Objekt- und Szenenerkennung: Einblicke in kognitive Prozesse

Object- and Scene Recognition: Insight in Cognitive Processing

Herbert Hagendorf
Institut für Psychologie
Humboldt Universität zu Berlin
Rudower Chaussee 18
12489 Berlin
Tel. 030-20939382
E-mail: herbert.hagendorf-rz.hu-berlin.de

Zusammenfassung
Der Abruf von Bildinformation ist eine Thematik, an der neben den Computerwissenschaften die kognitiven Neurowissenschaften und die Psychologie mit unterschiedlichen Zielstellungen arbeiten. Die Forschung an biologischen Systemen hat als eine Aufgabe, auf der Grundalge der relativ langsamen Neurone effiziente Erkennungsalgorithmen zu entwickeln. Als Vorstufe für eine solche Diskussion werden Ergebnisse zu kognitiven Prozessen der Objekt- und Szenenerkennung vorgestellt, die Restriktionen für solche biologisch orientierte algorithmische Modellierungen in der Bildverarbeitung darstellen. Die Ergebnisse zeigen, dass ein hierarchischer Erkennungsprozess, der mit den Merkmalen beginnend über Objekte und Szenen fortschreitet, nicht gehalten werden kann. Die angesichts der temporalen Verarbeitungsmerkmale der Neurone überaus schnelle und ohne Aufmerksamkeit erfolgende Bereitstellung semantischer Information über eine Szene verweisen auf zwei mögliche Lösungen: Zugriff auf Semantik über globale Merkmale einer Szene oder Beschleunigung der Verarbeitung durch Rückkopplung von höheren Ebenen.

Abstract
In research on retrieval of picture information computer science, cognitive neuroscience, and psychology are engaged. One of the aims of the research in biological systems is to gain knowledge which can be used for the development of algorithms based on low temporal constants of neurons. As a preliminary for such a research program results on cognitive processes on object- and scene recognition are presented. The results are not in accordance with a hierarchical recognition process starting with feature analysis and resulting in recognition of objects which are the basis for scene recognition. Taking into consideration the slow temporal processing characteristics of neurons the extraordinary fast retrieval of semantic information of scenes from memory are in accordance with two possible processing architectures: retrieval of semantic information on the basis of fast processing of correlated global features of scenes or making processing more efficient by recurrent processing.

Einleitung
Der Abruf von Bildinformation ist eine Thematik, an der neben den Computerwissenschaften die kognitiven Neurowissenschaften und die Psychologie mit unterschiedlichen Zielstellungen arbeiten. Dabei könnte durchaus die Frage diskutiert werden, ob die Erarbeitung von Algorithmen nicht auch von Einsichten in die Arbeitsweise biologischer System profitieren kann. Als Vorstufe für eine solche Diskussion werden Ergebnisse zu kognitiven Prozessen der Objekt- und Szenenidentifikation vorgestellt, die Restriktionen für eine solche biologisch orientierte algorithmische Modellierungen in der Bildverarbeitung darstellen. Szenen sind nach Henderson und Hollingworth (1999)

semantisch kohärente Ansichten einer realen Umwelt, die aus Hintergrundelementen und mehreren Objekten in einer spezifischen räumlichen Anordnung gebildet werden.
Ausgangspunkt ist die allgemein akzeptierte These, dass die Wahrnehmung ein aktiver und konstruktiver Vorgang ist. Aus der psychologischen Forschung der letztem 30 Jahre ist eine Reihe von experimentellen Ergebnissen vorgelegt worden, welche Einblicke in die bis heute so schwer erklärbare und nachvollziehbare Leistungsfähigkeit der bildbezogenen menschlichen Wahrnehmungs- und Gedächtnisleistungen liefern.

