Chapter 1

Effects of spatial and temporal prediction during prolonged learning of novel objects

1.1 Introduction

The core challenge of object recognition is concerned with solving the invariance problem (DiCarlo, Zoccolan, & Rust, 2012). Essentially, object identity must remain be invariant across large changes in a an object's visual position, scale, rotation, and viewpoint to generate successful behavior. Understanding how exactly the brain solves this problem has been a major focus of the object recognition literature with the bulk of data and models suggesting that it is solved gradually by a hierarchy of neural processing mechanisms from V1 through inferior temporal (IT) cortex that extract increasingly complex features at each stage with increasing tolerance to transformations (Riesenhuber & Poggio, 1999; Serre, Oliva, & Poggio, 2007; O'Reilly, Wyatte, Herd, Mingus, & Jilk, 2013).

One question that remains to be fully answered is how invariance is learned in the first place. One intriguing hypothesis is that a temporal association rule might form associations between multiple samples of a single object as it undergoes transformations (Stringer, Perry, Rolls, & Proske, 2006; Wallis & Baddeley, 1997; Isik, Leibo, & Poggio, 2012). It has been demonstrated that some

neurons can form temporal associations between arbitrary pairs of stimuli (Sakai & Miyashita, 1991), including a population in monkey IT cortex (Meyer & Olson, 2011). Experiments by DiCarlo and colleagues have indicated that these temporal associations can build new invariance for specific object transformations including changes in position and size (Cox, Meier, Oertelt, & DiCarlo, 2005; Li & DiCarlo, 2008, 2010). This new invariance was learned without supervised reward, suggesting that it could be a natural consequence of generic neural processing mechanisms given the spatiotemporal statistics of the physical world (Li & Dicarlo, 2012).

Evidence of invariance due to temporal associations has yet to be demonstrated in IT neurons for three-dimensional changes in viewpoint (although see Wallis & Bulthoff, 2001; Wallis, Backus, Langer, Huebner, & Bulthoff, 2009, for relevant human behavioral work). IT neurons typically have a tuning curve of approximately 90 degrees for newly acquired three-dimensional objects (Logothetis, Pauls, Bulthoff, & Poggio, 1994; Logothetis, Pauls, & Poggio, 1995), but recognition is possible in a viewpoint invariant manner especially after prolonged learning (Wallis & Bulthoff, 1999). Intuitively, predictable motion from one moment to the next could be considered important for encoding three-dimensional objects (Lawson, Humphreys, & Watson, 1994; Stone, 1998; Vuong & Tarr, 2004; Balas & Sinha, 2009BalasSinha09c; Chuang, Vuong, & Bulthoff, 2012), and thus a temporal association rule could plausibly be used to learn viewpoint invariance from naturalistic spatial structure of objects.

The work described in this chapter investigated the role of predictable spatiotemporal information during a novel object learning task. In the context of the LeabraTI model (Chapter **??**) as well as several other theories of sensory prediction (Arnal & Giraud, 2012; Giraud & Poeppel, 2012), spatial structure might be learned from predictions about incoming sensory information made at regular temporal intervals. To test this hypothesis, both the spatial and temporal predictability of changes in objects' viewpoint were manipulated during a training period followed by a series of same-different judgements over static test views.

Somewhat surprisingly, the results of the experiment indicated that accuracy was lowest when stimuli were learned in a combined spatially and temporally predictable context and highest

when learned in a completely unpredictable context. Reaction times were also slower when objects were learned with spatial predictability.

1.2 Methods

1.2.1 Participants

A total of 62 students from the University of Colorado Boulder participated in the experiment (ages 18-22, mean=19.11 years; 30 male, 32 female). All participants reported normal or corrected-to-normal vision and received course credit as compensation for their participation. Informed consent was obtained from each participant prior to the experiment in accordance with Institutional Review Board policy at the University of Colorado.

1.2.2 Stimuli

Novel "paper clip" objects similar to those used in previous investigations of three-dimensional object recognition (Bulthoff & Edelman, 1992; Edelman & Bulthoff, 1992; Logothetis et al., 1994; Logothetis et al., 1995; Sinha & Poggio, 1996) were used as stimuli (see Chapter **??** Methods). A total of eight objects were used – four as targets and four as distractors. The four target objects were also used in the Chapter **??** experiment. Target and distractor objects were paired together for the purposes of the experiment. All objects are shown in Figure 1.1.

1.2.3 Procedure

The experiment was divided into 16 blocks, each containing a training period followed by a series of test trials (Figure 1.2). During the training period of a given block, participants observed one of the target objects rotate about its y-axis. The object either rotated coherently (i.e., spatially predictable, S+ conditions) or in a random manner (S- conditions). Coherent rotation was composed of adjacent views spaced 12 degrees apart. The object made four complete rotations during the study period. All views of the object were still presented four times each in the random case.

