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 (DicarloZoccolanRust12). 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 (RiesenhuberPoggio99SerreOlivaPoggio07OReillyWyatteHerdEtAl13).

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 (StringerPerryRollsEtAl06WallisBaddeley97IsikLeiboPoggio12). It has been demonstrated that some neurons can form temporal associations between arbitrary pairs of stimuli (SakaiMiyashita91), including a popula-

tion in monkey IT cortex (MeyerOlson11). 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 (CoxMeierOerteltEtAl05LiDiCarlo08LiDiCarlo10). This new invariance can be 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 (LiDiCarlo12).

Evidence of invariance due to temporal associations has yet to be demonstrated in IT neurons for three-dimensional changes in viewpoint (although see Wallis-Bulthoff01WallisBackusLangerEtAl09, for relevant human behavioral work). IT neurons typically have a tuning curve of approximately 90 degrees for three-dimensional objects (LogothetisPauls-BulthoffEtAl94LogothetisPaulsPoggio95), but these objects can be recognized irrespective of viewpoint after prolonged exposure(WallisBulthoff99EdelmanBulthoff92TarrGauthier98). Intuitively, predictable motion from one moment to the next could be considered important for encoding three-dimensional structure (LawsonHumphreysWatson94Stone98VuongTarr04BalasSinha09bBalasSinha09cChuangVuongBulthoff12), and thus a temporal association rule could plausibly be used to group together multiple viewpoints 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 (ArnalGiraud12GiraudPoeppel12), 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 (BulthoffEdelman92EdelmanBulthoff92LogothetisPaulsBulthoffEtAl94LogothetisPaulsPoggio95SinhaPoggio96) 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.

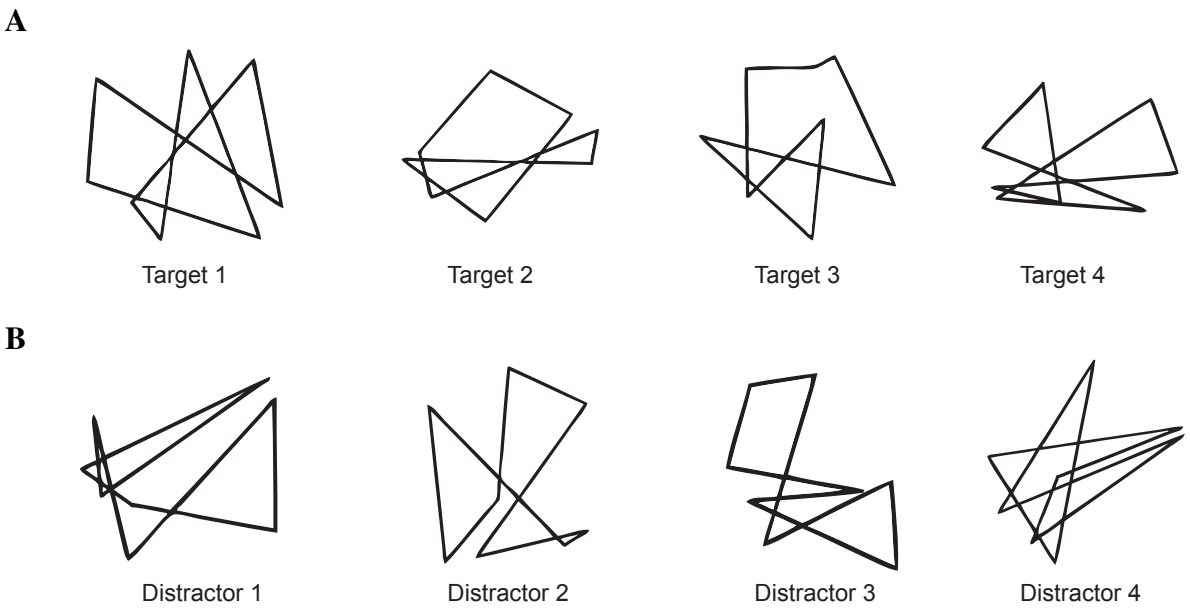

Figure 1.1: Novel "paper clip" objects

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 (Brainard97Pelli97). 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 (TownshendAshby78TownshendAshby83). These behavioral measures are plotted in Figure 1.3.

