Space as a Scaffold for Temporal Generalisation

Jacques Pesnot Lerousseau (jacques.pesnotlerousseau@psy.ox.ac.uk)
Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom

Christopher Summerfield (christopher.summerfield@psy.ox.ac.uk)
Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom

Abstract

Object recognition relies on invariant representations. A longstanding view states that invariances are learned by explicitly coding how visual features are related in space (Biederman, 1987; Marr, 1982). Here, we asked how invariances are learned for objects that are defined by relations among features in time (temporal objects, TO). We trained people to classify auditory, visual and spatial temporal objects composed of four successive features into categories defined by sequential transitions across a two-dimensional feature manifold, and measured their tendency to transfer this knowledge to categorise novel objects with rotated transition vectors. Rotation-invariant temporal objects could only be learned if their features were explicitly spatial or had been associated with a physical spatial location in a prior task. Thus, space acts as a scaffold for generalising information in time.

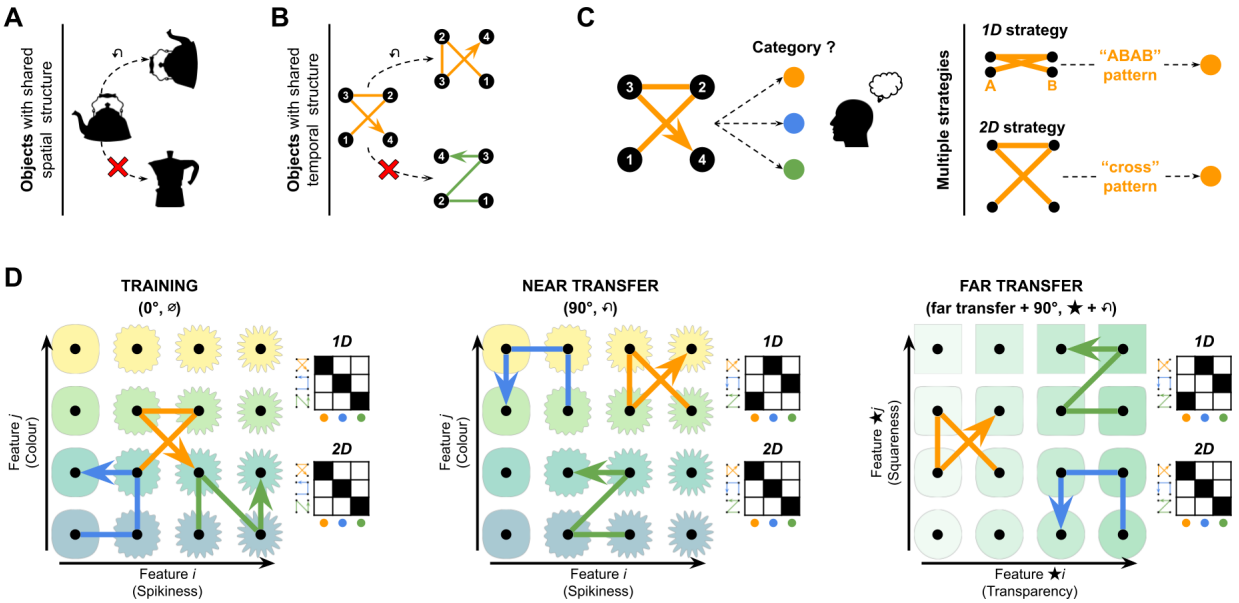
Main text

We operationalised TOs as a sequence of four auditory, visual or spatial stimuli (a quadruplet) drawn from one of 16 points on a continuously varying 2D feature manifold (e.g., spikiness and colour in the visual domain; see Fig. 1D). We define a "temporal object invariance" as a regularity in the TO transition vectors that is independent of both translation and rotation.

Figure 1. A. An upside-down teapot can still be recognised by the relative spatial relations among its handle, body, and spout. **B.** Here, we asked whether TOs can be recognised when their associated transition vectors are rotated. **C.** Each TO was constructed by first sampling a random point on the 2D feature manifold, and then iteratively choosing three further adjacent feature pairs to generate the quadruplet. Each category was defined by a set of transition vectors. **D.** During training, participants learned to assign canonical (0° rotation) TOs to one of three categories using a button press, receiving fully informative feedback after each response. During transfer, participants performed transfer trials involving categorisation of TOs whose feature transition vectors were rotated by 90°, 180° or 270°. These novel TOs were either sampled from the same 2D feature manifold (near transfer condition) or a new 2D feature manifold in the same modality (far transfer condition). Transfer trials received no feedback, allowing us to infer what knowledge was being generalised between training and transfer.

Temporal objects defined by spatial locations, but not auditory or visual features, are rotation-invariant.

