AUFMERKSAMKEIT UND AUGENBEWEGUNGEN

Sehseminar 2011 - Augenbewegungen - Funktion und Anwendungen

Robert Böser
robert.boeser@uni-ulm.de
Universität Ulm, 18.06.2011

1. Einführung

Die Aufgabe, zu entscheiden, wo sich ein bestimmtes Objekt in einer Szene befindet, kann unterschiedlich schwer sein. Betrachtet man beide Fälle in Bild 1, in welchen das gesuchte Objekt jeweils ein horizontaler roter Balken ist, so findet man dieses Objekt im linken Teilbild sehr schnell, da es sich alleine in der Farbe von den anderen Objekten unterscheidet und geradezu aus der Menge der Objekte heraussticht. Dies wird als *pop-out-Effekt* bezeichnet. Dabei ist die Anzahl der Objekte unerheblich. Im rechten Teilbild ist es schwer, das gesuchte Objekt zu finden. Man muss nach einer Kombination von Merkmalen suchen. Merkmalskombinationen fallen nicht sofort auf und die Objekte müssen nacheinander betrachtet werden, wodurch die benötigte Zeit mit der Anzahl der vorhandenen Objekte zunimmt.

Im Folgenden wird die Ursache dieses Phänomens näher untersucht und ein Modell vorgestellt, welches das beobachtete Phänomen simuliert und mit dessen Hilfe die notwendigen Verarbeitungsschritte erklärt werden.

2. Der biologische Hintergrund

Die visuelle Information wird im Gehirn von zwei parallelen Pfaden, dem dorsalen und dem ventralen Pfad verarbeitet. Der ventrale Pfad dient dem Erkennen und der Identifikation von Objekten, wo hingegen der dorsale Pfad der Lokalisierung von Objekten dient[4]. Hier werden prägnante Bildbereiche bestimmt. Da im dorsalen Pfad aber nur Aussagen zum Ort gemacht werden können, müssen diese Bildbereiche im ventralen Pfad analysiert werden. Die Aufmerksamkeit und damit der Blick des Betrachters wird deshalb nacheinander auf die im dorsalen Pfad bestimmten prägnanten Bildbereiche gerichtet.

In Bild 1 kann links das gesuchte Objekt schnell gefunden werden, da es sich ausschließlich in einem Merkmal, der Farbe, von den restlichen Objekten abhebt. Es hat auf Grund der Einzigartigkeit dieses Merkmals eine hohe Prägnanz und die Aufmerksamkeit wird direkt auf das gesuchte Objekt gelenkt. Im rechten Teilbild zeichnet sich das gesuchte Objekt durch eine Kombination von Merkmalen aus. Da diese Kombination aber nicht im dorsalen Pfad ausgewertet werden kann, wird die Aufmerksamkeit nacheinander auf prägnante Regionen gelenkt und diese im ventralen Pfad ausgewertet, bis das gesuchte Objekt gefunden wird. Durch diese sequenzielle Betrachtung des Bildes erhöht sich die benötigte Suchzeit. Befindet sich das gesuchte Objekt nicht in der Szene, so müssen alle Objekte betrachtet werden, um zu einer Antwort zu gelangen. Ist es hingegen vorhanden, so findet man dieses sogar bei zufälliger Auswahl der Bildbereiche im Durchschnitt nach der Hälfte der betrachteten Objekte.

Die schnelle Bestimmung prägnanter Bildbereiche für die eingabegetriebene Lenkung der Aufmerksamkeit im dorsalen Pfad wird durch Neuronen im primären visuellen Kortex ermöglicht, die direkt auf bestimmte Merkmale reagieren und diese parallel verarbeiten. Im Folgenden werden diese Merkmale, dazu gehören die Kontraste der Helligkeit, der Farben rot-grün sowie blau-gelb und die Kontrast-Orientierung, aber auch die Richtung und die Geschwindigkeit bewegter Objekte[3,4], als Primärmerkmale bezeichnet.

