Revealing the Time-Course of Mid-Level Feature Representations in Scenes Using Rendered Stimuli and Ground-Truth Annotations

***Agnessa Karapetian (agnek95@zedat.fu-berlin.de)**
Department of Education and Psychology, Freie Universität Berlin,
Habelschwerdter Allee 45, 14195 Berlin, Germany

***Alexander Lenders (lxlenders@gmail.com)**
Department of Education and Psychology, Freie Universität Berlin,
Habelschwerdter Allee 45, 14195 Berlin, Germany

Vanshika Bawa (vanshikabawa9@gmail.com)
Faculty of Biology, Albert-Ludwigs-Universität Freiburg,
Schänzlestr. 1, 79104 Freiburg, Germany

Martin Pflaum (contact@martinpflaum.com)
RWTH Aachen University,
Templergraben 55, 52062 Aachen, Germany

Raphael Leuner (raphael.leuner@web.de)
Department of Mathematics and Computer Science, Freie Universität Berlin,
Takustraße 9, 14195 Berlin, Germany

Gemma Roig (roig@cs.uni-frankfurt.de)
Department of Computer Science, Goethe University Frankfurt
Robert-Mayer-Str. 11-15, 60325 Frankfurt, Germany

***Kshitij Dwivedi (kshitijdwivedi93@gmail.com)**
Department of Computer Science, Goethe University Frankfurt
Robert-Mayer-Str. 11-15, 60325 Frankfurt, Germany

***Radoslaw M. Cichy (rmcichy@zedat.fu-berlin.de)**
Department of Education and Psychology, Freie Universität Berlin,
Habelschwerdter Allee 45, 14195 Berlin, Germany

*****indicates equal contribution

Abstract:

Scene perception is a key function of the human visual brain that follows a hierarchical processing stream from low- to mid- to high-level features. While the processing of low- and high-level features is well-researched, mid-level features and their temporal dynamics are still under-investigated, partly due to a lack of appropriate stimuli to probe them. To address this gap, we used a rendering software to create a rich stimulus set of images and short videos of scenes in which persons perform different actions. We also obtained the corresponding ground-truth annotations for five postulated mid-level features (reflectance, lighting, world normals, scene depth and skeleton position), as well as one low-level feature (edges) and one high-level feature (action). We collected electroencephalography (EEG) data during the presentation of these stimuli and applied encoding models to predict the EEG data from the ground-truth feature annotations. We observed that the encoding accuracy of our mid-level feature annotations peaked between ~100 ms and ~250 ms after stimulus onset, framed by the low- and high-level feature representations. This suggests that the postulated mid-level features play an intermediary role in the transformation of low-level inputs into high-level semantic information, providing insight into their place in the scene processing hierarchy.

Keywords: scene perception; mid-level features; EEG, encoding.

Introduction

Humans come to interact with the world by processing scene information of their immediate surroundings, starting from low-level features of objects (e.g., edges) and culminating with high-level semantic features (e.g., actions) (Groen et al., 2017). However, the temporal dynamics of the computations between low- and high-level features, i.e. the mid-level features, are incompletely understood, in part due to the lack of appropriate stimuli to probe mid-level features. Here, we addressed this challenge by using a 3D rendering software to create a stimulus set of naturalistic scenes and their ground-truth annotations for visual features. We considered five mid-level features based on theoretical models of object recognition (Biederman, 1987; Marr, 1982) and computer vision literature (Zamir et al., 2018): reflectance, lighting, world normals, scene depth and skeleton position. To frame our results, we further considered the low-level feature edges and the high-level feature action. Additionally, we created both images and short videos (300 ms) to explore the role of stimulus input in mid-level feature processing. Using these stimuli, we collected EEG data from human participants during stimulus presentation. Afterwards, we applied encoding models (Kriegeskorte & Douglas, 2019; Naselaris et al., 2011) to predict the EEG data from ground-truth feature annotations at every time point, thereby revealing the time courses with which low-, mid-, and high-level visual features emerge.

Methods

Stimulus Set and Data Collection

We collected EEG data from human participants while they performed a target detection task during the viewing of scene images (N=15) and 300-ms videos (N=20) from a stimulus set that we created in a game engine (Epic Games, 2019) (Fig. 1). The stimulus set was composed of 1440 rendered scene images and videos and of ground-truth visual feature annotations for every frame, for five mid-level features (low- and high-level features were computed separately). The stimuli were created by sampling from 3 characters, 20 rooms, 4 camera angles, and 6 actions. For the encoding analyses, the stimuli were split based on the rooms into training, test and validation sets, each containing respectively 1080, 180 and 180 samples.

