Learning sequence of views of three-dimensional objects: the effect of temporal coherence on object memory

Taosheng Liu

New York University

Please address correspondence to:

Taosheng Liu
Department of Psychology
New York University
6 Washington Place, 8th floor
New York, NY 10003

Email: taosheng.liu@nyu.edu
Fax: 212.995.4349
ABSTRACT

How humans recognize objects remains a contentious issue in current research on high-level vision. Here I test the proposal by Wallis and Bülthoff (1999) suggesting that object representations can be learned through temporal association of multiple views of the same object. Participants first studied image sequences of novel, 3-dimensional objects in a study block. On each trial, the images were from either an orderly sequence of depth-rotated views of the same object (SS), a scrambled sequence of those views (SR), or a sequence of different objects (RR). Recognition memory was assessed in a following test block. A within-object advantage was consistently observed—greater accuracy in the SR than the RR condition in all four experiments, greater accuracy in the SS than the RR condition in two experiments. Furthermore, spatiotemporal coherence did not produce better recognition than temporal coherence alone (similar or less accuracy in the SS compared to the SR condition). These results suggest that the visual system can use temporal regularity to build invariant object representations, via the temporal association mechanism.
People can easily recognize familiar objects from many different viewpoints under everyday conditions. How the visual system achieves such object constancy is a critical problem that any adequate theory of object recognition needs to account for.

One approach to achieving such object constancy is via viewpoint invariant representations, such as the geon structural description in Biederman’s recognition-by-component theory (Biederman, 1985; Hummel & Biederman, 1992). The use of abstract structural descriptions will naturally lead to viewpoint independent performance, as has been demonstrated in several studies experimentally (Biederman & Bar, 1999; Biederman & Gerhardstein, 1993).

However, a large body of literature exists now to question whether object representations are truly viewpoint-independent (e.g., Hayward & Tarr, 1997; Humphrey & Khan, 1992; Newell & Findlay, 1997; Tarr, 1995; Tarr & Bülthoff, 1995). Proponents of the view-based theories have proposed that the underlying representation is viewpoint specific. According to such theories, object representations are first established for particular learned views, and interpolation and extrapolation mechanisms are employed to recognize objects in novel viewpoints (e.g., the multiple-views theory, Bülthoff & Edelman, 1992).

Although the view-based theories have received some strong empirical support, one outstanding issue remains: if object representation is based on individual images, then how does the visual system ‘know’ which images belong to the same object? The fact that people can recognize multiple views of the same object suggests that some link between these different views must be established. One proposal by Wallis and Bülthoff (1999) suggests that such a link can be established via temporal association. That is,
different views over time can be associated to form invariant object representations. This *temporal association hypothesis* is based on the heuristic that whenever a series of temporally contiguous images occur, they are likely from the same object. Similar ideas about temporal association have been discussed in early studies of learning more than half a century ago (Pitts & McCulloch, 1947). The existence of temporal association was supported by a neurophysiological study (Miyashita, 1988), which found that neurons became selectively responsive to novel shapes in a particular fixed presentation sequence over many trials during learning. Temporal association has also been implemented in neural network models to achieve invariance in learning object representations (Edelman & Weinshall, 1991; Földiak, 1991; Wallis & Rolls, 1997).

Despite the theoretical and physiological work showing the role of temporal association in object recognition, behavioral evidence is somewhat sparse in the literature.

A relevant topic is the role of dynamic information in object recognition. For example, it has been shown that dynamic information can be also used to recognize biological movements (Johansson, 1973), facial identity (Knappmeyer, Thornton, Bülthoff, 2003; Thornton & Kourtzi, 2002), categories of novel objects (Newell, Wallraven, Huber, 2004), and identities of novel objects (Liu & Cooper, 2003; Stone 1998, 1999; Vuong & Tarr, 2004, 2006). The last set of studies demonstrated that a reversal in rotation direction impaired recognition of rotating objects, and suggested that observers can associate specific dynamic information—spatiotemporal signatures— to the identity of an object. Although relevant to the temporal association hypothesis, these
studies focus on the usefulness of dynamic information per se in object recognition, and says little about how static visual information can be associated by temporal proximity.