In Exp. 1, we recruited three cohorts of online participants (N = 50 each) to perform the task in the auditory (Exp. 1a), visual (Exp. 1b) and spatial (Exp. 1c) modalities. These conditions differed only in how the feature manifold was defined: e.g., fundamental frequency and modulation frequency for auditory features; e.g., spikiness and colour for visual features; e.g., horizontal and vertical position for spatial locations. Our main question was how participants would generalise learning to novel, rotated temporal objects. To test this, we fit a family of models to the transfer trials which variously assumed that participants had learned to map from features to categories using a single dimension (1D



models), both dimensions (2D model), or were simply responding randomly (R models). The 1D models assume that participants map events onto categories using either (but not both) of the two dimensions, forming a representation that resists invariance to rotation. By contrast, the 2D model predicts that participants will assign rotated objects to the same category as their unrotated counterparts, indicating that they have learned a rotation-invariant representation. For brevity, we only present data in near transfer here, but all the major differences between conditions were replicated in far transfer trials (unreported data).

Exp. 1 revealed a striking dissociation between modalities in generalisation over rotation: all non-random participants in the auditory and visual modality learned a 1D generalisation strategy, whereas the vast majority in the spatial modality were best fit by a 2D generalisation strategy (Bayes Factor in favour of a difference in model frequencies between groups [BF] > 100).

Spatial pre-training provides a scaffold to learn rotation-invariant TOs in the auditory and visual modalities.

Next, in Exp. 2 and 3, we tested the prediction that space can be used as a scaffold for learning rotation invariances for non-spatial TOs. We recruited new cohorts of participants (N = 50 each) to perform a multi-phase task that unfolded over two days. On day 1, participants received pre-training trials in the pre-training modality. These trials matched those in Exp. 1 for the corresponding modality (spatial or visual) except that they comprised both canonical (0°) TOs and 90° rotated TOs, but not those rotated by 180° or 270°. We included examples of rotated TOs in the training set in Exp. 2 to encourage generalisation, but as shown in Exp. 3 (which did not include rotated TOs in pre-training), results do not depend on this choice. Subsequently, participants performed a multimodal association task, in which they learned the association between each of the 16 stimuli in the pre-training modality and their corresponding stimulus in a different testing modality, where the corresponding stimulus occupied an equivalent position on the 2D feature manifold (mapping task). Then on day 2, after some refresher pre-training and mapping trials, participants performed the same task as in Exp. 1 in the testing modality, again with the exception that training trials also included 90° rotated temporal objects (in both Exp. 2 and 3).

Our prediction for Exp. 2 and 3 was that when space was the pre-training modality, participants would now learn using a predominantly 2D strategy in both auditory (Exp. 2a and 3a) and visual (Exp. 2b and 3b) testing modalities. In other words, by learning the association between auditory or visual features and a corresponding spatial location, TOs composed of exclusively auditory or visual features could now be generalised over rotations in a way not exhibited by a single participant in Exp. 1a or Exp. 1b. By way of control, however, we predicted that when the pre-training involved visual features, no such benefit would occur, and participants would fail to show rotation invariance (Exp. 2c and 3c).

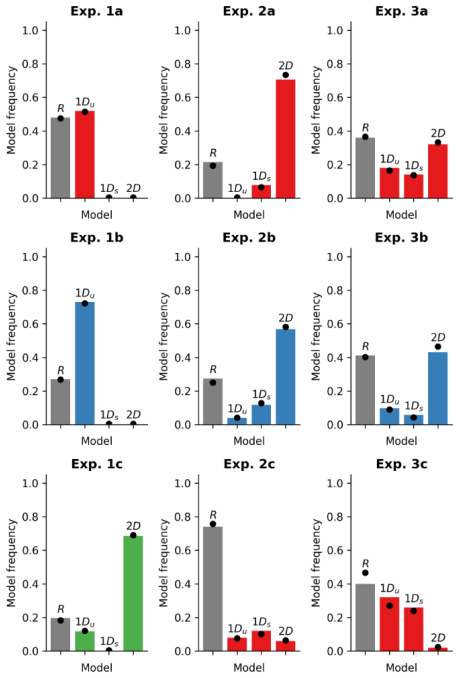


Figure 2. Model fits in “near transfer” trials. Bars are model frequencies in the cohort (N = 50 each). Colours correspond to the testing modality (red: auditory, blue: visual, green: spatial).

This is exactly what we found: with spatial pre-training, almost all non-random participants were best fit by a 2D strategy when audition was the testing modality and when vision was the testing modality. By contrast however, participants who underwent visual pre-training failed to show a benefit when audition was the testing modality (BF > 100 in Exp. 2 and 3). As predicted, participants who enjoyed spatial pre-training were still prone to use a 2D strategy when audition was the testing modality as well as when vision was the testing modality. Thus, spatial pre-training provides a scaffold to learn rotation-invariant representation of auditory and visual temporal objects even when rotation is never explicitly shown during pre-training.

Acknowledgments

We thank Jean Daunizeau for technical help with modelling. Work supported by Fondation Pour l'Audition FPA RD-2021-2 (J.P.L.) and European Research Council Consolidator Grant n° 725937 – CQR01290.CQ001 (C.S.).

References

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, *94*(2), 115–147. https://doi.org/10.1037/0033-295X.94.2.115

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (W. H. Freeman, Ed.).