Streszczenie

Percepcja otaczającego nas świata jest zdumiewająco szybka i bezwysiłkowa. Neuronalne i obliczeniowe podstawy tej niezwykłej sprawności wciąż pozostają jednak niepoznane. Badania neurobiologiczne zidentyfikowały w ludzkim systemie wzrokowym częściowo odrębne ścieżki odpowiedzialne za rozpoznawanie scen i obiektów, ale nadal nie wiadomo, jak i na jakich etapach przetwarzania szlaki te współdziałają. Niniejsza praca doktorska bada interakcje między nimi, opierając się przewidywaniach ugruntowanych modeli teoretycznych oraz wykorzystując komputerowy model widzenia, by ocenić w jakim stopniu zaobserwowane efekty może odtworzyć prostsza architektura dół–góra. We wszystkich badaniach naturalistyczne fotografie posłużyły jako substytut rzeczywistego świata wizualnego.

Pierwsze badanie testowało który element, kontekst sceny (tło) czy obiekt, jest przetwarzany szybciej i w jakiej kolejności oddziałują one na siebie. Uczestnicy klasyfikowali sceny i obiekty w dwóch zadaniach: go/no-go oraz wymuszonego wyboru (two-alternative forced-choice, 2AFC). Czasy reakcji analizowano równolegle ze zintegrowaną miarą łączącą szybkość i dokładność (Balanced Integration Score), aby kontrolować kompromisy między tymi wymiarami. Wyniki nie wykazały istotnej przewagi czasowej dla żadnej z reprezentacji. Wpływy były wzajemne: niezgodne obiekty spowalniały rozpoznawanie sceny, a niezgodny kontekst spowalniał rozpoznawanie obiektów.

Drugie badanie sprawdzało, czy reprezentacje obiektów mogą przyczyniać się do rozpoznawania niejednoznacznych scen. W tym celu zastosowano chronometryczną przezczaszkową stymulację magnetyczną (transcranial magnetic stimulation, TMS), aby określić, czy obszar bocznej części płata potylicznego (lateral occipital complex, LOC), wyspecjalizowany w przetwarzaniu obiektów, wywiera istotny i czasowo specyficzny wpływ na kategoryzację takich scen – analogiczny do udokumentowanej roli obszaru OPA (occipital place area) w identyfikacji niejednoznacznych obiektów. Wyniki w postaci trafności, czasu reakcji oraz zintegrowanej miary łączącej szybkość i dokładność (LISAS) porównano dla trzech okien czasowych stymulacji. Uzyskane dane potwierdzają, że obszar LOC ma istotny, przyczynowy udział w kategoryzacji niejednoznacznych scen, choć nie udało się ustalić dokładnego momentu, w którym ten wpływ zachodzi.

Trzecie badanie miało na celu ustalenie, czy rozpoznanie sceny na podstawie obiektu zależy od jej spójnej struktury, czy może zachodzić nawet wtedy, gdy zachowane są jedynie jej niskopoziomowe statystyki. Dodatkowo sprawdzono, czy zaobserwowany wynik jest specyficzny dla ludzi, czy też można go odtworzyć w sztucznej sieci neuronowej. Uczestnicy klasyfikowali obiekty umieszczone na trzech rodzajach tła: niejednoznacznym, zdegradowanym i neutralnym. Wyniki wykazały, że spójna struktura sceny jest kluczowa dla

rozpoznania sceny opartego na obiekcie. Skuteczność klasyfikacji tych samych bodźców przez model Places365-GoogLeNet nie różniła się istotnie od zachowania ludzi.