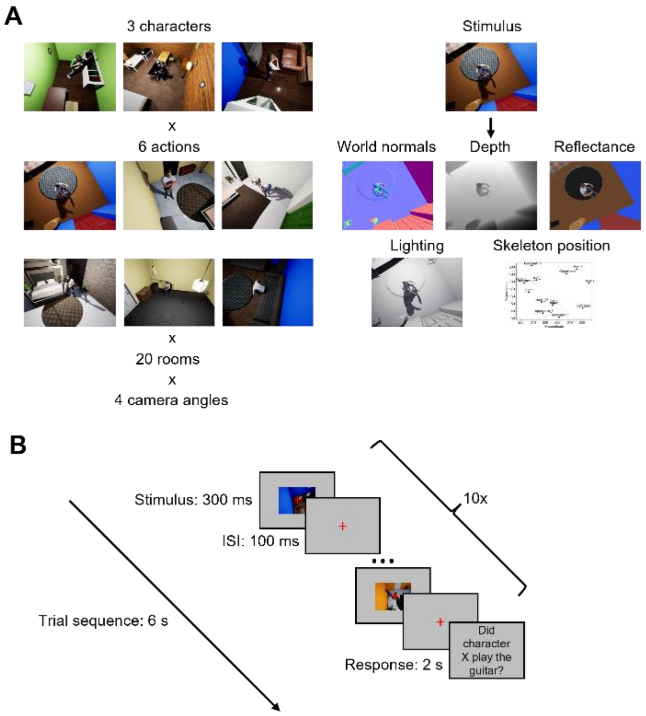

Figure 1: Methods. A. Stimulus set: examples of stimuli and ground-truth annotations for mid-level features. **B.** Experimental paradigm.

Encoding Analysis

To predict the EEG signal from the ground-truth annotations, we used linearizing encoding models (Kriegeskorte & Douglas, 2019; Naselaris et al., 2011). We performed the analysis on subject-level EEG data for each of the low-, mid- and high-level features and for images and videos separately. First, we trained a multiple linear ridge regression model on the training set to predict the EEG signal in each of the 19 posterior EEG channels using the annotations as predictors, independently for each of the 70 EEG time points. Then, we estimated the λ hyperparameter using the validation set, separately for every subject and feature. Afterwards, using the model with the optimized hyperparameter values, we predicted the EEG data from the test set of annotations. Finally, for every subject and every feature, we correlated the predicted EEG data with the true EEG data from the test set. Averaging the correlation over channels and subjects, we obtained a time-course per feature depicting the processing of low-, mid- and high-level features in scene images and videos.

Results

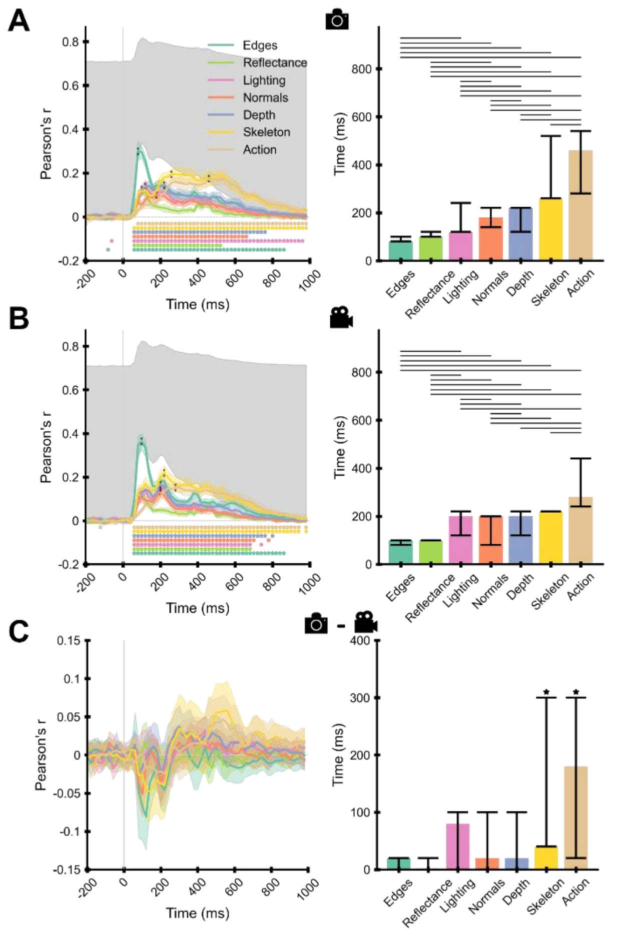

Figure 2: EEG encoding results. Time-courses (left) and peak latencies (right) of the annotations-based encoding accuracies for images (**A**), videos (**B**) and their difference (**C**).

Mid-level feature representations peaked between low- and high-level feature representations Using encoding models, we first observed that for both images and videos, our ground-truth annotations predicted all low-, mid- and high-level features in the EEG data significantly from ~60 ms on (p<0.05, FDR-corrected) (Fig. 2A and B, left). This shows that our annotations are suitable for predicting the EEG data. Second, we observed that mid-level features peaked between ~100 ms and ~250 ms after stimulus onset, for both images and videos (Fig. 2A and B, right). All investigated mid-level features except reflectance peaked significantly (p<0.05) after the low-level feature, edges (on average, 90 ms later), and all mid-level features peaked significantly before the high-level feature, action (on average, 100 ms earlier). This suggests that the selected mid-level features play an intermediary role in the transformations between low- and high-level features.