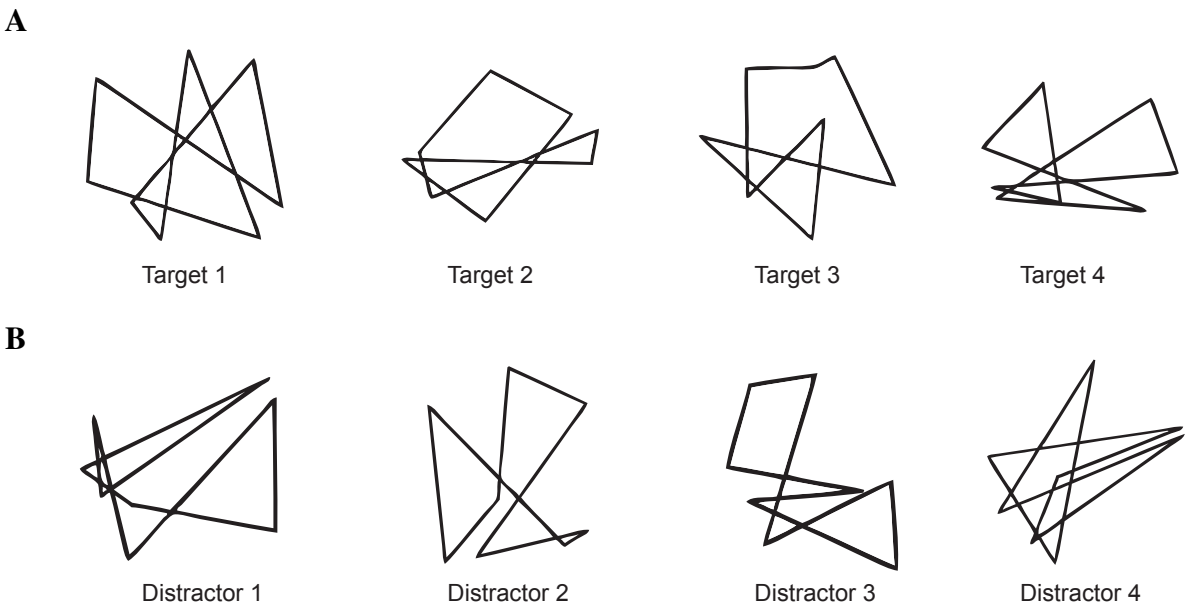

Figure 1.1: Novel "paper clip" objects

Four target (**A**) and four distractor object pairs (**B**) used in the experiment. See Chapter **??** Methods for additional information.

The presentation rate during the study period was either 10 Hz with a 50 ms on time and 50 ms off time (i.e., temporally predictable, T+ conditions) or variable with a 50 ms on time and off times ranging from 16.67-400 ms (T- conditions).

The S+/- and T+/- conditions were crossed and each of the target-distractor object pairs was assigned to one of the four conditions. These assignments were approximately counterbalanced across participants (Assignment 1: N=15; Assignment 2: N=17; Assignment 3: N=15; Assignment 4: N=15). Each block condition with its target-distractor pairing was repeated for four blocks during the experiment. Block order randomized was randomized for each participant.

During each block, participants were instructed to study the target object during the training period and then complete a series of 30 test trials. On each test trial, either the target object or its paired distractor was presented. Participants were instructed to respond "same" if they believed the object depicted the trained target object or "different" if they believed it depicted the distractor object. Half of the test trials contained 15 views of the target object spaced 24 degrees apart, and

the other half contained 15 views of the distractor, also spaced 24 degrees apart. Test trials were shown in a random order and feedback was withheld to prevent participants from changing their response criteria over the course of a block.

The experiment was displayed on an LCD monitor at native resolution operating at 60 Hz using the Psychophysics Toolbox Version 3 (Brainard, 1997; Pelli, 1997). All stimuli were presented at central fixation on an isoluminant 50% gray background and subtended approximately 5 degrees of visual angle. Test trials began with a fixation cross (200 ms) followed by a blank (400 ms) followed by the probe stimulus (100 ms). Participants were required to respond within 2000 ms. Subsquent test trials were separated by a variable intertrial interval of 1000-1400 ms.

The experiment began with a practice block to ensure that participants understood the task. The training period during the practice block was always spatially and temporally predictable and used a reserved target object and distractor that were not further used in any of the experimental blocks. During the practice test trials, participants received feedback after responding according to whether they were correct or incorrect. After completing the practice block, participants were informed that future training periods could be presented in spatially and/or temporally unpredictable manners.

1.3 Results

Three subjects were excluded from behavioral analysis for accuracy 2.7σ (or further) below mean accuracy across subjects. All three excluded subjects were assigned condition-object 3 resulting in the final counterbalancing – Assignment 1: N=15; Assignment 2: N=14; Assignment 3: N=15; Assignment 4: N=15. The remaining 59 subjects were submitted to a 2x2 ANOVA with spatial and temporal predictability as within-subjects factors and counterbalancing assignment as a between-subjects factor. Accuracy and reaction times were collected during the experiment and were used to compute *d'*, a measure of sensitivity that takes into account response bias, and inverse efficiency, a measure that combines accuracy and reaction times (Townshend & Ashby, 1978, 1983). These behavioral measures are plotted in Figure 1.3.