3. Das Modell

Itti und Koch haben ein Modell entwickelt, welches die eingabegetriebene Lenkung der Aufmerksamkeit simuliert. Es arbeitet auf einzelnen Standbildern. Aspekte der Bewegung und des stereoskopischen Sehens können somit vernachlässigt werden. In Bild 2 wird dieses Modell schematisch dargestellt, und die einzelnen Verarbeitungsschritte werden im Folgenden näher erklärt.

Zuerst werden 42 sogenannte Merkmalskarten berechnet. Die verwendeten Merkmale sind, in Anlehnung an die in Abschnitt 2 beschriebenen menschlichen Primärmerkmale, der Helligkeitskontrast, die Farbkontraste rot-grün und blau-gelb sowie vier Kontrastorientierungen im Winkel von 0°, 45°, 90° und 135°. Zu jedem Primärmerkmal werden jeweils sechs Merkmalskarten auf unterschiedlichen Skalen berechnet. Durch die Verkleinerung der Eingabe durch verschiedene Skalen kann ein einzelner Filter ein Merkmal in unterschiedlichen Größen detektieren.

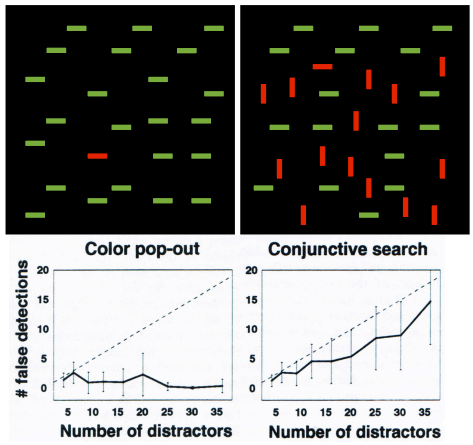

Bild 1[3]: gesucht ist jeweils ein horizontaler roter Balken.
links: Das Ziel zu finden, ist leicht, da es sich nur in einem Merkmal von seiner Umgebung unterscheidet. Die Suchzeit ist unabhängig von der Objektanzahl nahezu konstant.
rechts: Unterscheidet sich das Ziel aber in einer Kombination von Merkmalen, so wird die Suche aufwändiger. Die einzelnen Objekte müssen nacheinander betrachtet werden und die erforderliche Zeit steigt mit deren Anzahl.

3.1. Berechnung unterschiedlicher Skalen

Die Verkleinerung des Eingabebildes um den Faktor 2 wird schrittweise durch Glättung und anschließende Unterabtastung durchgeführt[1]. Die Eingabe wird dazu in Quadrate von 2x2 Pixeln zerlegt. Die Pixel der nächsten Skala berechnen sich aus dem Durchschnittswert dieser Quadrate:

$$S_k(x,y) = \frac{1}{4}\left(\sum_{i=2x}^{2x+1}\sum_{j=2y}^{2y+1} S_{k-1}(i,j)\right) \quad (k = 1, 2, ..., 8)$$

Dieser Schritt wird acht mal wiederholt. Daraus ergeben sich die Skalen $k = 0$ (Originalbild) bis $k = 8$ (Reduktionsfaktor 1:256).

3.2. Die Merkmalsfilter basieren auf einer Zentrum-Umfeld-Operation

In diesem Modell werden Zentrum-Umfeld-Operationen (ZU) durch einfache Differenzen zwischen verschiedenen Skalen des Eingabebildes realisiert. Es werden folgende Skalen-Paare (z, u) verwendet (2, 5), (2, 6), (3, 6), (3, 7), (4, 7) und (4, 8)[3]. Das Ergebnis der ZU, interpoliert auf Skala 0, ergibt sich nun durch:

$$Y(x,y) = Z_z\left(\left\lfloor \frac{x}{2^z} \right\rfloor, \left\lfloor \frac{y}{2^z} \right\rfloor\right) - U_u\left(\left\lfloor \frac{x}{2^u} \right\rfloor, \left\lfloor \frac{y}{2^u} \right\rfloor\right)$$

z und u sind dabei die Skalen des Zentrums und des Umfeldes. Die Gauß-Klammern stellen dabei sicher, dass es sich bei den Pixel-Positionen um ganzzahlige Werte handelt.