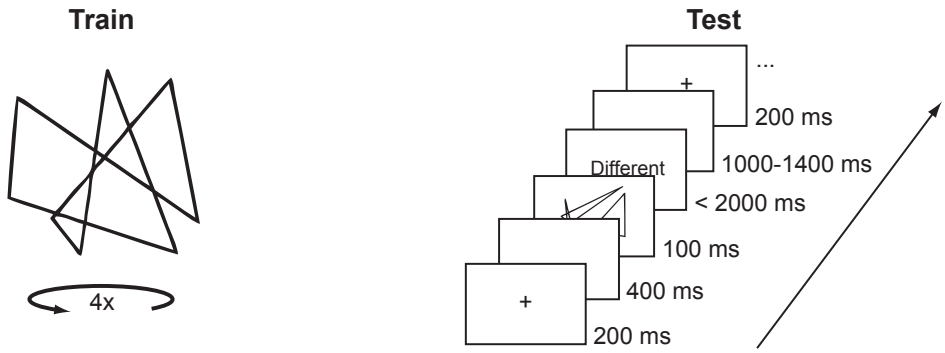

Figure 1.2: Experimental procedure

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.

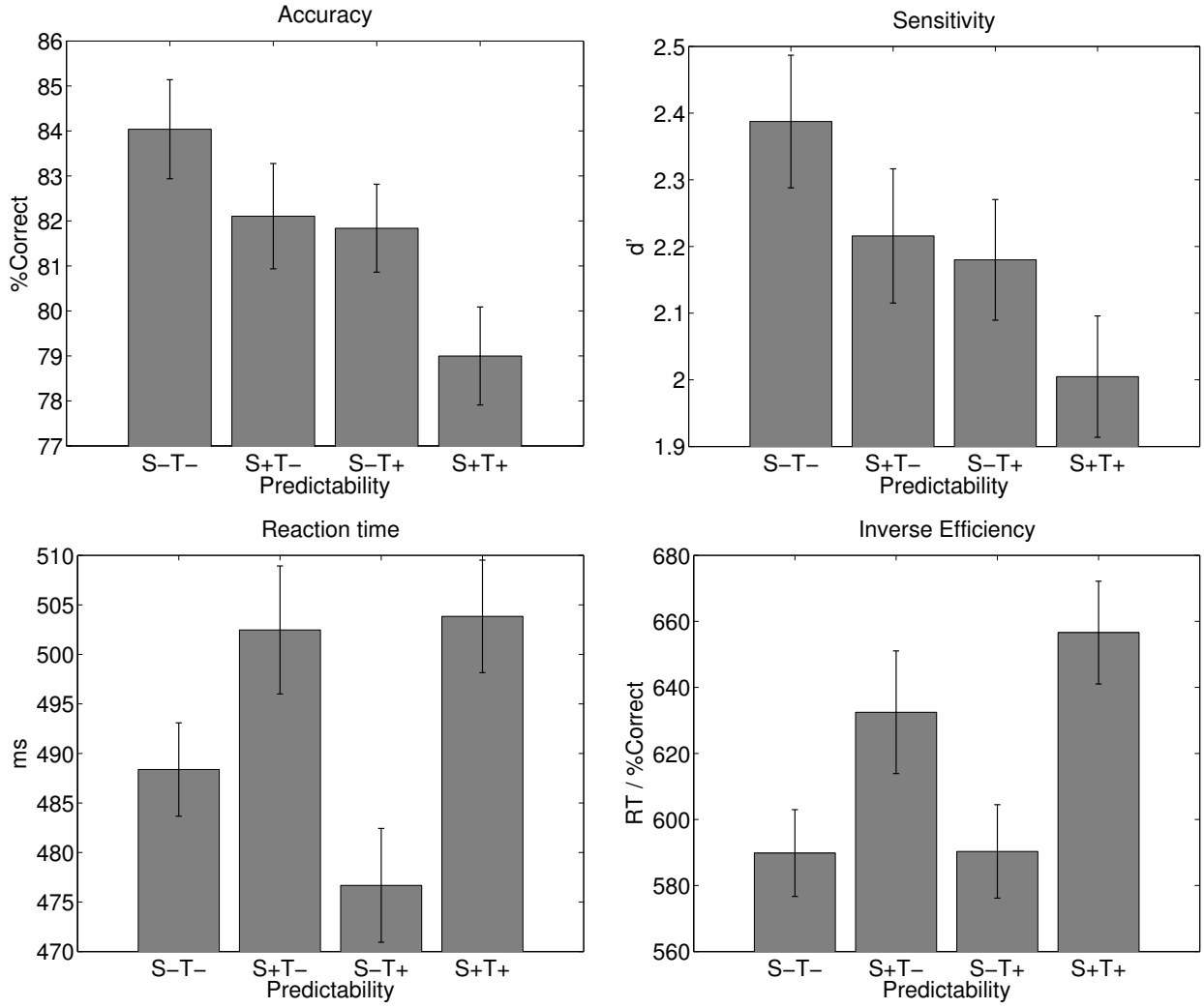

Figure 1.3: Behavioral measures of spatial and temporal predictability

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 Cousineau05) 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 (BalasSinha09b). 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

