001) Natural image statistics for computer graphics, Technical Report UUCS 01-002, School of Computing, University of Utah.
18. Renegal, P., Zador, A. M. (1999) Natural Scene Statistics at the Center of Gaze, Network: Comput. Neural Syst. 10, 1-10.
19. Ruderman, D.L. (1994) The statistics of natural images, Network: Computation in Neural Systems 5:517-548.
20. Scharff, L.F.V., Ahumada, A.J., Hill, A.L. (1999) Disciminability Measures for Predicting Readability, Human Vision and Electronic Imaging I, Proceedings of SPIE Electronic Imaging, Vol. 3644.
21. van Hateren, J.H. & van der Schaaf, A. (1998) Independent component filters of natural images compared with simple cells in primary visual cortex. Proc. Royal Society of London B 265:359-366.
22. Wandell, B. (1995) The Foundations of Vision, Sinaur Associates Inc.
23. Wichmann, F. A., Sharpe, L. T., Gegenfurtner, K. R. (2002) The contributions of color to recognition memory for natural scenes, Journal of Experimental Psychology: Learning, Memory and Cognition. Vol. 28, No. 3 509-520.
24. Yarbus, A.L. (1967) Eye Movements and Vision (B. Haigh, Trans.) New York: Plenum Press. (Original work published in 1956).
25. Tootell, R.B.H., Hadjikhani, N. (2000) Attention – brains at work!, Nature Neuroscience Vol. 3 No.3. 206-208.
26. Murphy, H., Duchowski, A.T. (2002) Modeling Visual Attention in VR: Measuring the Accuracy of Predicted Scanpaths, *EuroGraphics*.
27. Einhäuser, W., König P. (2003) Does luminance-contrast contribute to a saliency map for overt visual attention?, Eur J Neurosci. 17(5): 1089-97
28. Magnussen, S. (2000). Low-level memory processes in vision. Trends in Neurosciences, 23(6), 247-251.
29. Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. Nature, 390, 279-281.

