Una implementación computacional de un modelo de atención visual *Bottom-up* aplicado a escenas naturales

A Computational Implementation of a Bottom-up Visual Attention Model Applied to Natural Scenes

David F. Ramírez-Moreno[(1)*], Juan F. Ramírez-Villegas[(2)*]
[(1)] PhD. dramirez@uao.edu.co
[(2)] BEng. juanfelipe.rv@gmail.com
[*] Universidad Autónoma de Occidente. Cali. Colombia

Recibido julio 13 de 2010, aprobado noviembre 18 de 2011.

Palabras claves
Neurobiología computacional, visión, sistema visual.

Key words
Computational neurobiology, vision, visual system.

Resumen
El modelo de atención visual *bottom-up* propuesto por Itti et al., 2000 [1], ha sido un modelo popular en tanto exhibe cierta evidencia neurobiológica de la visión en primates. Este trabajo complementa el modelo computacional de este fenómeno desde la dinámica realista de una red neuronal. Asimismo, esta aproximación se basa en la existencia de mapas topográficos que representan la prominencia de los objetos del campo visual para la formación de una representación general (mapa de prominencia), esta representación es la entrada de una red neuronal dinámica con interacciones locales y globales de colaboración y competencia que convergen sobre las principales particularidades (objetos) de la escena.

Abstract
The bottom-up visual attention model proposed by Itti et al. 2000 [1], has been a popular model since it exhibits certain neurobiological evidence of primates' vision. This work complements the computational model of this phenomenon using a neural network with realistic dynamics. This approximation is based on several topographical maps representing the objects saliency that construct a general representation (saliency map), which is the input for a dynamic neural network, whose local and global collaborative and competitive interactions converge to the main particularities (objects) presented by the visual scene as well.

Introducción

El cerebro de los primates emplea algún procesamiento visual en serie, de la mano con el procesamiento masivo en paralelo [1]. Existe mucha evidencia experimental acumulada a favor de la existencia de dos mecanismos de control sobre los que la atención visual se desarrolla [2-5]. El primero de ellos es conocido como procesamiento *bottom-up* o proceso de pre-atención dependiente de la prominencia de los rasgos de los objetos e independiente de la tarea. El segundo es conocido como procesamiento *top-down* o proceso de atención, mucho más lento que el anterior, controlado por la voluntad y por tanto, dependiente de la tarea específica en ejecución. Dichos procesamientos se dan sobre la concepción de una representación neuronal de orden alto que selecciona parte de la información sensorial disponible, probablemente para reducir la complejidad del análisis de las escenas [6], esta selección se hace en forma de una región espacial circunscrita conocida como "foco de atención", éste cambia de una locación a otra de forma serial.

En los últimos años se han establecido diferentes modelos de procesamiento *bottom-up* [1], [6]-[13], que reproducen el comportamiento del mecanismo neurobiológico dadas las hipótesis establecidas por Treisman et al. [3], según las cuales las diferentes propiedades del espacio son codificadas en mapas de características diferentes en distintas regiones del cerebro. De acuerdo a este modelo, para resolver el problema de las vinculaciones (asociaciones) de los rasgos codificados por separado, hay un mapa de prominencia (*saliency map*) que codifica conjunciones de características en la imagen. Este mapa maestro recibe entradas desde todos los mapas de características, pero retiene solamente las que distinguen el objeto de lo que lo rodea, de modo tal que las características específicas y detalladas se quedan en los mapas de caracte-

rísticas iniciales (las que sirven para reconocer el objeto). De igual manera, la escena es susceptible a un proceso de atención o búsqueda fina, sólo después de que las características hayan sido asociadas en una porción del mapa maestro. La mayoría de modelos asumen que el fenómeno de atención visual opera sobre representaciones primarias, i.e., mapas corticales topográficos que codifican el espacio visual [14].