Another related finding is that naming familiar objects is easier when participants viewed structured sequences compared to random sequences of different object views (Lawson, Humphreys, & Watson, 1994). However, naming a familiar object relies on pre-existing representations and probably activates semantic associations, thus the results cannot be readily used to assess the temporal association hypothesis, which describes a process to build new object representations. To directly test the temporal association hypothesis, Wallis and Bülthoff (2001, also see Wallis, 2002) showed depth-rotated face images in a sequence that was composed of either the same individual’s face or faces from different individuals. They found that participants were impaired in discriminating different individuals’ faces in a later test block if those images were studied in the same sequence. This observation suggests that participants had associated the face images in a sequence as if they were from the same person, probably using the temporal association heuristic.

However, face recognition is special in the sense that recognition occurs on the subordinate level, instead of the basic level (Rosch et al., 1976), and it might involve specialized mechanisms (Palmeri & Gauthier, 2004). Furthermore, the discrimination task did not test for explicit recognition. To further test the temporal association hypothesis, it is necessary to generalize the results to other experimental situations (e.g., stimuli and task). Importantly, one such study using novel, 3D objects did not find any beneficial effect of temporal association on object recognition (Harman & Humphrey, 1999). In the experiments, participants first studied image sequences of depth-rotated objects in one of
the three conditions. On each trial, they saw either an ordered sequence of the depth-rotated views of the same object (SS), a scrambled sequence of those views (SR), or a random sequence of different objects (RR). Accuracy in a subsequent recognition task was not affected by the study manipulation, whereas reaction time was slightly faster for the random condition (RR) than the ordered condition (SS).

The authors treated these results with caution, pointing out some potential confounding factors in the experimental design. For example, 7 images were shown for 1 sec each on every trial during encoding. Furthermore, the similarity between views was relatively high due to a small rotation angle (20°) between adjacent views. These factors could potentially lead to less attention and encoding effort in the SS condition compared to the RR condition where every image was from a different object (a more ‘interesting’ condition from participants’ point of view). Thus the experimental conditions might not be optimal to reveal the effect of temporal coherence on object recognition.

The present experiments were conducted to test whether these results generalize to other experimental circumstances or whether the effect of temporal coherence on recognition can be observed under more favorable conditions. Experiments 1 and 2 used shorter exposure durations with reduced inter-view similarity to reduce potential confound in attention and effort; Experiment 3 used long exposure durations to mimic the conditions in Harman and Humphrey (1999); Experiment 4 employed an incidental encoding task to manipulate encoding effort. A within-object benefit was observed in all experiments, consistent with a temporal association mechanism during shape learning.
Experiment 1

As discussed above, the lack of effect of temporal coherence in Harman & Humphrey (1999) could be explained by a lack of attention and effort during encoding the temporally coherent sequences, as the long exposure time and high similarity between adjacent views in these sequences made them less interesting to participants. In this experiment, short exposure times and small inter-view similarity were used, to make the temporally coherent sequences less repetitive. Another rationale for using shorter exposures came from ecological considerations. In the real world, different views of an object are often seen in a rapid succession (e.g., when turning an object around in one’s hand). These are also scenarios under which temporal association is presumed to operate (Edelman & Weinshall, 1991; Földiak, 1991; Wallis & Bülthoff, 1999).

Methods

Participants.

Seventy-five undergraduate students at New York University participated in this experiment for course credit. All participants reported normal or corrected-to-normal vision and gave informed written consent. The experimental procedures were approved by the Institutional Review Board at New York University.

Visual Stimuli.

Thirty-two novel, 3D objects were designed and rendered on an Apple Macintosh G3 computer using custom-written C routines (see the Appendix for the full stimulus set). The objects contained only flat surfaces and were rendered as solids via depth shading without external light sources in the scene. The object image was fit into a window subtending a visual angle of 4.9°.
Design and Procedures.

There were three between-subject groups, corresponding to three different temporal coherence conditions in the study phase as described below, with 25 participants in each group. A study-test block design was employed for all groups in the experiment. A between-subject design was used because it is impractical to test long-term memory of novel objects in a within-subject design – this would entail learning many more novel shapes at once; the memory load is simply too high.