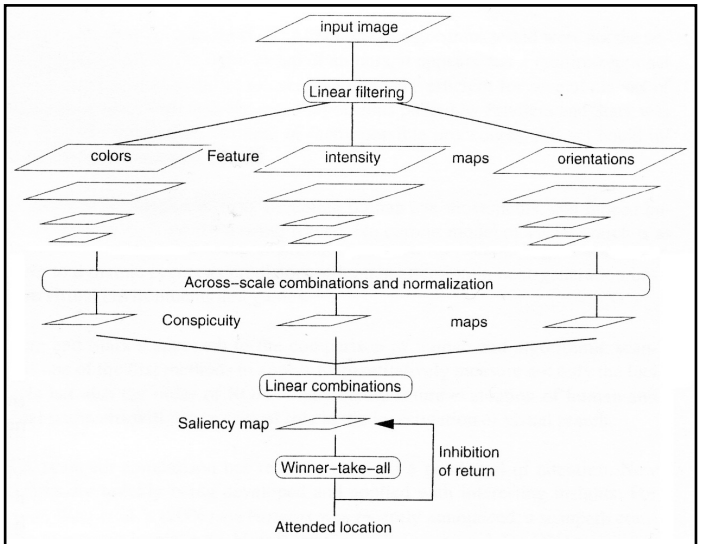

Bild 2[3]: Schematisches Diagramm des vorgestellten Modells. Die einzelnen Verarbeitungsschritte sind Skalierung und Filterung der Eingabe, Normalisierung und Verbesserung der Filterergebnisse, Berechnung der einzelnen Merkmalskarten, Kombination der Merkmale zur Saliency Map und Ausgabe der markantesten Stelle mit Unterdrückung zuvor besuchter Positionen.

3.3. Der Faltungsoperator

Lineare Filter können mit Hilfe der Faltung formuliert werden. Lineare Filter können bei Operationen wie Glättung oder Kontrast-Detektion Anwendung finden.

Die Faltung ist ein Operator, der zwei Funktionen f und g in eine dritte Funktion $(f * g)$ überführt.

$$(f * g)(x, y) = \sum_{i=0}^{X-1} \sum_{j=0}^{Y-1} f(i, j) \cdot g(x-i, y-j)$$

Dabei ist f das Eingabebild und g die Filtermaske. X und Y sind die Breite und Höhe der Eingabe[2]. Anschaulich kann die Faltung dadurch beschrieben werden, dass die Filtermaske über das Eingabebild bewegt und an jedem Punkt die gewichtete Summe der Umgebungswerte ausgewertet wird.

Bild 3[2]:
a) Laplace-Filter für die 2. Ableitung: mit dieser Filtermaske kann die Kontraständerung detektiert werden. Dies geschieht durch Subtraktion des gemittelten Umgebungswertes vom Zentrum.

$$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

b) orientierter Kontrastfilter: Diese Anordnung der Gewichte ermittelt den Kontrast in einer bestimmten Richtung. Die hier gezeigte Maske detektiert horizontale Kontrastkanten.

$$\begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

3.4. Berechnung der Merkmalskarten

Die Berechnung der Merkmalskarten des Helligkeits-Merkmals erfolgt mittels einer ZU durch einfache Differenz der Intensitätskanäle der Skalenpaare in Abschnitt 3.2.

Der Kontrast der Farben rot und grün wird ebenfalls mit einer ZU durch den Differenzbetrag

$$Kontrast_{rot\text{-}grün}(x, y) = |(R_z(x_z, y_z) - G_z(x_z, y_z)) - (G_u(x_u, y_u) - R_u(x_u, y_u))| \quad mit\ x_s = \left\lfloor \frac{x}{2^s} \right\rfloor;\ y_s = \left\lfloor \frac{y}{2^s} \right\rfloor$$

ermittelt. Dabei sind R und G die Farbkanäle rot bzw. grün der jeweiligen Skala, z und u sind die Skalen des Zentrums bzw. der Umgebung, und es werden die Skalenpaare aus Abschnitt 3.2. verwendet. Der Kontrast der Farben blau und gelb wird auf die gleiche Weise berechnet.