En este trabajo se implementa un modelo de atención visual *bottom-up* basado en representaciones primarias, i.e., una variedad de mapas que codifican diferentes características elementales (orientación, color e intensidad) y su convergencia sobre una representación general o mapa de prominencia. Adicionalmente, en ausencia de procesamiento *top-down* y con el ánimo de hacer el modelo lo más biológicamente plausible, se usa una red neuronal *winner-take-all* (WTA), cuya dinámica interna garantiza la generación de los cambios de atención sobre diferentes locaciones del mapa de prominencia.

Método

El modelo de atención visual *bottom-up* propuesto en este trabajo se basa en la aproximación de Itti et al. [1], [6] (figura 1). Sin embargo, difiere en dos puntos importantes: como primera medida, se establece un modelo complementado del sistema de color doble-oponente propuesto, dada la evidencia experimental establecida por Conway, 2004 [15].

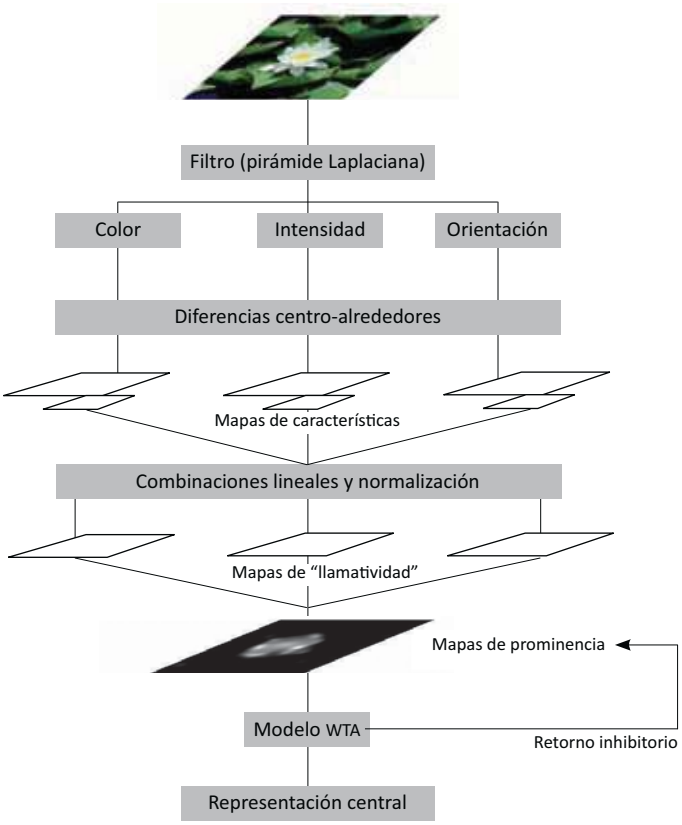

Figura 1. Esquema general del modelo de atención visual propuesto

El segundo se refiere a que, en tanto el modelo de atención visual descrito por Itti et al. [1], [6] en ausencia de procesamiento *top-down*, implementa una red WTA basada en el modelo LIF (*Leak Integrate and Fire*). En este trabajo se usa una red WTA utilizando la función de Naka-Rushton [5], [16] para describir los disparos de potencial de las neuronas, lo que resuelve el problema de los cambios en las locaciones atendidas por el modelo.

Extracción de características visuales primarias

Las características visuales de bajo nivel son extraídas directamente de la imagen en el color original sobre distintas escalas espaciales, para esto se utilizan filtros lineales en forma de pirámide, i.e., pirámides Gaussianas [17]. Una vez que se han calculado las pirámides Gaussianas, cada característica es calculada en una estructura centro-alrededores (*center-surround*) relacionada estrechamente con los campos receptivos visuales. Las diferencias centro-alrededores son realizadas entre escalas amplias y finas para cada característica específica: El centro receptivo corresponde a un píxel al nivel $c \in \{2, 3, 4\}$ en la pirámide y los alrededores al píxel correspondiente en el nivel $s = c + \delta$, con $\delta \in \{3, 4\}$.