Erkennung als sequentieller Prozess
Die Verarbeitung visueller Information durch die Retina unterliegt der Restriktion, dass das räumliche Auflösungsvermögen in der Retina nur in einem Raumwinkel von wenigen Grad hoch ist. Daraus folgt schon, dass die Verarbeitung eines Bildes über mehrere Fixationen in einem räumlich und zeitlich erstreckten Prozess erfolgt. Dabei ist noch immer unklar, wie die interne Repräsentation einer Szene aussieht. Ergebnisse zur Veränderungsblindheit, d.h. der Entdeckung einer Veränderung in einem Bild, wenn ein kleiner zeitlicher Unterschied zwischen den beiden Bildern existiert (Zeit zwischen zwei Fixationen, Bildschnitt in einem Film, experimentell gesetzter zeitlicher Abstand von etwa 100 ms), führt dazu, dass Personen Schwierigkeiten haben, solche Unterschiede (Veränderungen in Merkmalen, Objekten) zu bemerken. Diese wirft die Frage auf, welche Information tatsächlich in einem Bild gespeichert ist. Gesichert ist, dass zur Erkennung solcher Veränderungen zwischen zwei Bildern fokussierte Aufmerksamkeit am Ort der Veränderung notwenig ist (Simons und Rensink, 2005).

Bildgedächtnis
Dieser scheinbar geringen Gedächtnisleistung aus Arbeiten zur Veränderungsblindheit stehen andererseits beeindruckende Gedächtnisleistungen gegenüber (Shepard, 1967; Standing, 1977). Das Bildgedächtnis gilt als sehr robust. Die Vergessenrate ist gering. Shepard zeigte Personen über 600 Bilder jeweils für 6 s und testete die Wiedererkennungsrate, indem er jeweils aus zwei Bildern eins auswählen ließ. Die Erkennungsrate lag bei 98%. Standing dehnte diese Untersuchung auf 10 000 Bilder aus und fand eine Erkennungsrate von 83 %. Auch hier blieb allerdings die Frage offen, was ist die Grundlage solche Erkennungsleistungen: Merkmale, Bedeutung, verbale Umschreibungen?

Schnelle Bilderkennung ohne Aufmerksamkeit
Szenen und Objekte lassen sich auf unterschiedlichen Ebenen beschreiben, die auf ein hierarchisches Erkennungsmodell führen: Gestaltfilter zur Verarbeitung sensorischer Information, Ermittlung von Merkmalen der Objekte, Erfassung von Objekten und Objektkategorien und schließlich Erfassung der räumliche Anordnungen von Objekten zu Szenen. Einerseits sind solche hierarchischen Beschreibungen aus Kategorisierungsleistungen bekannt (Tversky & Hemenway, 1983), andererseits ist seit den gestaltpsychologischen Arbeiten immer wieder dokumentiert worden, dass die Erkennung einer Ganzheit („Wald“) nicht über die Erkennung einzelner Objekte („Bäume“) laufen muss (Navon, 1977). Dies führt auf die Frage, welche Information eigentlich bei der Betrachtung eines Bildes zu welchem Zeitpunkt bereitgestellt wird.
In den letzen Jahren ist eine Reihe von Ergebnissen vorgestellt worden, die zeigen, dass in sehr kurzer Zeit und ohne fokussierte Aufmerksamkeit eine semantische Klassifikation von Objekten erreicht werden kann. Wenn Bilder von Objekten oder Szenen für eine kurze Zeit in schneller Abfolge (6 Bilder pro Sekunde) dargeboten werden, ist die Wiedererkennungsrate nach etwa 2 Minuten sehr schlecht. Allerdings sagt dies nichts über die Tiefe der Verarbeitung in diesen 167 ms aus. Wenn die Personen bei vorgegebener verbaler Beschreibung oder vorgegebenem Bild in einer solchen schnellen Abfolge ein Bild suchen mussten, war diese Erkennungsleistung trotz der schlechten Gedächtnisleistung über 75 %. Diese zeigt, dass bei extrem kurzer Darbietung ein semantische Verarbeitung eines Bildes gelingt (Potter et al. 2004).
Bei Darbietungszeiten von 20 ms kann eine Person das Vorhandensein eines Tieres auf einer Fotografie überzufällig in weniger als 300 ms mit einem Tastendruck anzeigen. Diese Leistung wird auch durch Training mit diesen Bildern nicht beeinflusst (vanRullen & Thorpe, 2001). Aus der