Um das Orientierungs-Merkmal zu bestimmen, werden zunächst die unterschiedlichen Skalen des Eingabebilds mit einer orientierten Kontrast-Filtermaske (siehe Bild 2b) gefaltet:

$$Orientierung_{s,o}(x, y) = (Eingabe_s * Filtermaske_o)(x, y)$$

Dabei ist s die verwendete Skala und o die jeweilige Orientierung. Aus diesen Filterergebnissen werden durch ZU-Operationen analog der Berechnung des Helligkeits-Merkmals die Merkmalskarten berechnet. Da ein Orientierungsfilter aber nur eine bestimmte Orientierung detektieren kann, müssen mehrere unterschiedliche Filtermasken angewendet werden.

3.5. Normierung und Kontrasterhöhung der Merkmalskarten

Die Merkmalskarten werden auf einen festen Dynamikbereich zwischen 0 und 1 normiert[3]:

$$Karte_{normiert}(x, y) = \frac{Karte_{roh}(x, y)}{\max(Karte_{roh})}$$

Diese Normierung ist notwendig, um die Ergebnisse der unterschiedlichen Merkmalsfilter vergleichen zu können. Um schwache Auffälligkeiten in den Merkmalskarten besser detektieren zu können und gleichmäßige Regionen abzuschwächen, wird der Kontrast der normierten Merkmalskarten verstärkt:

$$Karte_i = |Karte_{i-1} + Karte_{i-1} * Filtermaske - C|_{\geq 0} \quad (i = 1, ..., 10)$$

Jede normierte Merkmalskarte wird mit einer Filtermaske ähnlich Bild 2a gefaltet und das Ergebnis zur alten Merkmalskarte addiert. Dadurch werden Bereiche mit hohem Kontrast verstärkt. Durch die Subtraktion eines Bias $C = 0{,}02$ werden homogene Bereiche abgeschwächt. Negative Werte werden durch $|.|_{\geq 0}$ auf 0 gesetzt[3]. Diese Kontrastverbesserung muss über mehrere Iterationen i durchgeführt werden, da in jedem Schritt nur kleine Änderungen der Werte auftreten.

3.6. Zusammenfassung der Merkmalskarten und Bestimmung des prägnantesten Bereichs

Die normalisierten Merkmalskarten werden über die verschieden Skalen eines Merkmals zu den so genannten *Conspicuity Maps* zusammengefasst. Es gibt drei dieser *Conspicuity Maps*, jeweils eine für Farbe, Helligkeit und Orientierung. Alle Merkmalskarten des jeweiligen Merkmals werden hierfür punktweise aufaddiert. Wegen der These von Itti und Koch, dass die Prägnanz von Bildbereichen bei ähnlichen Merkmalen konkurriert, unterschiedliche Merkmale jedoch einen unabhängigen Teil zur Gesamtprägnanz beitragen[3], wird auf die *Conspicuity Maps* die in Abschnitt 3.5. beschriebene Normierung und Kontrasterhöhung angewendet.

Die *Saliency Map* stellt eine universelle und von den Merkmalen unabhängige Repräsentation der Prägnanz der Objekte im Eingabebild dar. Mit ihrer Hilfe wird der Bereich der höchsten Prägnanz bestimmt. Die *Saliency Map* berechnet sich aus der punktweisen Summe der drei normierten *Conspicuity Maps*. Der prägnanteste Bildbereich wird durch das Maximum der Werte in der *Saliency Map* repräsentiert. Um die Stelle des Maximums zu ermitteln, verwenden Itti und Koch ein sogenanntes *Winner-take-all-Netzwerk* (*WTA*). Das *WTA*, ein zweidimensionales einschichtiges künstliches neuronales Netzwerk, enthält für jeden Punkt der *Saliency Map* ein Neuron. Jedes dieser Neurone hemmt alle anderen Neurone des Netzes, so dass das Neuron, welches aus der *Saliency Map* den stärksten Input erhält, die Ausgaben aller anderen Neurone unterdrückt und als Gewinner hervorgeht. Die Aufmerksamkeit wird nun auf den Bildbereich gerichtet, der durch die Ausgabe des Gewinnerneurons markiert wird.