On each trial in the study block, 3 images were shown in succession. These images were taken from viewpoints on a viewing circle around the object. Figure 1 depicts the geometry of the view arrangement. For each object, 3 depth-rotated views were defined around a semi-circle, with 50° between adjacent views (corresponding to views 2, 4, 6 in Figure 1). All 3 views were shown during study in three levels of temporal coherence, as illustrated in Figure 2. In the first condition (Same object, Sequential view, or SS), 3 views from the same object were presented in ascending order one after another. In the second condition, the 3 views from the same object were shown in a scrambled order (Same object, Random view, or SR). In the third condition, 3 images of different objects were shown in different viewpoints (Random object, Random view, or RR). Each view was shown for 440 ms followed by an inter-stimulus interval.
Learning sequence of views

(ISI) of 440 ms\(^1\) (Figure 2). Participants were told that they would see a sequence of 3 images on every trial, and that they should remember the objects for a later memory test. They pressed a key to start each trial. The 16 objects (48 images) were shown twice in separate random orders. The total number of images shown in the 3 conditions during study was identical (48 images). In addition, there were 2 filler objects at the beginning of the study block.

Insert Figure 2 about here

Insert Figure 2 about here

The test block was the same for all three groups. At the end of the study block, participants were told that they were going to see one image on each trial and they should make an old/new judgment. They were informed that the studied objects would be tested only in their original orientations but that multiple views of the nonstudied objects would also be presented and they were to respond old only when an object appeared in the study phase. For both the studied and nonstudied objects, the 3 views in the middle of the viewing circle were tested (views 2, 4, and 6, see Figure 1). For each participant, 16 objects were randomly selected from the pool of 32 objects as studied objects, and the remaining served as nonstudied objects. All the trials in the test block were presented in a different random order for each participant. The participant pressed the space bar to start each trial, after which an image appeared and the participant responded old or new by pressing the ‘/’ or the ‘z’ key, respectively. Participants were instructed to respond as accurately and as quickly as possible, with the emphasis on accuracy. The object image

---

\(^1\) A pilot study was conducted to ensure that participants perceived no apparent motion in the SS condition at this ISI. With shorter ISIs, apparent motion might be perceived in the SS, but not (or less so) in the SR.
either remained on the screen until the response, or disappeared after 2 s, whichever came first. There were 4 filler objects at the beginning of the test block; responses to the filler items were excluded from data analyses. The whole experiment lasted about 25 min, after which participants were told the purpose of the experiment and provided with written explanations.

Data Analysis.

For all experiments, preliminary signal detection analyses were performed to derive sensitivity ($d'$) and criterion ($C$). No significant effect was found with criterion in any experiment, indicating that the experimental manipulation did not introduce systematic response bias. When accuracy is high, as occurred in certain individual participants, $d'$ measures tend to be unstable (Macmillan & Creelman, 2005). Hence the difference between hit and false alarm rates were used as the measure of recognition accuracy; results based on $d'$ and hit minus false alarm rates were in agreement with each other in all experiments. Although accuracy was the main dependent measure of performance, response latency was also analyzed for the sake of completeness. Only correct recognition trials ('hit') were used in this analysis, and reaction times greater than 3 standard deviations from the mean of each participant were removed from the analysis (<1% trials for all experiments). For all experiments, preliminary analyses including test view as a factor did not reveal any significant effect of test view and its interaction with study condition. Hence the results reported here were collapsed across test views.

Results and Discussion

Temporal coherence affected recognition accuracy in this experiment (Figure 3a). The effect of study condition was significant in one-way ANOVA ($F(2, 72) = 5.12$, MSE and RR conditions, which would introduce a potential confound in comparing these conditions.
Learning sequence of views

\(= .039, p < .01\). Post-hoc paired comparison (Tukey HSD test) revealed that accuracy in RR was significantly lower than that in both the SS and SR conditions (\(p < .05\)), with no significant difference between the latter two. There was no significant effect of study condition on reaction time (\(F < 1\), Figure 3b).

---

Insert Figure 3 about here

---

These results demonstrate that temporal association can indeed benefit object recognition under appropriate conditions. Compared to Harman & Humphrey (1999), both the inter-view similarity and total study time were reduced: the angular difference between views was 50° and the total exposure duration during study for each object was 2.64 s (440 ms per view, 3 views per object, 2 repeats), as opposed to 20° and 21 s (1 s per view, 7 views per object, 3 repeats) in Harman and Humphrey’s Experiment 1. This result is consistent with the idea that reduced study time and inter-view similarity made the SS and SR conditions less repetitive and tedious, and hence participants were presumably more attentive and effortful during study.

