The role of semantics in similarity judgements of scene stimuli

Katerina Marie Simkova[a], Jasper van den Bosch[a], Ian Charest[a, b]

[a]School of Psychology, University of Birmingham, [b]cerebrUM, Département de Psychologie, Université de Montréal

contact: k.m.simkova@bham.ac.uk

Introduction

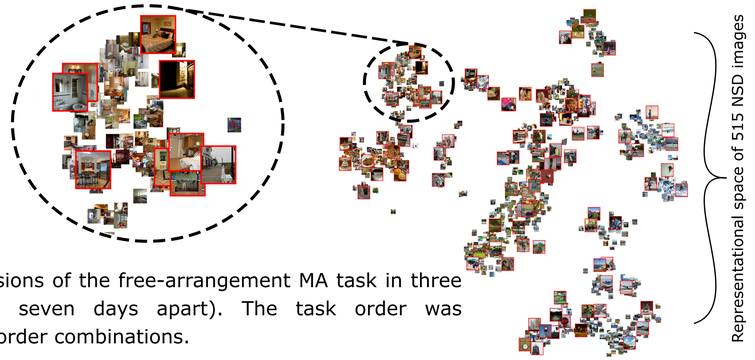
Classic accounts of human vision suggest functional segregation in two distinct cortical pathways: a "where" and a "what" pathway[1, 2], a dorsal stream specialised in spatial information ("where") and a ventral stream for category[3–6] or conceptual information[7]. The ventral stream could therefore be seen as a distributed system where overlapping feature maps[8] encode specific dimensions about particular objects[9, 10]. This view might stem from simple experimental paradigms involving single objects or simplified stimuli, but it is not clear whether this view accurately generalises to naturalistic viewing situations with complex scenes containing multiple objects and concepts.

Here we set out to investigate how human observers perceive semantic information that is being communicated via visual vs lexical stimuli using the multiple arrangements (MA) task: we isolated the semantic information by adapting the MA method into three different task formats - arrangements of images, hidden images, and captions - and explored how the similarity judgements relate to state-of-the-art deep neural networks and brain data.

Methods

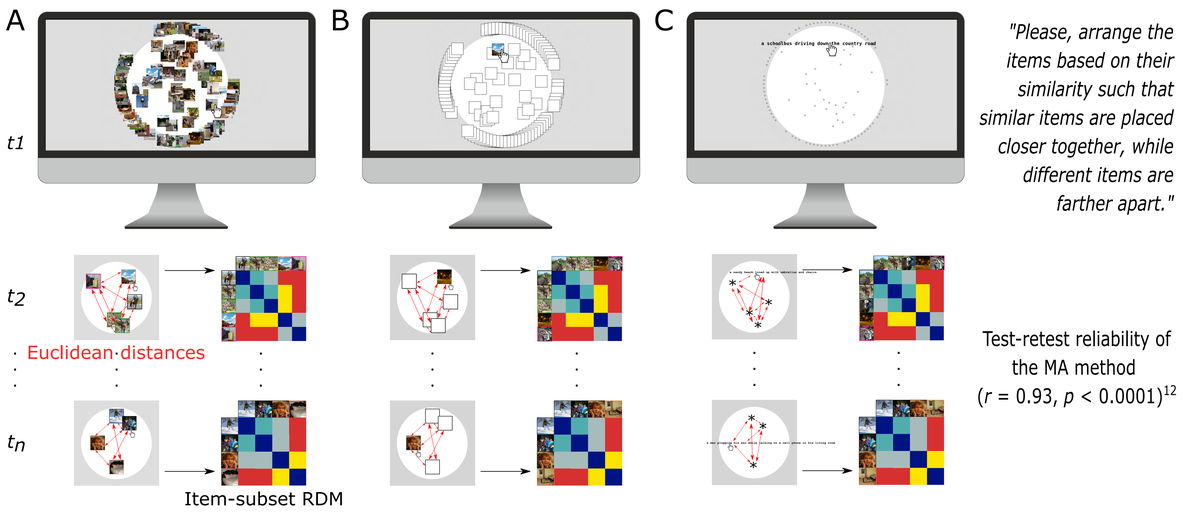
Stimulus set

100 images (in red boxes) were sampled from the Natural Scenes Dataset (NSD)[11] so that they span the full representational space (as characterised from GUSE) of the NSD shared 515 images.



Multiple arrangements task

Participants (n = 63) completed three versions of the free-arrangement MA task in three lab-based sessions (scheduled at least seven days apart). The task order was counterbalanced yielding six possible task order combinations.



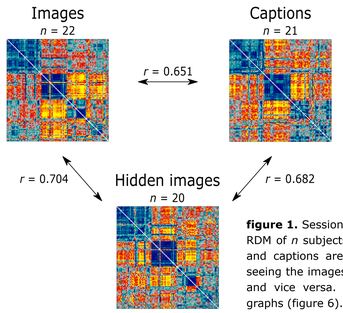
"Please, arrange the items based on their similarity such that similar items are placed closer together, while different items are farther apart."

Test-retest reliability of the MA method ($r = 0.93$, $p < 0.0001$)[12]

The relative Euclidean distances between items were iteratively adjusted by the weighted-averaging approach: subsets of stimuli carrying the weakest weighted evidence were presented in subsequent n trials - reducing potential placement errors and yielding the final subject-level task-specific RDM. Trials terminated once 45 minutes elapsed.