Podsumowując, uzyskane wyniki nie potwierdzają założeń ścisłych modeli hierarchicznych w odniesieniu do percepcji scen naturalistycznych – żadna z reprezentacji nie ma czasowej przewagi, a ich wpływy są wzajemne. Badania wykazały również, że reprezentacje obiektów w obszarze LOC mają przyczynowy udział w rozpoznawaniu niejednoznacznych scen, choć nie udało się precyzyjnie określić okna czasowego tego efektu. Ustalenie przebiegu czasowego tego wpływu pozostaje kluczowe dla wyjaśnienia czy interakcje scena–obiekt są wspierane przez dwukierunkowy mechanizm predykcyjny. Rozpoznawanie sceny oparte na obiekcie okazało się zależeć od spójnej struktury sceny, a poprawność klasyfikacji w poszczególnych warunkach nie różniła się istotnie miedzy ludźmi a sztuczną siecią neuronową. Kwestią otwartą pozostaje więc w jakich warunkach dochodzi do rozbieżności między przetwarzaniem ludzkim a modelowym i jakie dokładnie procesy obliczeniowe za te różnice odpowiadają.

Słowa kluczowe: percepcja wzrokowa, scena naturalistyczna, kontekst sceny, obiekt, przezczaszkowa stymulacja magnetyczna, głęboka sieć neuronowa

Abstract

Perception of the visual world is strikingly fast and seemingly effortless. How such efficiency is neurally and computationally implemented remains a long-standing question. Neuroscientific research has identified partially segregated scene- and object-selective pathways in the human visual system, yet it is still unclear how and when these pathways interact. This thesis investigated their interaction in human vision, drawing on existing theoretical accounts. As a complementary benchmark, a computer model of human vision was evaluated to assess the extent to which the observed effects can be reproduced in a feedforward artificial architecture. In all studies, naturalistic photographs served as proxies for real-world input.

Study 1 tested whether one of the representations – scene context or objects – exhibits a temporal processing advantage and in what sequence these elements influence one another. Participants classified scenes and objects in two tasks: a go/no-go task and a two-alternative forced-choice (2AFC) task. Reaction times were analysed alongside the integrated speed–accuracy measure (Balanced Integration Score) to control for speed–accuracy trade-offs. After controlling for these trade-offs, no reliable temporal advantage was observed for either representation. The influences were mutual, with incongruent objects slowing scene context recognition and incongruent context slowing object recognition.

Study 2 examined whether object representations play a causal role in disambiguating scenes. Using chronometric TMS, the study assessed whether the object-selective lateral occipital complex (LOC) makes a causal, time-specific contribution to the categorization of ambiguous scenes, in parallel to the established role of the scene-selective occipital place area (OPA) in the classification of ambiguous objects. It was further assessed whether the effective temporal window of the LOC coincides with the one previously reported for the OPA. Accuracy, reaction times, and integrated speed–accuracy measure (LISAS) were compared across three stimulation windows. Results support a causal contribution of the LOC to the disambiguation of scenes, although its precise timing could not be conclusively established.

Study 3 investigated whether object-facilitated scene recognition requires a coherent scene layout or can still occur when only low-level scene statistics are preserved. It further examined whether this dependency is uniquely human or can also be instantiated within a feedforward artificial architecture. Human participants classified scenes with objects placed on ambiguous scenes, phase-scrambled scenes, and neutral backgrounds. Performance showed that a coherent layout is critical for object-based facilitation. The same images were classified by Places365-GoogLeNet, and model performance was not significantly different from human behaviour under these manipulations.

Taken together, the results provide little support for strictly hierarchical (“object-first” or “scene-first”) accounts of real-world scene perception – there is no temporal processing

advantage for either scene or object representation, and their influences are mutual. Critically, it is shown for the first time that object representations in the LOC make a causal contribution to the recognition of ambiguous scenes. Pinpointing the precise temporal window of this contribution will be essential for evaluating whether scene–object interactions are implemented via a shared bidirectional predictive processing mechanism. It is further demonstrated that object-based facilitation of scene recognition depends on a coherent scene layout, and that classification accuracy across conditions does not differ significantly between human observers and the artificial neural network. An important task for future work will therefore be to specify the conditions under which human and model processing diverge, and to identify the computational mechanisms that give rise to these differences.

Keywords: visual perception, real-world scene, scene-context, object, transcranial magnetic stimulation (TMS), deep neural network (DNN)