Interestingly, accuracy is similar, but not higher, for the SS compared to the SR condition, indicating spatiotemporal coherence did not further improve performance over temporal coherence alone. This observation suggests that a pure temporal association mechanism underlies the current results.

**Experiment 2**

Computational studies have demonstrated that temporal association could help form invariant object representations (Edelman & Weinshall, 1991; Földiak, 1991; Wallis...
& Rolls, 1997). Although Experiment 1 showed a beneficial effect of temporal coherence, the effect was restricted to the specific views during study, as the study and test views were identical. If temporal association indeed facilitates the formation of invariant representations, it should also support the recognition of novel views. Experiment 2 tested this prediction by testing recognition memory for novel views after studying sequences of varying temporal coherence.

Methods

Participants.

Seventy-five undergraduate students at New York University participated in this experiment for course credit. None of the participants had participated in Experiment 1.

Visual Stimuli.

Stimuli were identical to those in Experiment 1.

Design and Procedures.

The design and procedures were identical to those of Experiment 1 except for the object views shown in the test phase. In the test phase, 4 views of each object were shown, corresponding to views 1, 3, 5, 7 in Figure 1 – these were novel views that participants had not seen in the study phase. The instruction of the study phase also informed participants that the objects would be tested in novel viewpoints.

Results and Discussion

Temporal coherence facilitated recognition accuracy of novel views (Figure 4a), as reflected by a significant effect of study condition in one-way ANOVA (F(2, 72) = 5.64, MSE = .034, p < .01). Post-hoc paired comparison (Tukey HSD test) revealed that accuracy in RR was significantly lower than that in both the SS and SR conditions (p <
.05), with no significant difference between the latter two. There was no significant effect of study condition on reaction time (F < 1, Figure 4b).

---

The results resembled those from Experiment 1, i.e., both the SS and SR condition showed a higher recognition rate than the RR condition. Thus, a higher temporal coherence led to better recognition of both the studied views and novel views of an object. These results support the notion that temporal association facilitates the construction of invariant object representations.

Experiment 3

Results from Experiments 1 and 2 stand in contrast to those from Harman and Humphrey (1999), who did not find any effect of temporal coherence. In this experiment, the reasons for such a discrepancy were explored. As discussed above, one significant change from Harman and Humphrey’s experiments to Experiments 1 and 2 was the repetitiveness of the study sequence, as controlled by both the inter-view similarity and exposure duration. In Experiment 3, the repetitiveness of the study sequence was increased, to mimic the conditions in Harman and Humphrey (1999). If repetitiveness in the study sequence reduces participants’ attention and effort, recognition performance in the temporally coherent conditions should be reduced.

Methods

Participants.
Seventy-five undergraduate students at New York University participated in this experiment for course credit. None of the participants had participated in the previous experiments.

**Visual Stimuli.**

Stimuli were identical to those in Experiment 1.

**Design and Procedures.**

The design and procedures were identical to those of Experiment 1 except for a number of modifications of the study phase. As illustrated in Figure 5, on each trial in the study block, 7 images were shown, each for 1 s with an ISI of 750 ms. These images corresponded approximately to views 1-7 in Figure 1, except that there was a 20° separation between adjacent views (instead of 25°). The 16 objects (112 views) were shown 3 times in different random orders during the study phase. During the test phase, views 2, 4, 6 were shown in the *old/new* recognition task. All other aspects of the procedures were identical to those of Experiment 1.

Results and Discussion

Recognition accuracy was affected by the temporal coherence manipulation (Figure 6a). A one-way ANOVA with study condition (SS, SR, RR) as factor showed a significant effect ($F(2, 72) = 10.0$, $MSE = .045$, $p < .001$). Post-hoc paired comparisons (Tukey HSD test) indicated that accuracy in the SR condition was higher than that in the SS and RR condition ($p < .05$), which did not differ from each other. Mean reaction
times across participants are plotted in Figure 6b; there was no significant effect of study condition on reaction time (F(2, 72) = 1.84, MSE = .18, p > .15).

--------------------------------------
Insert Figure 6 about here
--------------------------------------

Compared to Experiment 1, overall recognition accuracy is higher, presumably due to increased study time. In addition, the pattern of results across conditions was also different: the SR condition led to the best recognition performance, while the SS condition was not statistically different from the RR condition, although the numerical trend was present.