Da das Modell aber auf Standbildern arbeitet, würde die Aufmerksamkeit auf diesem Punkt verharren. Die Position mit der nächst niedrigeren Prägnanz wird ermittelt, indem eine Rückkopplung in die *Saliency Map* stattfindet. In einer Umgebung um die oben bestimmte Maximalstelle werden die Werte der *Saliency Map* vermindert. Das *WTA* wertet im nächsten Schritt die *Saliency Map* erneut aus und wählt nun die Region mit der zweithöchsten Prägnanz aus. Auf diese Weise wird die Aufmerksamkeit zu Bildregionen absteigender Prägnanz gelenkt.

3.7. Validierung des Modells

Das Modell wird auf Bildern wie z.B. Bild 1 getestet. Im linken Teilbild weisen alle Balken eine horizontale Orientierung auf, was sich in einer gleichmäßigen Antwort des Orientierungsfilters zeigt. Durch die Normierung der Merkmalskarten wird dieser gleichmäßige Anteil abgeschwächt. Der Filter für den rot-grün-Kontrast hingegen zeigt einen deutlich erhöhten Wert an der Stelle des roten Balkens. Da sich die grünen Balken nicht von ihren Nachbarn unterscheiden, liefern sie kleinere Werte. Durch die Normierung tritt der rot-grün-Kontrast des roten Balkens in der *Saliency Map* deutlich hervor und die Aufmerksamkeit wird auf dieses Objekt gelenkt.

Im rechten Teilbild treten Objekte verschiedener Farben und Orientierungen auf. Dadurch enthält keine der Merkmalskarten nach der Normierung einen herausragenden Wert an der Position des roten horizontalen Balkens. Sequenziell wird die Aufmerksamkeit durch Maximums-Findung in der *Saliency Map* auf Bildbereiche absteigender Prägnanz gelenkt und diese analysiert, bis das gesuchte Objekt gefunden wird.

4. Zusammenfassung

Der *pop-out-Effekt* im ersten Beispiel beruht also auf dem Auftreten eines exklusiven Wertes in einer der Merkmalskarten. Dieser prägnante Wert wirkt sich durch Normierung und Kontrastverbesserung als Maximum der *Saliency Map* aus und die Aufmerksamkeit wird direkt auf diesen Bereich gelenkt. Die Suchzeit ist somit unabhängig von der Anzahl der Objekte.

Im zweiten Beispiel treten Objekte in verschiedenen Farben und Orientierungen auf. Die Vielzahl der rot-grün-Kontraste und die Kontrastverbesserung der Merkmalskarten schwächen den zuvor markanten Wert an der Stelle des gesuchten Objekts ab. Diese Abschwächung wirkt sich mit zunehmender Objektanzahl stärker aus. Die Position des Zielobjekts tritt somit nicht als deutliches Maximum aus der *Saliency Map* hervor, der *pop-out-Effekt* bleibt aus. In diesem Fall muss die Aufmerksamkeit wie ein Scheinwerfer nacheinander auf verschiedene Bildbereiche gelenkt werden um diese analysieren zu können und das Zielobjekt zu finden. Die Suchzeit nimmt also mit der Objektanzahl zu.

5. Quellenverzeichnis

[1] P.J. Burt, E.H. Adelson. The Laplacian pyramid as a compact image code. IEEE Trans. on Communications, 31(4): 532-540, 1983

[2] R.C. Gonzalez, R.E. Woods. Digital Image Processing. Addison-Wesley. Reading, MA, 1993.

[3] L. Itti, C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research 40 (2000) 1489 - 1506

[4] L. Itti, C. Koch. Computational modelling of visual attention. Nature Reviews Volume 2 (02.2001)