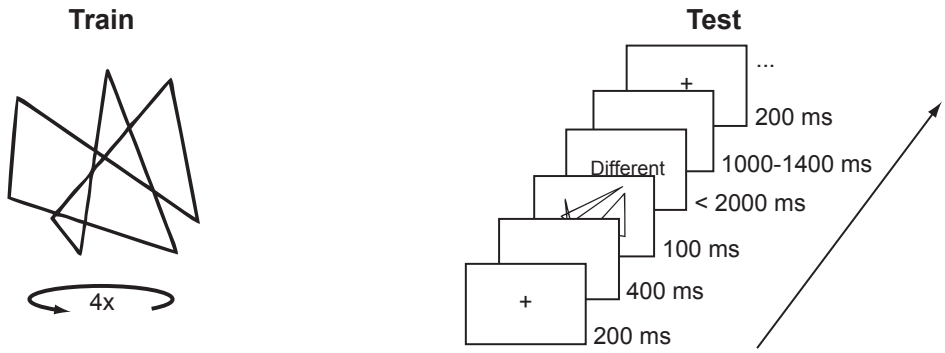

Figure 1.2: Experimental procedure

Experimental trials were composed of a training period followed by a testing period. The training period depicted a target object rotating a total of four times around its vertical axis. Rotation was either spatially and temporally predictable, spatially predictable or temporally predictable only, or completely unpredictable. The test period contained 30 trials that depicted either the training object or its paired distractor at 15 viewing angles each.

Overall, subjects were less accurate when the training period was spatially predictable ($F(1, 57) = 4.50$, $p = 0.038$) or temporally predictable ($F(1, 57) = 4.20$, $p = 0.046$). The interaction between spatial and temporal predicability failed to reach significance ($F(1, 57) = 0.20$, $p = 0.659$). Subjects were least accurate for the combined spatial and temporal predictability condition (denoted S+T+ in Figure 1.3). This condition significantly differed from the completely unpredictable condition (S-T-) ($t(58) = -2.8587$, $p = 0.001$), and trended toward significance for conditions with only spatial or only temporal predictability (S+T+ versus S+T-, $t(58) = -1.60$, $p = 0.116$; S-T- versus S+T+ versus S-T+, $t(58) = -1.77$, $p = 0.082$).

When responses are transformed into *d'*, effects of spatial predictability and temporal predictability during the training period trended toward significance (spatial, $F(1, 57) = 3.07$, $p = 0.085$; temporal, $F(1, 57) = 3.00$, $p = 0.089$). The interaction between spatial and temporal predicability failed to reach significance ($F(1, 57) = 0.00$, $p = 0.985$). The pattern of results as a function of predictability during the training period was the same as for accuracy, and thus this failure to reach critical significance likely reflects the loss of power when transforming responses into *d'* due to discarding misses and correct rejections.