Although the stimulus timing and viewing parameters were modeled on Experiment 1 of Harman and Humphrey (1999), a different pattern of results emerged. It is not clear what the cause might be. One possibility concerns the visual stimuli used: objects in Harman and Humphrey (1999) shared a common structure – all had a main axis of elongation with ‘geon’-like parts attached to the main body, whereas objects in the present experiment did not share an explicit structural design (see Appendix for the stimuli set). Structural similarity among objects might make temporal association less effective in constructing representations of individual objects, as views of different objects could be erroneously linked due to shared structural similarity. At present, however, this remains a conjecture and awaits further research that explicitly tests the effect of structural similarity on temporal association.

The decreasing trend in RT from SS to RR conditions resembled those of Harman and Humphrey (1999). However, RT data are hard to interpret when accuracy is not near
Learning sequence of views

ceiling (Pachella, 1974), and the effect of study condition on RT was not statistically significant in the current experiment. Nevertheless, the numerical trend was consistent with reduced attention and effort in the SS and SR conditions, leading to a slower RT during recognition.

Thus, there might be two opposing factors in this experiment: on the one hand, temporal association facilitated recognition for the SS and SR relative to the RR conditions; on the other hand, lack of attention and effort had the opposite effect. The SR condition yielded the best recognition performance perhaps because it did not suffer as much as the SS condition in terms of loss of attention and effort, as a jumbled sequence of the same object might be considered to be ‘more interesting’ than an ordered sequence.

Regardless of the presence of any confound related to attention and effort, these results clearly demonstrated a within-object advantage, that is, learning a sequence of images from the same object produced better memory than learning a sequence of images from different objects. This indicates that temporal coherence is indeed effective in building long-term memory representations of objects. The next experiment further explores the issue of attention and effort during encoding.

Experiment 4

Although a diminished attention and effort due to increased exposure time and inter-view similarity in Experiment 3 can explain the results in the SS condition, the experiment manipulated neither attention nor effort directly. Other stimulus-related factors may also account for the results as the physical stimuli (images) were not identical across these two experiments. In addition, the fact that the overall level of recognition performance changed across the two experiments may also complicate direct
comparisons across experiments. Another way to assess the effect of attention and effort is via changing task instructions while holding the stimulus presentations constant. In this experiment, stimulus presentations were identical to those in Experiment 1, but participants performed an incidental encoding task during study – they made an aesthetic judgment on the image sequences. The incidental task reduced encoding effort as participants were not actively committing the items to memory. Thus, if effort played a role in Experiment 3, one would expect a similar pattern of results in the present experiment, even with short exposure durations during study.

Methods

Participants.

Seventy-five undergraduate students at New York University participated in this experiment for course credit. None of the participants had participated in previous experiments.

Visual Stimuli.

Stimuli were identical to those in Experiment 1.

Design and Procedures.

The design and procedures were identical to those of Experiment 1 except for the instruction during the study block. Participants were told that they would see a sequence of three images on every trial, and that they should rate how much they liked a sequence as a whole on a 3-point scale. They pressed one of three keys to indicate their response – like, neutral, or dislike. The criteria for the likeness judgment were left to participants (for another example of the task, see Rock & Gutman, 1981). All other aspects of the procedures were identical to those of Experiment 1. At the end of the experiment,
participants were queried as to whether they knew during the study phase that their memory would be tested later.

Results and Discussion

Post-experiment inquiry confirmed that no participant had the foreknowledge during encoding that their memory would be tested. The rating response was coded as follows: 1 = dislike, 2 = neutral, 3 = like. For each participant, a mean rating was calculated and the group averaged rating for the SS, SR, and RR condition was 2.04±.06, 2.02±.05, 2.02±.05 (mean±s.e.m.), respectively. There was no significant effect for the ratings in one-way ANOVA (F < 1). These results suggest that participants were more or less neutral toward a sequence.

For the recognition data in the test phase, once again, temporal coherence affected recognition accuracy (Figure 7a). Study condition had a significant effect on recognition accuracy (F(2, 72) = 6.24, MSE = .019, p < .01). Post-hoc paired comparison (Tukey HSD test) revealed that the only significant difference was between the accuracy of SR and RR conditions (p < .05). There was no significant effect of study condition on reaction time (F < 1, Figure 7b).