Suche eines Bildes in einer Menge von dargebotenen Bildern ist bekannt, dass es Konstellationen gibt, bei denen diese Suche nicht von der Anzahl der Bilder in dieser Menge abhängt, also keinen sequentiellen Aufmerksamkeitsprozess verlangt. Inzwischen ist auch getestet, dass die erwähnte Kategorisierungsleistung keine Aufmerksamkeit erfordert. In einer Doppelaufgabenanforderung hatten Personen gleichzeitig zwei Kategorisierungsaufgaben zu lösen. Eine am Fixationsort auszuführende Buchstabenunterscheidung wurde begleitet von einer peripher dargebotenen Kategorisierungsaufgabe mit Bildern (Li et al., 2005). Probanden können auf diesen peripher dargebotenen Bildern Tiere oder Verkehrsmittel entdecken, ohne dass diese durch die zentrale Buchstabenaufgabe beeinträchtigt würde. Diese Ergebnisse besagen, dass eine semantische Repräsentation eines Bildes auch außerhalb des Fokus der Aufmerksamkeit abgerufen werden kann. Dies spricht dafür, dass sehr früh in der Verarbeitung ein Zugriff auf semantische Information erfolgt.
Die frühe Rolle der semantischen Information wird auch bei der Objekterkennung deutlich. Immer wieder wurde gefunden, dass die Erkennung von Objekten durch einen top down Prozess (top down Information in diesem Kontext ist jede Information, die nicht in dem zu verarbeitenden Reiz enthalten ist) beeinflusst wird. In einer Untersuchung wurden Objekte in einer semantisch konsistenten Umgebung (Tänzerin auf einer Bühne) oder einer semantisch inkonsistenten Umgebung (Tänzerin in einem Schwimmbad) gezeigt. Bei einer Darbietungszeit von 80 ms gibt es für die Benennung des Objektes einen Konsistenzeffekt. Dies gilt auch für die Erkennung des Hintergrundes: ein konsistentes Objekt verkürzt die Erkennung des Hintergrundes. Das Ergebnis spricht für eine interaktive Verarbeitung von Objekt und Hintergrundinformation.

Augenbewegungen und Szenenerkennung

Im Prinzip zeigen diese Untersuchungen, dass schon während der ersten Fixation auf einem Bild semantische Information bereitgestellt wird (Anm.: Die Darbietungszeit der Bilder war so kurz, dass eine Refixation des Auges nicht möglich war). Da nun bei freier Betrachtungszeit Sequenzen von Augenbewegungen wegen der räumlichen Begrenzung der hohen Auflösung notwendig sind, kann auch aus der Folgen von Augenfixationen auf einem Bild entnommen werden, welche Information zu welchem Zeitpunkt aufgenommen wird. Es ist seit den ersten Untersuchungen bekannt, das solche Augenbewegungen natürlich sehr stark von den Absichten des Betrachters abhängen. Bekannt ist auch, dass die erste Fixation auf einem Bild nicht zufällig erfolgt. Der erste Zielort ist ein Bildbereich, der informativ vor dem Hintergrund der Betrachtungsaufgabe ist. Versuche, diese Regularität in den Fixationen durch Salienzkarten zu erfassen, die über Kantendichte, Raumfrequenzen oder Kontraste definiert sind, müssen als fehlgeschlagen angesehen werden (Henderson, 2004). Vielmehr ist heute klar, dass die Kontrolle der Sequenz der Augenbewegungen und damit der Szenenerkennung von verschiedenen Faktoren abhängen: vom visuellen Eingangssignal, von früher betrachteten Informationen eines Bildes und von langfristig gespeichertem Wissen über visuelle, räumliche und semantische Regularitäten einer Szene. Also auch aus diesen Untersuchungen folgt, dass eine Interaktion von lokaler und globaler Information zur Kontrolle der Augenbewegungen bei der Szenenerkennung ausgenutzt wird.