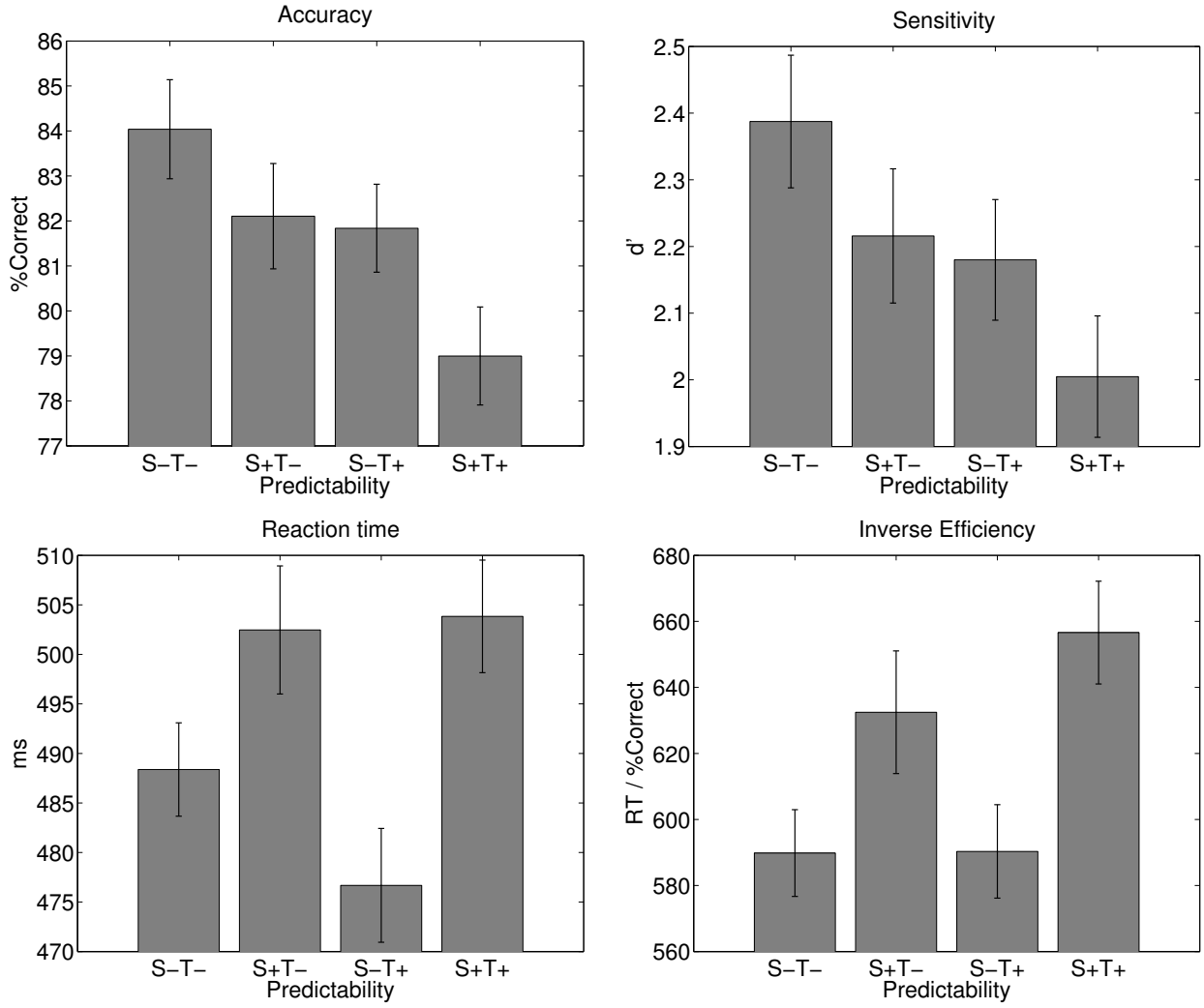

Figure 1.3: Behavioral measures of spatial and temporal predictability

Accuracy, *d'* (sensitivity), reaction time, and inverse efficiency (reaction time divided by percent correct) as a function of predictability during the training period. S-/+ refers to spatially unpredictable and predictable, T-/+ to temporally unpredictable and predictable. Error bars depict within-subjects error using the method described in Cousineau (2005) adapted for standard error.

Subjects were slower to respond when the training period was spatially predictable ($F(1, 57) = 10.99$, $p = 0.002$). A similar effect for temporal predictability failed to reach significance ($F(1, 57) = 0.53$, $p = 0.471$), nor did the interaction between spatial and temporal predictability ($F(1, 57) = 1.21$, $p = 0.276$). Effects on inverse efficiency (defined as reaction time divided by percent correct) were similar. Inverse efficiency was highest when the training period was spatially

predictable ($F(1, 57) = 9.64$, $p = 0.003$), but did not significantly differ as a function of temporal predictability ($F(1, 57) = 0.45$, $p = 0.507$), nor when considering the interaction between spatial and temporal predictability ($F(1, 57) = 0.71$, $p = 0.403$).

Effects were highly variable across target objects (Figure 1.4). Target-condition assignment did not significantly affect any of the behavioral measures (all p's > 0.05), but often interacted with predictability effects and their interactions. One reason for this variability regards the orthographic projection used to render the objects. Previous research has indicated that recognition accuracy fluctuates as a function of how well the two-dimensional projection of an object captures its full three-dimensional structure (Balas & Sinha, 2009). For example, when there is a large amount of foreshortening in the projection, it could be difficult to infer the length of line segments that compose the object, impairing recognition. These degenerate projections are generally diametrically opposed on the object.

To investigate this hypothesis, accuracy was computed as a function of viewing angle for each target object to investigate whether it interacted with predictability during the training period (Figure 1.5). Only accuracy was considered in this analysis as each data point only corresponded to four trials per subject and thus transformation to *d'* was not plausible. Test trials during which distractor objects were presented were also excluded from this analysis since there is no consistent relationship between the targets and distractors across viewing angles and thus they would only contribute noise. With the exception of target object 1, all objects indicated fluctuations in accuracy as a function of viewing angle with two diametrically opposed degenerate views. The most consistent differences in accuracy between training conditions appeared to be localized to the troughs of the accuracy function, corresponding to these degenerate views.

Standard statistical tests did not have enough power to detect differences between conditions for degenerate views due to the low trial counts for each data point. To address this design limitation, a bootstrapping method was used to resample the available data in these cases. The completely unpredictable (S-T-) and combined spatial and temporal predictability (S+T+) were used to assess differences due to training context since these two conditions elicited the greatest