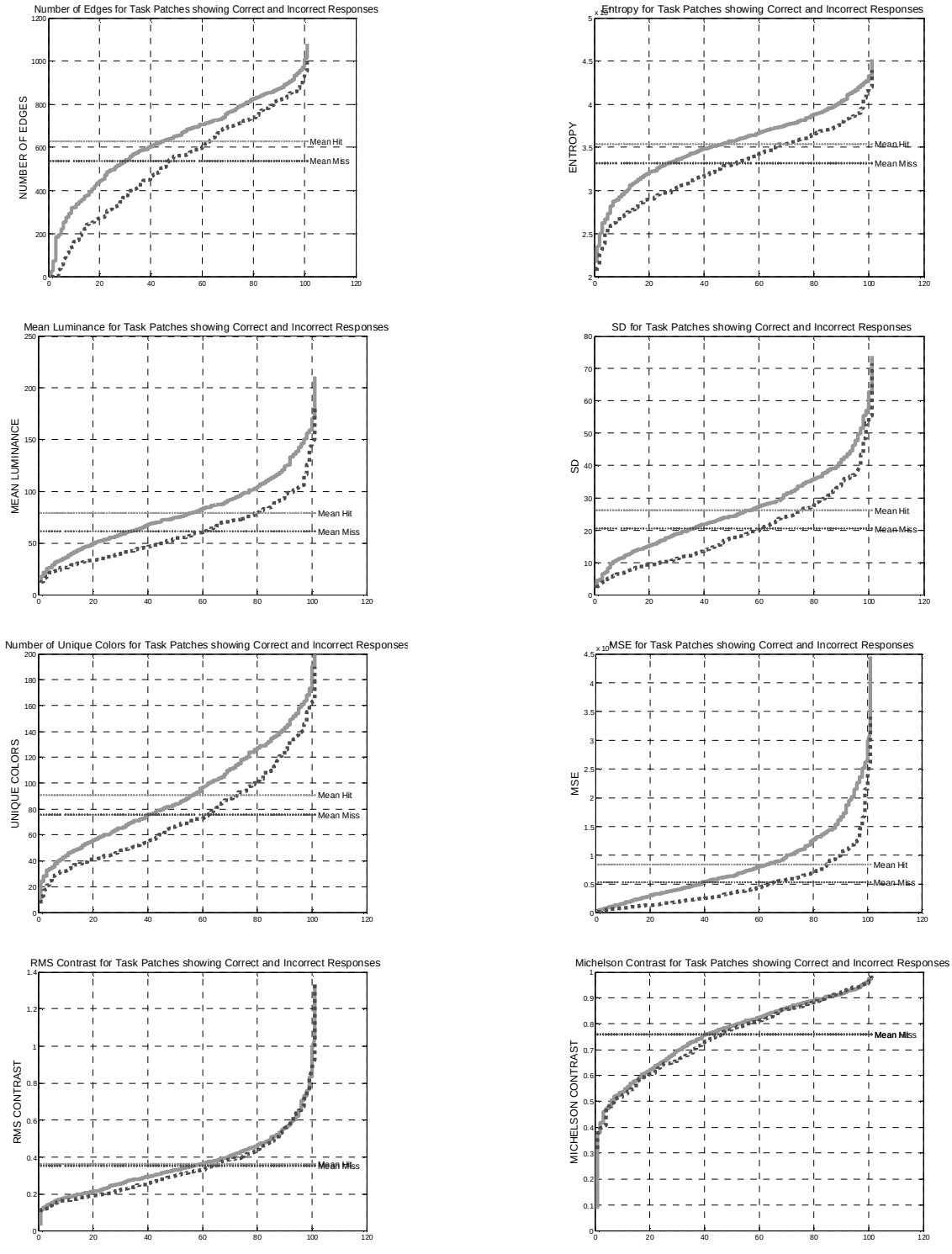

Fig. 5. Sorted Patch Statistics for Real Patches (HIT [solid line] and MISS [dotted line] Responses).

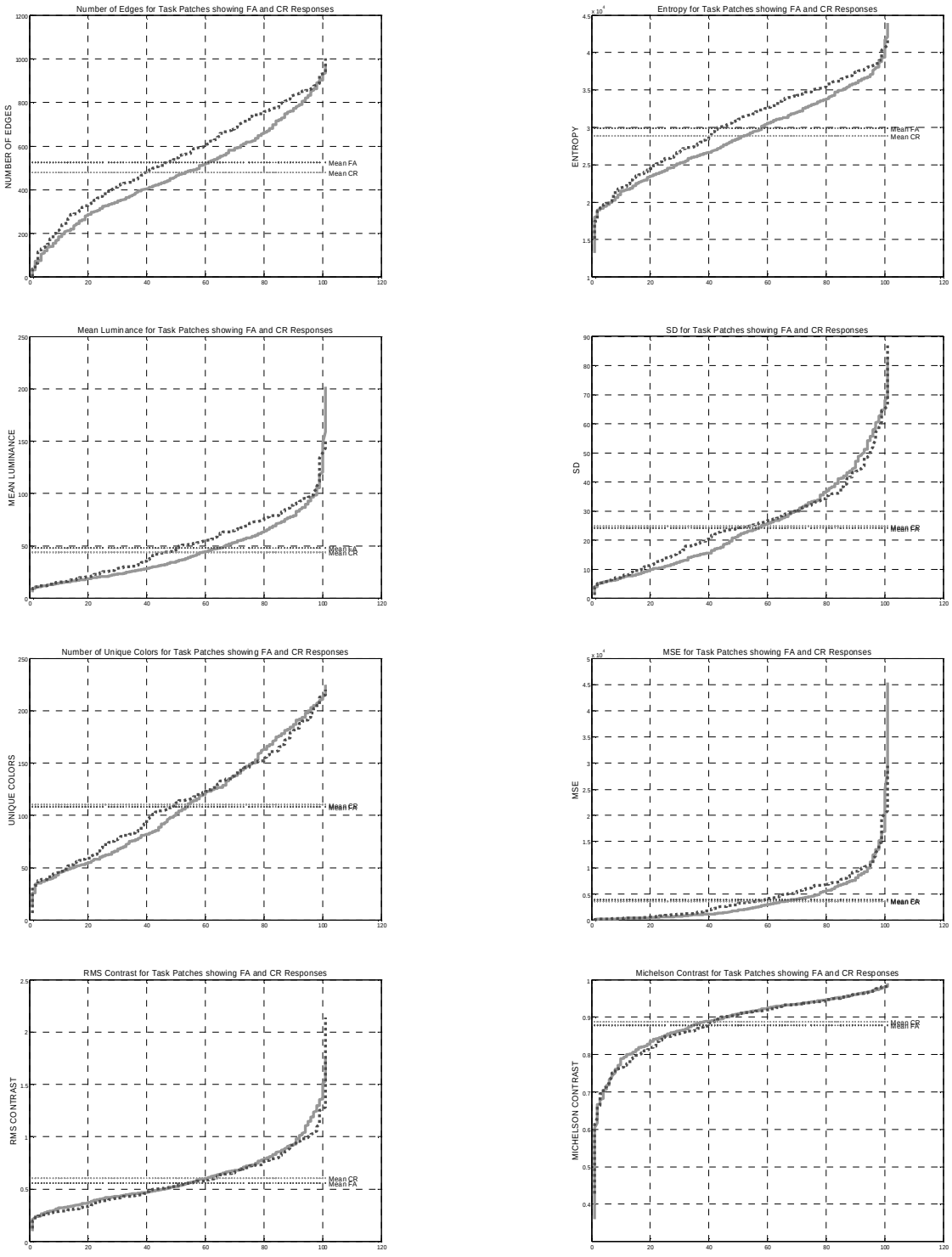

Fig. 6. Sorted Patch Statistics for Decoy Patches (FALSE ALARM [dotted line] and CORRECT REJECTION [solid line] Responses).