Mapas de Intensidad y Orientación

El primer grupo de mapas de características está relacionado con la intensidad de contraste, que en mamíferos es detectado por neuronas sensibles a centros oscuros sobre fondos luminosos o viceversa. Estos dos tipos de sensibilidad son calculados utilizando (1).

$$I(c, s) = | I(c) \ominus I(s) |, \quad (1)$$

Donde $I(c)$ es la señal de intensidad de centro, $I(s)$ es la señal de intensidad de alrededores y el símbolo "Θ" corresponde a la operación de resta punto a punto entre diferentes escalas, llevando la imagen al nivel más fino.

Los mapas de orientación son extraídos utilizando pirámides de Gabor $O(\theta, \sigma)$, donde $\theta \in \{0^o, 45^o, 90^o, 135^o\}$ [18]. De esta forma, se establece el contraste de orientación entre las escalas de centro y alrededores según (2).

$$O(c, s, \theta) = | O(c, \theta) \ominus O(s, \theta) |, \quad (2)$$

Donde $O(c, \theta)$ y $O(s, \theta)$ son las señales de orientación de centro y alrededores, respectivamente.

Sistema de Color Doble Oponente Complementado

La evidencia neurobiológica establece que los puntos de luz que selectivamente modulan cada clase de cono (L, M ó S, o de forma imprecisa rojo, verde o azul) son destellados alrededor de los campos receptivos de las células de color V1 para mapear la estructura espacial de las entradas. Evidencia experimental [15], [19], sugiere que el procesamiento del color es mediado por un mecanismo antagónico.

Conway [15], ha establecido y documentado la existencia de células sensibles a otra clase de contraste de color. Según sus experimentos, existen clases de células sensibles a longitudes de onda grandes en el centro (L), a la vez que sensibles a longitudes de onda medias y bajas en la periferia y viceversa, i.e., para estímulos rojo-*on* (L+) o amarillo-*on* (S-), el centro de la célula se excita siendo cromáticamente oponente al estímulo verde-*on* (M+), al igual que el estímulo azul-*on* (S+) en la periferia o alrededores; de forma similar ocurre para las células con centro sensible a estímulos verdes. Dado este antagonismo y que los estímulos para las células M y S se muestran alineados, se sugiere la existencia de células rojo-cyan, y de forma análoga se propone la existencia de células verde-magenta.

Teniendo en cuenta que el antagonismo se da entre los colores rojo-verde, azul-amarillo, rojo-cyan y verde-magenta, se establecen los canales de color correspondientes y se construyen los mapas $RG(c, s)$, $BY(c, s)$, $RC(c, s)$, y $GM(c, s)$ respectivamente, según (3) a (10).

$$R = r - \frac{(g + b)}{2}, \tag{3}$$

$$G = g - \frac{(r + b)}{2}, \tag{4}$$

$$B = b - \frac{(g + r)}{2}, \tag{5}$$

$$Y = \frac{(r + g)}{2} - \frac{|r - g|}{2} - b, \tag{6}$$

$$RG(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|, \tag{7}$$

$$BY(c, s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|, \tag{8}$$

$$RC(c, s) = |(R(c)+Y(c)-G(c)-B(c)) \ominus (G(s)+B(s)-R(s)-Y(s))|, \tag{9}$$

$$GM(c, s) = |(G(c)+Y(c)-R(c)-B(c)) \ominus (R(s)+B(s)-G(s)-Y(s))|, \tag{10}$$

Donde las variables R, G, B y Y corresponden a los canales de color rojo, verde, azul y amarillo, respectivamente. Para el cálculo de estos canales, todos los valores resultantes por debajo de cero son llevados automáticamente a cero. $R(c)$, $G(c)$, $B(c)$ y $Y(c)$ son las señales de centro correspondientes a los canales de color rojo, verde, azul y amarillo, respectivamente. De forma análoga, $R(s)$, $G(s)$, $B(s)$ y $Y(s)$ son las señales de alrededores correspondientes a los canales de color rojo, verde, azul y amarillo, respectivamente.