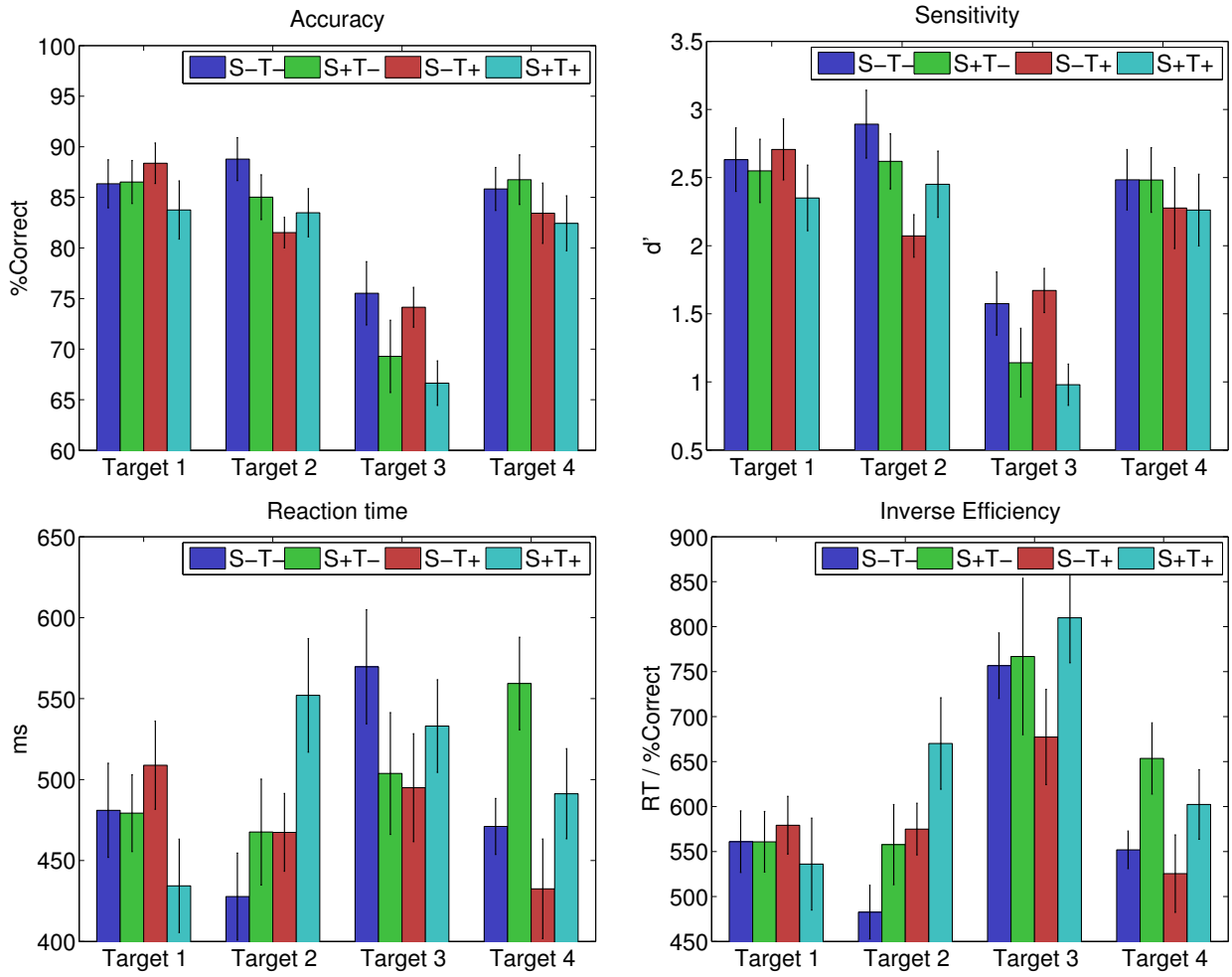

Figure 1.4: Behavioral measures for each target object

Accuracy, *d'*, reaction time, and inverse efficiency for each target object. Horizontal axes denote target and colors predictability during the training period. Error bars depict between-subjects standard error.

difference in average accuracy in the full analysis. The accuracy function over viewing angles was collapsed across conditions and the two minima associated with degenerate views were identified for each object. For target object 1, the two views were at $\theta = \{24°, 312°\}$, object 2: $\theta = \{48°, 240°\}$ object 3: $\theta = \{144°, 312°\}$, and object 4 $\theta = \{24°, 192°\}$. S-T- and S+T+ accuracy was for each object's degenerate views and resampled with replacement from the 59 subjects for 10000 iterations. This produced degenerate view accuracy distributions for spatiotemporally unpredictable and predictable training contexts (Figure 1.6). Accuracy for was lower for degenerate views for the