Mögliche Erklärungsansätze für die Szenenerkennung

Geht man von der mit verschiedenen Methoden dokumentierten Rolle des frühen Zugriffs auf semantische Information aus, dann erhebt sich die Frage, welche Information vermittelt diesen Zugriff. In einer globalen Betrachtung kommt Rensink (2000) zu einer Rahmenvorstellung, die von drei Systemkomponenten ausgeht: visuelle Verarbeitung auf elementarem Merkmalsniveau, aufmerksamkeitsgesteuerte Objektverarbeitung und Verarbeitung von semantische Information über die Bedeutung einer Szene (engl. gist) einschließlich des räumlichen Layouts. Diese Information dient als Kontrollinformation für die Ausrichtung der Aufmerksamkeit im Bild. Wie erfolgt der Zugriff auf diese semantische Information?
Die erwähnte Studie von Vogel et al. (2006) lässt den Schluss zu, dass lokale und globale Information genutzt wird. Ein Versuch, den Zugriff über globale Bildmerkmale zu zeigen, geht auf Oliva (2005) zurück. Aus Beschreibungen von Bildern bei kurzzeitiger Darbietung von 100 ms kann geschlossen werden, dass eine Vielzahl von Information auf verschiedenen Beschreibungsniveaus zugänglich wird. Personen erinnern die Bedeutung einer Szene, die Kategorie, einzelne Objekte

(das visuelle Arbeitsgedächtnis umfasst etwa 3-4 Objekte) und ausgewählte saliente Merkmale (z.B. Farben). Oliva ging von statistischen Eigenschaften von Kategorien natürlicher Szenen (Berge, Wüsten etc.) aus. Sie ermittelten aus Ratings von Personen verschiedene Dimensionen eines multidimensionalen Raumes, in dem eine Klassifikation von Szenen gelingt, z.B. Offenheit, Symmetrie, Tiefe. Diese Informationen über die perzeptiven Dimensionen wurden über Kombinationen linearer Filter, wie sie charakteristisch für die frühe Verarbeitung im visuellen System sind, ermittelt. Es hat sich gezeigt, dass die Wahrnehmungsschwelle für eine Beurteilung solcher Dimensionen bzw. für eine Kategorisierung diese Sichtweise bestätigen. Die Wahrnehmungsschwelle für diese globalen Dimensionen liegt mit 23 ms unter der Wahrnehmungsschwelle von 33 ms für die Kategorieinformation. Vor dem Hintergrund dieses Zugriffs auf semantische Information über globale Bildmerkmale, die durch eine Kombination von Raumfrequenzfiltern abgeleitet werden können, haben Vogel et al. (2006) eine Erweiterung vorgeschlagen. Ihre Untersuchungen zeigen, dass die Kategorisierung von Szenen besser gelingt, wenn globale Information und lokale Information interaktiv genutzt werden.

Affektive Merkmale von Szenen

Aus psychologischer Sicht sei ergänzend hinzugefügt, dass Bilder natürlich auch eine affektive Komponente haben. Diese bewertende Information wird auch in vergleichbar kurzer Zeit wie die kategorisierungsrelevante Information bereitgestellt. Die Klassifikation von Gesichtsausdrücken nach dem emotionalen Gehalt wird durch einen vorangestellten Szenenreiz beeinflusst, der positiv oder negativ von Personen bewertet worden ist. Dabei entfaltet ein solcher Reiz seine Wirkung, obwohl er mit einer Zeit unterhalb der Wahrnehmungsschwelle dargeboten wurde. Auch hier konnte Vessel und Biederman (2001) aus Urteilen von Personen ein mehrdimensionales Bewertungssystem aufstellen, das die Präferenzurteile der Probanden gut erklärt. Darüber hinaus konnte gezeigt werden, dass diese Präferenzen tatsächlich etwas mit einem neuronalen Belohnungssystem zu tun haben.