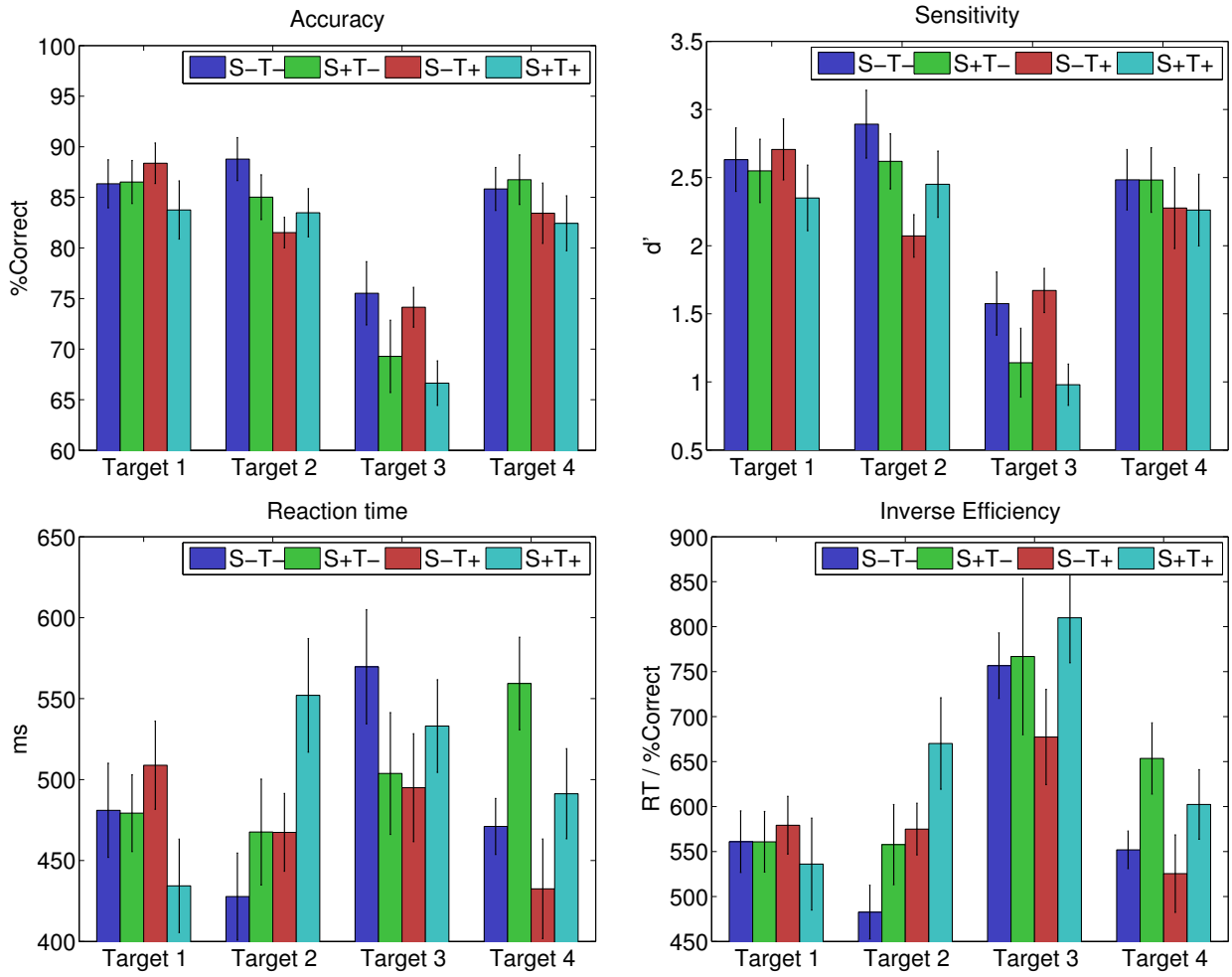

Figure 1.4: Behavioral measures for each target object

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