Experiment 4 was identical to Experiment 1 except the task at study. Yet the pattern of the results resembled more to that in Experiment 3 than to Experiment 1: highest accuracy for SR, followed by SS and RR. The overall level of recognition performance was also comparable between Experiments 4 and 1. Thus reducing the
encoding effort by performing an incidental task had an analogous effect on recognition memory with repeating the study items many times in highly similar views. This suggests that the lower accuracy for the SS relative to the SR condition observed in Experiments 3 and 4 was partially due to reduced effort during encoding.

Importantly, both the SS and SR conditions yielded higher recognition accuracy than the RR condition, with the SR significantly higher than the RR condition. This result shows that the benefit of temporal association on recognition is robust; it occurred under conditions where participants did not intentionally learn the visual objects.

General Discussion

The present results demonstrate that studying temporally coherent sequences of object views produced higher recognition performance than studying random sequences, as reflected in higher accuracy for the SS and SR conditions than the RR condition in all four experiments (only a numerical trend was observed in the comparison between SS and RR in two out of four experiments). This result is consistent with the temporal association hypothesis, according to which views appearing close in time can be associated to form object representations (Wallis & Bülthoff, 1999). Moreover, the advantage of the SS over RR was dependent on manipulations of the exposure time and participants’ intention during study, variables that presumably affected the amount of encoding effort.

In a previous study with a similar design, Harman and Humphrey (1999) did not find any beneficial effect of temporal coherence on recognition. These authors attributed their results to confounds in attention and effort. Because the exposure time was long (21
s) and the angular separation between views was small (20°), there was a high degree of information redundancy in the SS and SR sequences, which could potentially result in less attention and encoding effort. In Experiments 1 and 2 in the present study, such redundancy was reduced with shortened exposure time (< 3 s) and smaller inter-view similarity (50° angular separation). Shorter exposure times are also more ecologically relevant, given that we tend to see different views of an object in a relatively rapid sequence, when the temporal association mechanism is assumed to operate (Edelman & Weinshall, 1991; Földiak, 1991; Wallis & Bülthoff, 1999). In Experiments 1 and 2, these manipulations led to better recognition for both SS and SR compared to the RR condition. It is worth noting that Experiment 3 in Harman and Humphrey (1999) used a shorter total duration (12 s) and large angular separation between views (60°). In addition, the ISI between views was short such that participants perceived apparent motion. Still, the authors did not observe any benefit in recognition for the SS over RR condition. Thus the 12 s duration might still be too long or the perception of motion itself impaired the recognition of individual views for the SS condition.

When exposure duration and inter-view similarity were increased in Experiment 3, recognition performance in the SS condition became significantly lower than that in the SR condition, and statistically indistinguishable from that in the RR condition. This was consistent with an explanation based on encoding effort (Harman & Humphrey, 1999). Experiment 4 directly manipulated encoding effort by using an incidental task during study and found similar results as in Experiment 3, even with short exposure times and small inter-view similarity. These results suggest that encoding effort or intention might be a factor in determining performance in the SS condition in Experiments 3 and 4.
Learning sequence of views

The jumbled sequence in the SR condition may be more interesting to participants, counteracting the effect of redundancy in the image sequence, and consistently produced superior memory performance across all experiments.

Even when confounds in encoding effort were minimized (Experiments 1 and 2), performance in the SS condition did not exceed that in the SR condition, although the SS sequence maintained spatial, in addition to temporal coherence. Such a result suggests the existence of a pure temporal association mechanism, at least in the context of the present experimental conditions. This interpretation is consistent with an earlier neurophysiological study (Miyashita, 1988), which demonstrated that neurons in inferior temporal cortex became responsive to stimuli that were presented in close temporal order even though they were quite different shapes. However, Wallis and Bülthoff (2001, see also Wallis, 2002) obtained a different result using face images in a confusion/discrimination paradigm. They found temporal proximity alone was not sufficient to produce association, measured by the confusion of face identity (Experiment III, equivalent to the SR condition here). Instead, both spatial and temporal proximity was required to produce association (Experiment I, equivalent to SS condition here).