Schlussbemerkung

Zusammenfassend lässt sich sagen, dass eher klassische Auffassungen vom Wahrnehmungsprozess davon ausgehen, dass dieser Zugriff auf die Bedeutung einer Szene in einem hierarchischen sequentiellen Prozess erfolgt, der von elementaren Merkmalen über Objekte zu der Szeneninformation fortschreitet, nicht zu halten ist. Die angeführten Untersuchungen zeigen, dass die Information über ein komplexes Bild wie das einer Szene schon sehr früh verfügbar wird. Allein schon bestimmte elementare Informationen wie Raumfrequenzen oder Farben können diese Information vermitteln. Zum Ausdruck kommt dies in biologisch motivierten Modellen zu Aufmerksamkeitssteuerung bei der Bildbetrachtung (Siagian & Itti, 2006). Bleibt zu prüfen, ob die Ergebnisse, gewonnen an natürlichen Szenen, also an Reizen mit besonderer biologischer Relevanz, auf andere Bildklassen generalisiert werden können.

Literatur

Davenport, J.L. & Potter, M.C. (2004). Scene consistency in object and background perception. Psychological Science, 11, 559-564.

Greene, M.R. & Oliva, A. (2006). Natural scene categorization from conjunction of ecological global properties. Proceedings of the 28th Annual Conference of the Cognitive Science Society, *Vancouver, July (pp. 291-296).*

Henderson, J.M. & Hollingworth, A. (1999). High level scene perception. Ann. Rev. Psychol., 50, 243-271.

Henderson, J.M. (2004). Human gaze control during real world scene perception. Trends in Cognitive Science, 7, 498-504.

Li, F.F; VanRullen, R.; Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. PNAS, 99, 5596-5601.

Maljkovic, V. & Martini, P. (2005). Short–term memory for scenes with affective content. Journal of Vision, 5, 215-229.

Navon, D. (1977). Forest before trees: the precedence for global features in visual perception. Cognitive Psychology, 9, 353-383.

Oliva, A. (2005). Gist of the scene. In L. Itti; G. Rees, & J.K. Tsotsos (Eds.). Neurobiology of Attention.San Diego: Elsevier.

Potter, M.C.; Staub, A., & O´Connor, D.H. (2004). Pictorial and conceptual representation of glimpsed pictures. Journal of Experimental Psychology: Human Perception and Performance, 30, 478-489.

Rensink, R.A. (2000). The dynamic representation of scenes. Visual Cognition, 7, 17-42.

Simons, D.J.; Rensink, R.A. (2005). Change blindness: Past, present, and future. Trends in Cognitive Sciences, **9**: 16-20.

Shepard, R.N. (1967). Recognition memroy for words, sentences, and pictures. Journal of Verbal Learning and Verbal Behavior, 6, 156-163.

Siagian, C. & Itti, L. (2006).Rapid biologically inspired scene classification using features shared with visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence, in press.

Standing, L. (1973). Learning 10.000 pictures. Quarterly Journal of Experimental Psychology, 25, 207-222.

Tversky, B.; Hemenway, K. (1983). Categories of environmental scenes. Cognitive Psychology, 15, 121-149.

VanRullen, R. ; Thorpe, S.J: (2001). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual categories. Perception, 30 (6), 655-668.

Vessel, E.A.; Biederman, I. (2001). Why do we prefer to looking at some scenes rather than others? Presentation for OPAM Conference . New Orleans, Louisiana.

Vogel, J.; Schwaninger, A.; Wallraven, C.; Bülthoff, H. (2006) Categorization of natural scenes: local and global information. To appear in *Symposium on Applied Perception in Graphics and Visualization APGV 2006*, Boston, MA, USA, July 2006.