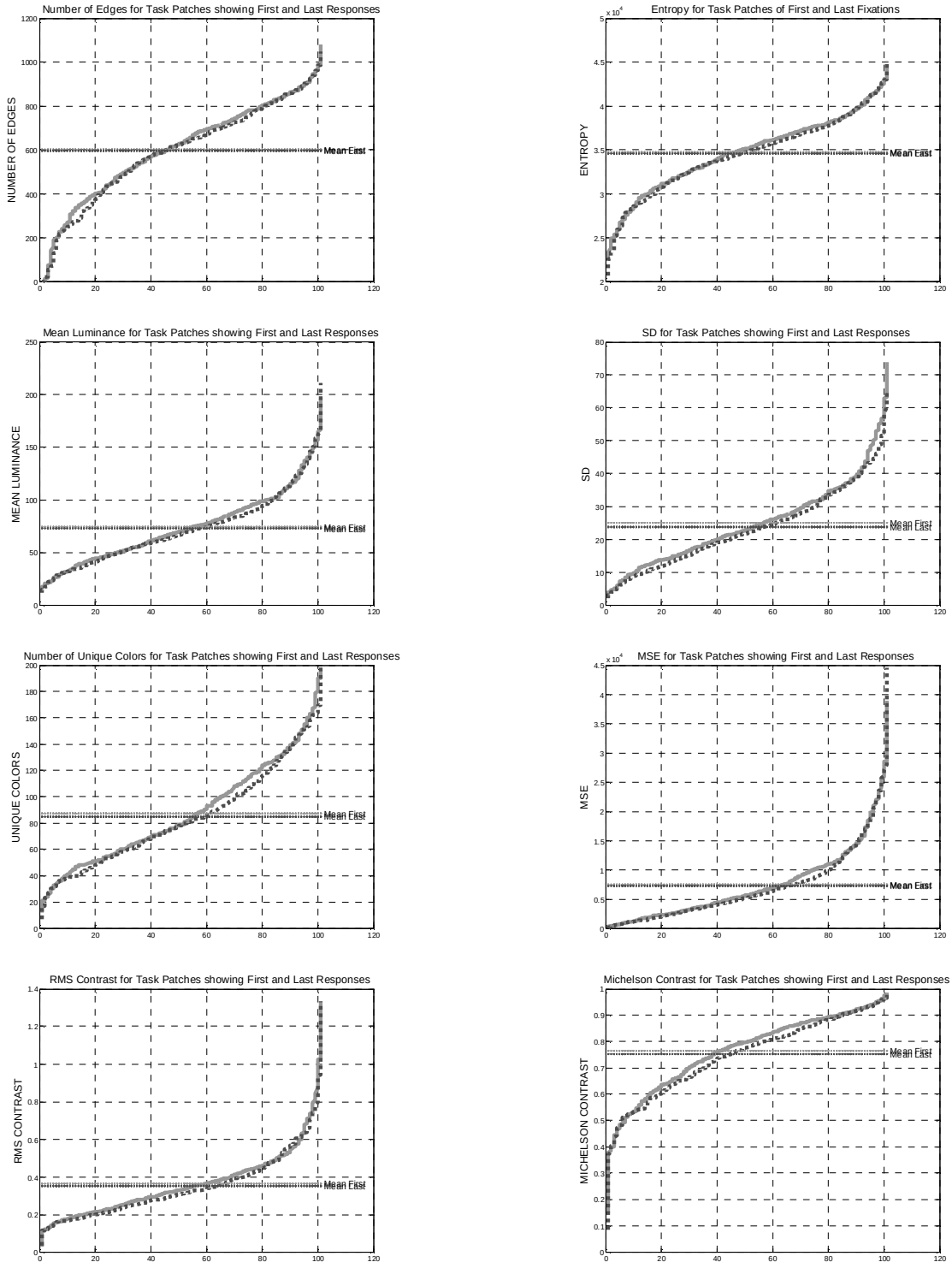

Fig.7. Sorted Patch Statistics for Real FIRST [solid line] and Real LAST [dotted line] Fixation.