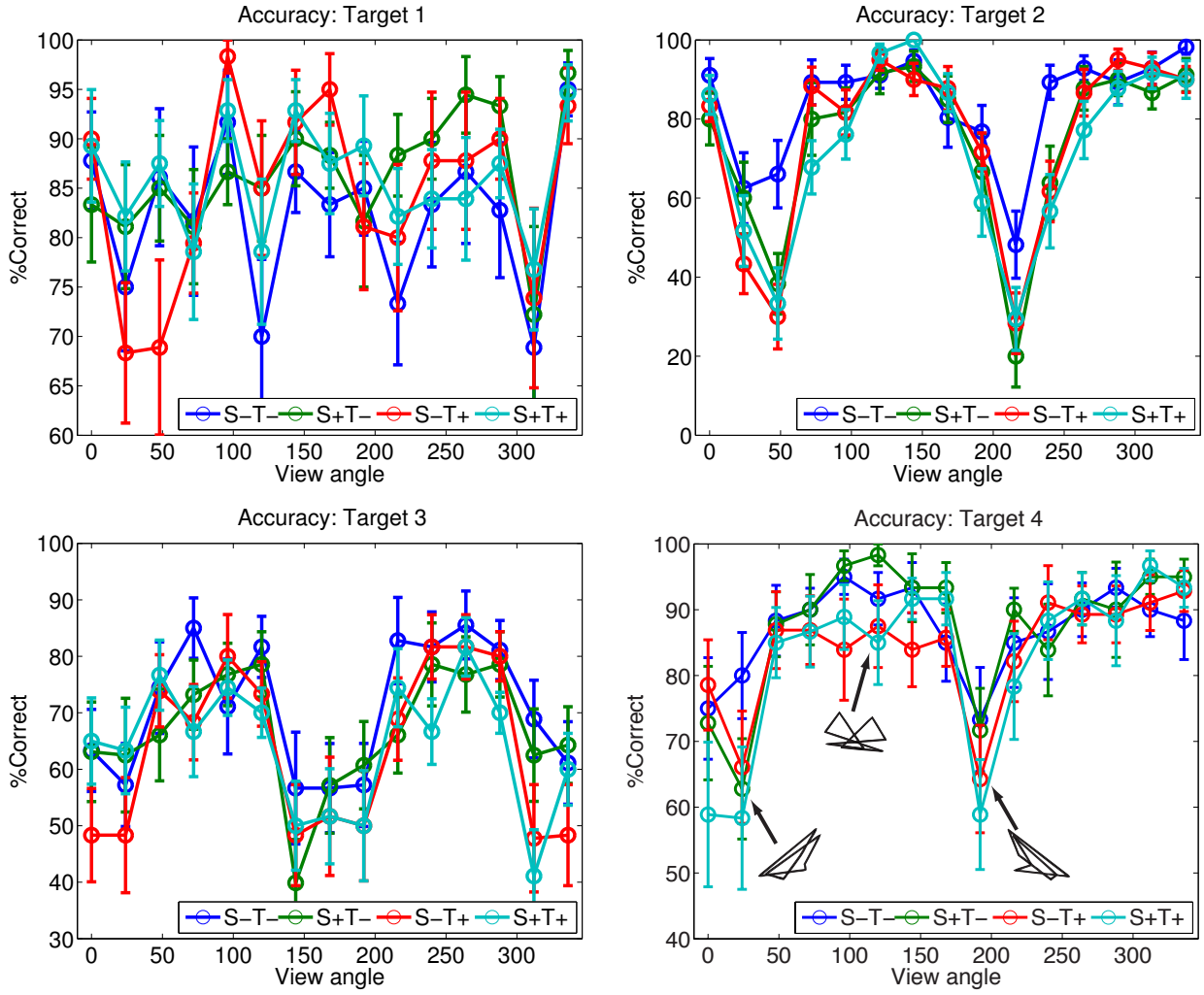

Figure 1.5: Accuracy as a function of viewing angle for each target object

Target accuracy at each viewing angle presented during the test periods. Horizontal axes denote viewing angle and colors predictability during the training period. Error bars depict between-subjects standard error. Diametrically opposed foreshortened views and one canonical view are shown for target object 4.

spatiotemporally predictable condition for all target objects except target 1, which didn't exhibit the patterned accuracy function that other targets did. The predictability difference in accuracy for degenerate views was significant at the 90% alpha level (i.e., the confidence interval of the difference between means did not include zero) for all target objects except target 1.