Results

Task order effects

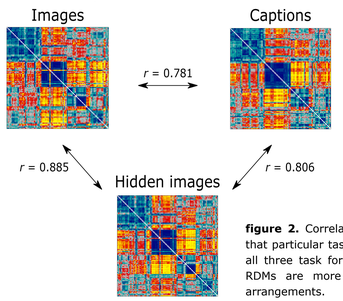


First session images and second session captions MA $r = 0.687$

First session captions and second session images MA $r = 0.646$

figure 1. Session 1 RDMs correlated for all three task formats (mean RDM of *n* subjects for that particular task). Additional *r*'s for images and captions are included to control for the arrangement bias of seeing the images first (session 1) and sentences second (session 2) and vice versa. The task order effect is more obvious from bar graphs (figure 6).

Overlap between tasks

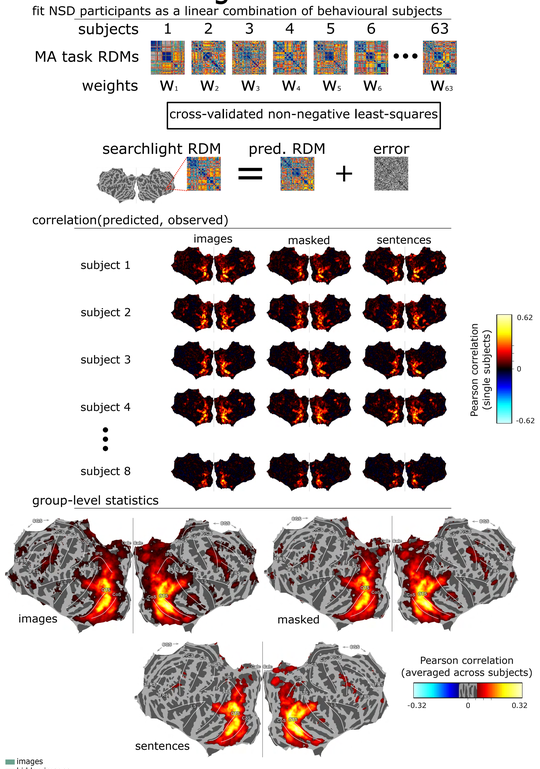


Within-subject $r = 0.268$

Between-subject $r = 0.166$

figure 2. Correlation of task RDMs (mean RDM of all subjects for that particular task) to explore the representational overlap between all three task formats. Additional *r*'s to compare whether the task RDMs are more likely to resemble between- or within-subject arrangements.

fMRI searchlight RSA



Semantic features vs visual features

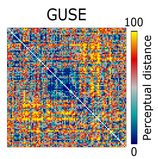


figure 3. RDM after feeding our stimulus set of 100 captions through the Google Universal Sentence Encoder (GUSE). *R*'s show correlation of GUSE with each task format.

Captions RDM $r = 0.407$

Masked images RDM $r = 0.339$

Images RDM r = 0.327

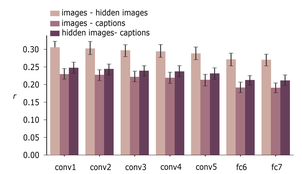


figure 4. Partial task-task RDM correlations with AlexNet layers as covariates. The analysis is based on all sessions.

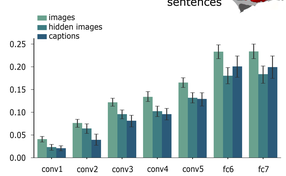


figure 5. Task RDM correlations with AlexNet layers based on first session only.

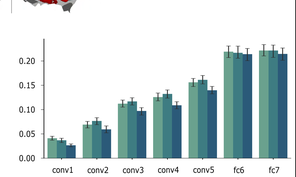


figure 6. Task RDM correlations with AlexNet layers based on all sessions.

Discussion

Here we investigated to what level human observers rely on purely visual versus semantic information when performing similarity judgements. Although the task order effects in the representational overlap is evident, analyses based on first session RDMs still reveal a significant level of correspondence between the tasks. The partial task-task RDM correlation with AlexNet layers show that the arrangements are not entirely dependent on visual features. This suggests that MA is an efficient method in capturing semantic level representations.

Moreover, the fMRI searchlight analyses, using a linear combination of behavioural participants to predict NSD participants, revealed strikingly similar brain regions across the three MA tasks. Altogether this suggests that the representations in the visual ventral stream share a representational format with visual and lexical based similarity judgements.

References

[1]Ungerleider, L. G. & Haxby, J. V. (1994). "What" and "where in the human brain. Curr. Opin. Neurobiol.
[2]Goodale, M. A. & Milner, A. D. (1992). Separate visual pathways for perception and action. Trends Neurosci.
[3]Kanwisher, N. et al. (1997). The Fusiform Face Area: A module in human extrastriate cortex specialized for face perception. J. Neurosci.
[4]Haxby, J. V. et al. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science.
[5]Kriegeskorte, N. et al. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron.
[6]Charest, I. et al. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity for real-world objects. PNAS
[7]Clarke, A. & Tyler, L. K. (2015). Understanding what we see: how we derive meaning from vision. TiCS.
[8]Op de Beeck, H. P. et al. (2008). Interpreting fMRI data: maps, modules and dimensions. Nat. Rev. Neurosci.
[9]Cichy, R. M. et al. (2019). The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. NeuroImage.
[10]Hebart, M. N., et al.(2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. Nat Hum Behav.
[11]Allen, E. J. et al. (2022). A massive 7-t fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nature Neuroscience.
[12]Kriegeskorte, N. & Mur, M. (2012). Inverse MDS: inferring dissimilarity structure from multiple item arrangements. Front. Psychol.