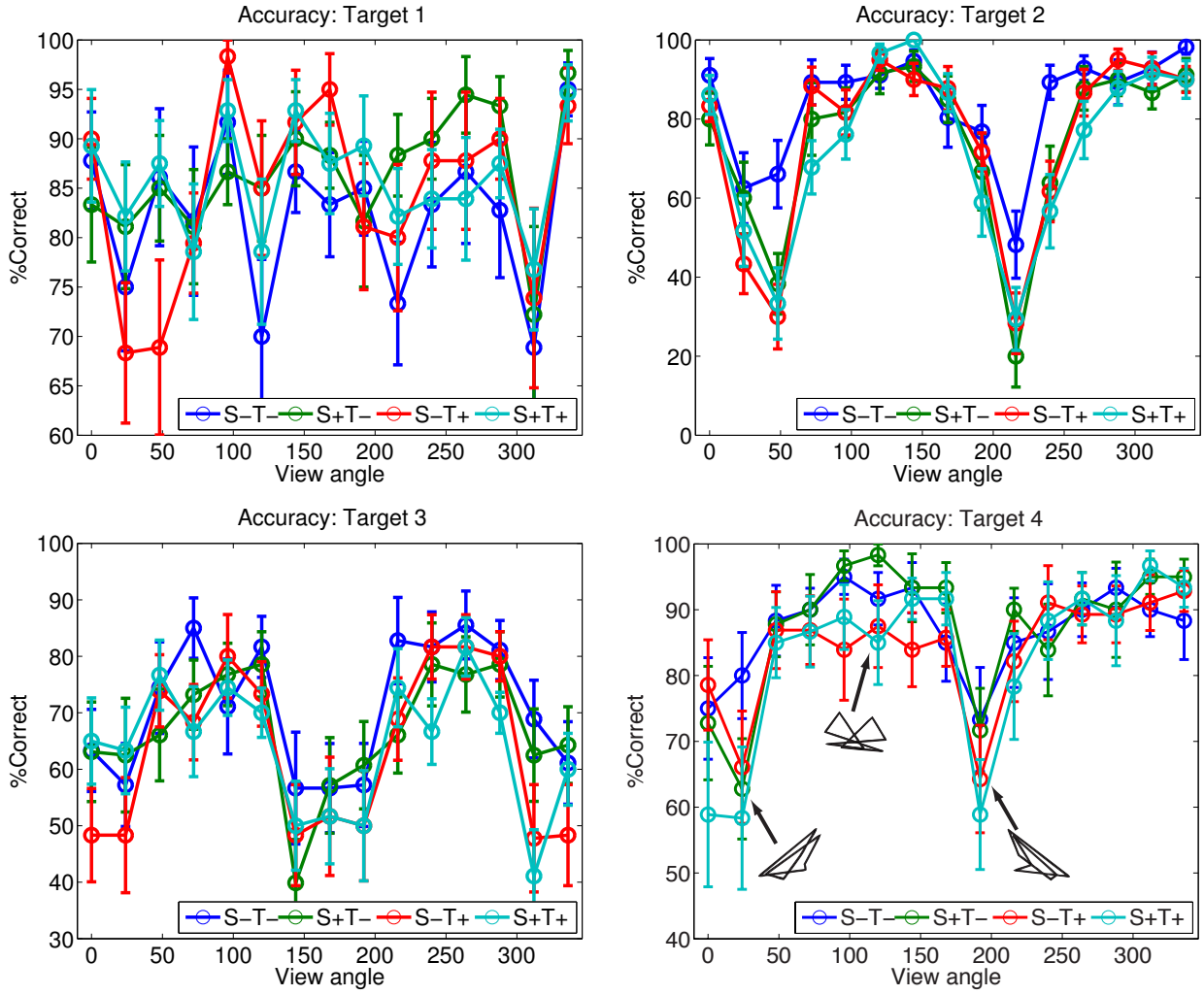

Figure 1.5: Accuracy as a function of viewing angle for each target object

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.