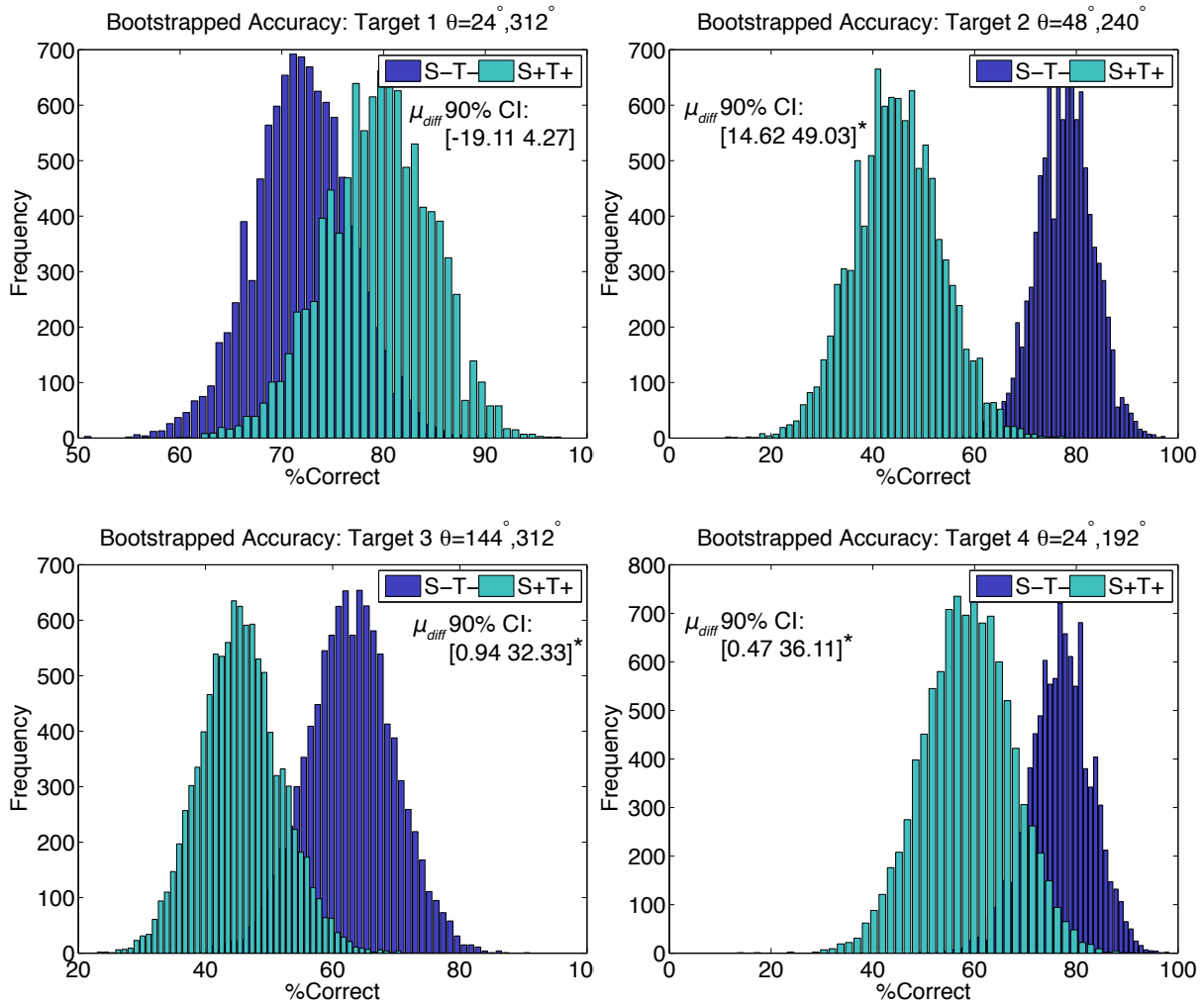

Figure 1.6: Bootstrapped accuracy for degenerate views

Average target accuracy for degenerate views resampled with replacement from the 59 subjects for 10000 iterations. Viewing angle for averaging is noted for each target object. Asterisks denote significant differences based on 90% confidence intervals.

1.4 Discussion

1.4.1 Summary of results

The work described in this chapter investigated how predictability biased learned representations of novel objects. The experimental paradigm used to address this question involved training participants to recognize novel objects while manipulating their spatial and temporal predictabil-

ity. Somewhat surprisingly, accuracy was lowest when stimuli were learned in a combined spatially and temporally predictable context and highest when learned in a completely unpredictable context. Reaction times were also slower when objects were learned with spatial predictability.

Behavioral measures were highly variable across objects. There was some indication that the principal differences between predictability conditions during training were driven primarily by degenerate viewing angles caused by three-dimensional foreshortening in the objects used. In three out of four objects, accuracy was lower for degenerate views learned in a spatiotemporally predictable context compared to a completely unpredictable one.

1.4.2 A behavioral disadvantage for spatial prediction during object learning

Intuitively, spatial predictability should be advantageous for learning the three-dimensional structure of objects given that it is the learning context within which the visual system evolved. However, the literature contains a mixture of contradictory effects regarding the utility of spatially predictable information during object learning and recognition. Initial experiments described in Lawson et al. (1994) with depth-rotated line drawings of familiar objects demonstrated the expected increase in recognition accuracy for spatially predictable sequences. A number of studies have found that studying depth-rotating sequences of novel objects under one ordering and then testing with a different ordering impairs recognition (Stone, 1998; Vuong & Tarr, 2004; Chuang et al., 2012), implying that learned predictability about the sequence is used to encode the object (BalasSinha09c). The foreshortening model advanced in (Balas & Sinha, 2009) also provided a better match to observers' data by incorporating spatial information (e.g., the first- and second-order derivatives of the foreshortening function over object views).

Some of the experiments described in Lawson et al. (1994), however, demonstrated better accuracy for sequences studied with weak spatial predictability (maximum of two spatially coherent frames in the sequence) than total spatial predictability. Experiments described in Harman and Humphrey (1999) failed to find any positive or negative effects of spatial predictability on accuracy. They did, however, increase in reaction time for objects learned in a spatially predictable

context, similar to the one reported here. One possible reason for the behavioral disadvantage for objects learned with spatial predictability is that less attention is necessary in these conditions. A constantly changing sequence of views might require more attentive processing to encode whereas the relatively low amount of change between views in spatially predictable sequences is comparatively "unsurprising" such that some views might be overlooked during encoding. However, there was some indication that the adverse effect of spatial predictability was driven primarily by the degenerate views of the stimuli used in the present work. A more focused experiment is clearly necessary to explicitly test the hypothesis that behavioral performance is impaired for degenerate views learned in a spatially predictable context but relatively intact for canonical views.

1.4.3 Viewpoint invariance for "paper clip" objects

The "paper clip" objects used in the current work have a long history of use in studies of three-dimensional viewpoint effects in human observers (Bulthoff & Edelman, 1992; Edelman & Bulthoff, 1992; Sinha & Poggio, 1996) as well as studies monkey physiology studies (Logothetis et al., 1994; Logothetis et al., 1995). The objects are easy to generate systematically and thus can be combined with a staircase procedure to titrate difficulty or can be generated en masse to find the parameters that elicit maximal responses from neurons. Various effects with the objects have been replicated using computational models of object recognition with identical stimuli (Riesenhuber & Poggio, 1999) and mathematical properties of the objects are known to capture a large amount of variability in behavior (Balas & Sinha, 2009). Thus, it can be reasonably concluded that paper clip objects are a useful class of stimuli.

However, other work brings under question the ecological validity of paper clip objects. The objects are constructed from thin line segments separated by empty space and thus self-occlusion of features is less of a problem than for three-dimensional volumetric objects with surfaces. This might imply that view invariance is not actually necessary to represent the full three-dimensional structure of paper clip objects, since the majority of features can be extracted from a static view. Accordingly, studies comparing three-dimensional objects composed of line segment with volu-

metric objects found that the line segment objects were not represented in a viewpoint invariant manner (FarrahRochlinKlein94PizloStevenson99).