Skeleton position and action peaked earlier in videos We observed significant differences between images and videos in peak latencies for skeleton position and action (Figure 2C, right). Skeleton position and action peaked respectively 40 ms and 180 ms earlier in videos than in images, suggesting that the dynamic changes in videos help resolve biological motion faster.

Discussion

First, we observed that mid-level features peak between ~100 ms and ~250 ms after stimulus onset, i.e., between low- (~90 ms) and high- (~370 ms) level features. This shows that mid-level features as identified by theoretical models and computer vision (Biederman, 1987; Marr, 1982; Zamir et al., 2018) occupy a place in the middle of scene processing. This finding complements previous research on the temporal (Grootswagers et al., 2019; Proklova et al., 2019; Wang et al., 2022) and spatial (Freeman et al., 2013; Roe et al., 2012; Tsao et al., 2003) dynamics of other mid-level features, such as texture, form and shape. Together this suggests that the mid-level features proposed by human and computer vision models build on low-level features and feed into high-level features to result in the successful processing of semantic information from a scene.

Second, we showed that skeleton position and action peaked earlier for videos than for images, meaning that the movement information in videos aids the processing of biological motion. Building on previous psychophysics (Johansson, 1973) and computer vision (Girish et al., 2020) research, this suggests that the movement contained in dynamic stimuli leads to a more efficient recognition of action information.

In sum, we revealed the temporal dynamics of mid-level features in static and dynamic stimuli and elucidated the role of mid-level features in scene perception.

Acknowledgments

We thank the HPC Service of Zedat, Freie Universität Berlin (Bennett et al., 2020) for computing resources. This study was supported by the European Research Council (ERC) starting grant (ERC-StG-2018-803370 to A.K. and R.M.C.), and the German Research Council (DFG) grants (CI 241/1-1, CI 241/3-1, CI 241/1-7, INST 272/297-1 to R.M.C., FOR 5368 ARENA, RO 6458/2-1, RO 6458/2-1 to G.R.).

References

Bennett, L., Melchers, B., & Proppe, B. (2020). *Curta: A General-purpose High-Performance Computer at ZEDAT, Freie Universität Berlin*. http://dx.doi.org/10.17169/refubium-26754

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

Epic Games. (2019). *Unreal Engine* (4.22.1) [Computer software]. https://www.unrealengine.com

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, *16*(7), Article 7. https://doi.org/10.1038/nn.3402

Girish, D., Singh, V., & Ralescu, A. (2020). *Understanding Action Recognition in Still Images*. 370–371. https://openaccess.thecvf.com/content_CVPRW_2020/html/w23/Girish_Understanding_Action_Recognition_in_Still_Images_CVPRW_2020_paper.html

Groen, I. I. A., Silson, E. H., & Baker, C. I. (2017). Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160102. https://doi.org/10.1098/rstb.2016.0102

Grootswagers, T., Robinson, A. K., Shatek, S. M., & Carlson, T. A. (2019). Untangling featural and conceptual object representations. *NeuroImage*, *202*, 116083. https://doi.org/10.1016/j.neuroimage.2019.116083

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*(2), 201–211. https://doi.org/10.3758/BF03212378

Kriegeskorte, N., & Douglas, P. K. (2019). Interpreting encoding and decoding models. *Current Opinion in Neurobiology*, *55*, 167–179. https://doi.org/10.1016/j.conb.2019.04.002

Marr, D. (1982). *Vision*. W. H. Freeman.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Proklova, D., Kaiser, D., & Peelen, M. V. (2019). MEG sensor patterns reflect perceptual but not categorical similarity of animate and inanimate objects. *NeuroImage*, *193*, 167–177. https://doi.org/10.1016/j.neuroimage.2019.03.028

Roe, A. W., Chelazzi, L., Connor, C. E., Conway, B. R., Fujita, I., Gallant, J. L., Lu, H., & Vanduffel, W. (2012). Toward a Unified Theory of Visual Area V4. *Neuron*, *74*(1), 12–29. https://doi.org/10.1016/j.neuron.2012.03.011

Tsao, D. Y., Vanduffel, W., Sasaki, Y., Fize, D., Knutsen, T. A., Mandeville, J. B., Wald, L. L., Dale, A. M., Rosen, B. R., Van Essen, D. C., Livingstone, M. S., Orban, G. A., & Tootell, R. B. H. (2003). Stereopsis activates V3A and caudal intraparietal areas in macaques and humans. *Neuron*, *39*(3), 555–568. https://doi.org/10.1016/s0896-6273(03)00459-8

Wang, R., Janini, D., & Konkle, T. (2022). Mid-level Feature Differences Support Early Animacy and Object Size Distinctions: Evidence from Electroencephalography Decoding. *Journal of Cognitive Neuroscience*, *34*(9), 1670–1680. https://doi.org/10.1162/jocn_a_01883

Zamir, A., Sax, A., Shen, W., Guibas, L., Malik, J., & Savarese, S. (2018). *Taskonomy: Disentangling Task Transfer Learning* (arXiv:1804.08328). arXiv. https://doi.org/10.48550/arXiv.1804.08328