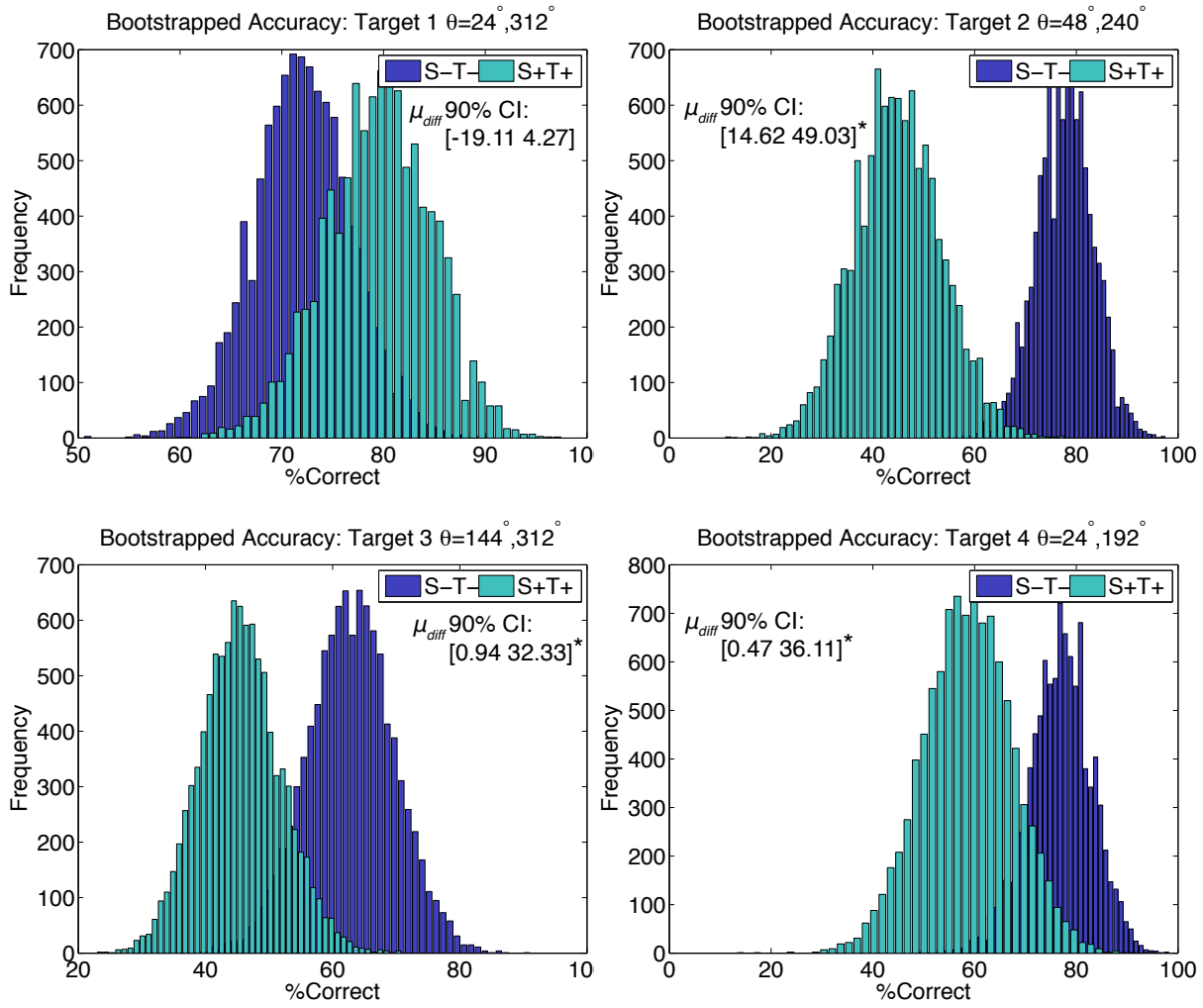

Figure 1.6: Bootstrapped accuracy for degenerate views

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 representations of novel objects over prolonged learning. The experimental paradigm used to address this question involved training participants to recognize novel objects while manipulating their

spatial and temporal predictability. 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. These findings were unexpected because the spatial structure of the physical world has been suggested to be a powerful learning mechanism when coupled with the putative temporal association rules of visual neurons (StringerPerryRollsEtAl06WallisBaddeley97IsikLeiboPoggio12SakaiMiyashita91MeyerOlson11CoxMeierOerteltEtAl05LiDiCarlo08

Behavioral measures were highly variable across objects used in the experiment. 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 LawsonHumphreysWatson94) 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 (Stone98VuongTarr04ChuangVuongBulthoff12), implying that learned predictability about the sequence is used to encode the object (BalasSinha09c). The foreshortening model advanced in (BalasSinha09b) 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 LawsonHumphreysWatson94), 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 HarmanHumphrey99) 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 (TarrGauthier98). 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 Building viewpoint invariance for three-dimensional objects

Given the cumulative literature considered here, the most reasonable interpretation of the experimental results is as follows: Three-dimensional objects are typically represented in a viewpoint-dependent manner (Wallis-Bulthoff99EdelmanBulthoff92TarrGauthier98LogothetisPaulsBulthoffEtAl94LogothetisPaulsPoggio95). Each of these views is associated with the given object's identity to the degree to which three-dimensional features are recoverable from the two-dimensional projection. This can lead to low accuracy for degenerate viewing angles caused by extreme three-dimensional foreshortening demonstrated here and in previous studies (BalasSinha09b).