Thus, it is is possible that a spatiotemporally predictable training context simply not optimal for learning to represent paper clip objects. Temporal association mechanisms (Stringer et al., 2006; Wallis & Baddeley, 1997; Isik et al., 2012) might bias the development of viewpoint invariance by forming unwanted associations between canonical and degenerate views, which could lower overall accuracy levels and slow reaction times. This explanation would require that these mechanisms were somehow not elicited in the unpredictable learning contexts.

References

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. Trends in Cognitive Sciences, 16(7), 390–398.

Balas, B. J., & Sinha, P. (2009). The role of sequence order in determining view canonicality for novel wire-frame objects. Attention, Perception & Psychophysics, 71(4), 712–723.

Brainard, D. (1997). The Psychophysics Toolbox. Spatial Vision, 10(4), 433–436.

Bulthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. Proceedings of the National Academy of Sciences of the United States of America, 89(1), 60–64.

Chuang, L. L., Vuong, Q. C., & Bulthoff, H. H. (2012). Learned non-rigid object motion is a view-invariant cue to recognizing novel objects. Frontiers in Computational Neuroscience, 6(26).

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Massons method. Tutorials in Quantitative Methods for Psychology, 1(1), 42–45.

Cox, D. D., Meier, P., Oertelt, N., & DiCarlo, J. J. (2005). 'Breaking' position-invariant object recognition. Nature Neuroscience, 8(9), 1145–1147.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? Neuron, 73(3), 415–434.

Edelman, S., & Bulthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. Vision Research, 32(12), 2385–2400.

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. Nature Neuroscience, 15(4), 511–517.

Harman, K. L., & Humphrey, G. K. (1999). Encoding 'regular' and 'random' sequences of views of novel three-dimensional objects. Neuropsychologia, 28(5), 601–615.

Isik, L., Leibo, J. Z., & Poggio, T. (2012). Learning and disrupting invariance in visual recognition with a temporal association rule. Frontiers in Computational Neuroscience, 6(37).

Lawson, R., Humphreys, G. W., & Watson, D. G. (1994). Object recognition under sequential viewing conditions: Evidence for viewpoint-specific recognition procedures. Perception, 23(5), 595–614.

Li, N., & DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science, 321(5895), 1502–1507.

Li, N., & DiCarlo, J. J. (2010). Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. Neuron, 67(6), 1062–1075.

Li, N., & Dicarlo, J. J. (2012). Neuronal learning of invariant object representation in the ventral visual stream is not dependent on reward. The Journal of Neuroscience, 32(19), 6611–20.

Logothetis, N., Pauls, J., Bulthoff, H., & Poggio, T. (1994). View-dependent object recognition by monkeys. Current Biology, 4(5), 401–414.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. Current Biology, 5(5), 552–563.

Meyer, T., & Olson, C. R. (2011). Statistical learning of visual transitions in monkey inferotemporal cortex. Proceedings of the National Academy of Sciences of the United States of America, 108(48), 19401–19406.

O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., & Jilk, D. J. (2013). Recurrent processing during object recognition. Frontiers in Psychology, 4(124).

Pelli, D. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision, 10(4), 437–442.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nature Neuroscience, 2(11), 1019–1025.

Sakai, K., & Miyashita, Y. (1991). Neural organization for the long-term memory of paired associates. Nature, 354(6349), 152–155.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Sciences of the United States of America, 104(15), 6424–6429.

Sinha, P., & Poggio, T. (1996). Role of learning in three-dimensional form perception. Nature, 384(6608), 460–463.

Stone, J. V. (1998). Object recognition using spatiotemporal signatures. Vision research, 38(7), 947–951.

Stringer, S. M., Perry, G., Rolls, E. T., & Proske, J. H. (2006). Learning invariant object recognition in the visual system with continuous transformations. Biological Cybernetics, 94(2), 128–142.

Townshend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In J. N. Castellan Jr., & F. Restle (Eds.), Cognitive Theory: Volume 3 (pp. 200–239). Hillsdale, NJ: Lawrence Erlbaum Associates.

Townshend, J. T., & Ashby, F. G. (1983). Stochastic Modeling of Elementary Psychological Processes. Cambridge: Cambridge University Press.

Vuong, Q. C., & Tarr, M. J. (2004). Rotation direction affects object recognition. Vision Research, 44(14), 1717–1730.

Wallis, G., Backus, B. T., Langer, M., Huebner, G., & Bulthoff, H. (2009). Learning illumination- and orientation-invariant representations of objects through temporal association. Journal of Vision, 9(7).

Wallis, G., & Baddeley, R. (1997). Optimal, unsupervised learning in invariant object recognition. Neural Computation, 9(4), 883–894.

Wallis, G., & Bulthoff, H. (1999). Learning to recognize objects. Trends in Cognitive Sciences, 3(1), 22–31.

Wallis, G., & Bulthoff, H. (2001). Effects of temporal association on recognition memory. Proceedings of the National Academy of Sciences of the United States of America, 98(8), 4800–4804.