One possible explanation for this discrepant finding pertains to the stimuli and levels of recognition. In Wallis and Bülthoff’s experiments, participants made facial identity discrimination, which is a subordinate-level (as opposed to basic level, Rosch et al., 1976) task entailing more fine-grained discrimination than the present experiments. It is conceivable that spatiotemporal proximity is necessary to influence recognition at this level, but temporal proximity alone is sufficient to influence recognition of more distinct objects. Furthermore, for familiar stimuli like faces, sequences with spatiotemporal
proximity might also contain more meanings than sequences with only temporal proximity, given that we have a lot of prior experience with facial movements. Another possibility concerns the tasks used in Wallis and Bülthoff (2001, also Wallis, 2002) and the present study. In the former studies, the effect of association was manifested as an impairment in discrimination performance. Such impairment would only occur if participants fully associated faces of different individuals, which seems to be a stringent test of the formation of associations. The present study, on the other hand, allows any effect of association to emerge as a facilitation in performance. Thus the task in the present study might be more sensitive to the effect of temporal association. Further studies are needed to shed more light on how temporal association is influenced by stimulus type and task.

Conclusion

The present study demonstrates that temporal coherence improves recognition memory for novel, 3D shapes. The results support theoretical arguments for the role of temporal association in object recognition (Wallis & Bülthoff, 1999). Temporal proximity likely constitutes another source of invariance that the visual system can exploit for learning.
References

Biederman I, 1985 “Human image understanding: Recent research and a theory” 
*Computer Vision, Graphics, & Image Processing* **32** 29-73

Biederman I, Bar M, 1999 “One-shot viewpoint invariance in matching novel objects” 
*Vision Research* **39** 2885-2899


Miyashita Y, 1988 “Neuronal correlate of visual associative long-term memory in the primate temporal cortex” Nature 335 817-820


Pitts W, McCulloch W S, 1947 “How we know universals: the perception of auditory and visual forms” Bulletin of Mathematical Biophysics 9 127-147


Vuong Q C, Tarr M J, 2004 “Rotation direction affects object recognition” *Vision Research* **44** 1717-1730

Vuong Q C, Tarr M J, 2006 “Structural similarity and spatiotemporal noise effects on learning dynamic novel objects” *Perception* **35** 497-510


Author Note

I thank Emily Langner, Alissa Cantone, and Megan Yoburn for testing the participants and Dr. William Hayward for helpful comments on an earlier version of the paper. I also thank Dr. Marisa Carrasco for helpful discussions and support for this research.

Correspondence should be addressed to Taosheng Liu, Department of Psychology, New York University, 6 Washington Place, 8th floor, New York, NY 10003 (Email: taosheng.liu@nyu.edu).
Figure captions

Figure 1. The geometry of the viewing circle. Viewpoints 1-7 are depicted on a circle in the \(xz\) plane (see the inset for the designation of 3 axes), with a 25° separation between views. The bottom row illustrates the appearance of an object in viewpoints 1-7.

Figure 2. Temporal structures of the three presentation conditions: SS, SR, and RR. V2, V4, and V6 are the 3 views depicted in Figure 1. The images in SS and SR are from the same object, whereas images in RR are from different objects (indicated by the different subscripts \(i, j, k\)), although some of the images might be from the same object by chance. The view sequences depicted in SR and RR conditions are only two out of many possible sequences.

Figure 3. Results for Experiment 1. a) Mean recognition accuracy (plotted as hit minus false alarm rates) for each condition. b) Mean response latency of ‘hit’ trials for each condition. Error bars are ±1 s.e.m., and asterisks indicate significant difference in paired comparisons (\(p < .05\)).

Figure 4. Results for Experiment 2, same notation as in Figure 3.

Figure 5. Temporal structures of the three presentation conditions in Experiment 3: SS, SR, and RR. Notations are similar to Figure 2. Note that the views approximately correspond to views 1-7 in Figure 1, but with a 20° separation between views.
Figure 6. Results for Experiment 3, same notation as in Figure 3.

Figure 7. Results for Experiment 4, same notation as in Figure 3.

Appendix: 32 objects used in the current experiments (only one view is shown for each object).
Learning sequence of views

Study views: 2, 4, 6
Exp 1 test views: 2, 4, 6
Exp 2 test views: 1, 3, 5, 7

Figure 1
Figure 2
Figure 3
Learning sequence of views

Figure 4
Learning sequence of views

Figure 5
Learning sequence of views

Figure 6
Learning sequence of views

Figure 7
Learning sequence of views

Appendix Figure