When three-dimensional objects are studied in a full spatiotemporally predictable context, temporal association mechanisms are invoked that build new invariance (CoxMeierOerteltEtAl05LiDiCarlo08LiDiCarlo10LiDiCarlo12). This invariance accounts for variability due to fore-

shortening but is actually *not optimal* for three-dimensional objects that need to be recognized from individual static views. If the full spatiotemporal sequence is available when the object needs recognized (as is typical in real world object recognition) accuracy is not impaired by degenerate views and activation of the newly invariant features might even be facilitated (BalasSinha09c). However, if only static views are available during recognition, or perhaps worse, the full sequence is available but not in its spatially predictable order, it is not possible to activate these invariant features, leading to impaired recognition (Stone98VuongTarr04ChuangVuongBulthoff12).

The corollary of this interpretation is that when three-dimensional objects are studied in a completely unpredictable context, temporal association mechanisms are actually *not invoked*. Further research is needed to determine if this assertion is plausible, and if so, precisely why these mechanisms would not be active. However, one possibility is that the temporal associations over samples of an input sequence at regular intervals (e.g., 100 ms, see Chapter **??**). When input sequences don't align to these sampling intervals, the associations between subsequent inputs cannot be formed. The result is that despite prolonged learning, the objects remain represented in a viewpoint-dependent manner, which facilitates recognition for static views.