Mapa de prominencia y representación central

Una vez obtenidos los 54 mapas de características resultantes, se realizan combinaciones lineales entre mapas del mismo tipo. De esta forma se obtienen 3 mapas de características llamativas (*conspicuity maps*) y de una combinación lineal de estos, se obtiene el mapa de prominencia final. Este procedimiento es mostrado por (11) a (14)

$$\square = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(I(c, s)), \tag{11}$$

$$\bar{C} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [N(RG(c, s)) + N(BY(c, s))], \tag{12}$$

$$\square = \sum_{\square \in \{0°, 45°, 90°, 135°\}} N\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(O(c, s, \square)) \right), \tag{13}$$

$$S = \frac{1}{3}(\square + \bar{C} + \square), \tag{14}$$

Donde $\square$, $\bar{C}$ y $\square$ son los mapas de características llamativas de intensidad, color y orientación, respectivamente y S es el mapa de prominencia final. El papel de la función $N(\cdot)$ dentro de las ecuaciones es normalizar cada uno de los mapas de prominencia iterativamente, de tal forma que se establezca competencia entre las unidades del mapa de prominencia completo.

En ausencia de procesamiento *top-down*, la presente aproximación implementa una red WTA en el espacio 2D de la imagen. El objetivo principal de esta red es dar una interpretación de los datos de búsqueda visual respecto al cambio del foco de atención. En un instante de tiempo determinado la máxima prominencia dentro del mapa final definirá la locación más saliente, a ésta estará dirigida la atención, luego, en instantes de procesamiento posteriores, dicha locación es inhibida para que la atención se desvíe hacia otros objetos o puntos salientes en el campo visual [20]-[21]. Si se utiliza la función de Naka-Rushton para describir las frecuencias de disparo de las neuronas, las ecuaciones de la red estarían dadas por (15) y (16).

$$\tau \frac{dT}{dt} = -T + \frac{C\left(E_T - kND + U_{SE}^{L} \square_j w_{ij}^{L} E_{T_j}\right)_+^2}{\sigma^2 + \left(E_T - kND + U_{SE}^{L} \square_j w_{ij}^{L} E_{T_j}\right)_+^2}, \tag{15}$$

$$\tau \frac{dD}{dt} = -D + \frac{C\left(E_D - k(N-1)D - kT + U_{SE}^{L} \square_j w_{ij}^{L} E_{D_j}\right)_+^2}{\sigma^2 + \left(E_D - k(N-1)D - kT + U_{SE}^{L} \square_j w_{ij}^{L} E_{D_j}\right)_+^2}, \tag{16}$$

Siendo: $C = 100.0$; $k = 3.0$; $\sigma = 120.0$; $\tau = 20.0$; $U_{SE}^{L} = 255.0$

Donde T es la respuesta de cualquier neurona que reciba información acerca del objetivo (locación de mayor prominencia), D es la respuesta de cada una de las N neuronas

$$w_{ij}^{L} = \frac{w_E}{2\pi\sigma_{E1}^2\sigma_{E2}^2}\exp\left[-\left(\frac{x^2}{2\sigma_{E1}^2}+\frac{y^2}{2\sigma_{E2}^2}\right)\right]-\frac{w_I}{2\pi\sigma_{I1}^2\sigma_{I2}^2}\exp\left[-\left(\frac{x^2}{2\sigma_{I1}^2}+\frac{y^2}{2\sigma_{I2}^2}\right)\right], \quad (17)$$

de distracción, la constante k determina la fuerza de la inhibición hacia atrás (competencia entre neuronas dentro del mapa), las constantes C, τ, σ y U_{SE}^{L} son parámetros ajustables del sistema, donde C es la frecuencia de disparo máxima del sistema, la constante de tiempo τ controla la frecuencia a la que las variables T y D tienden a su valor máximo (en milisegundos), el parámetro σ es la constante de semi-saturación del sistema y U_{SE}^{L} define la eficacia sináptica entre las conexiones neuronales intracapa. Para todos los efectos, el estímulo $E_T > E_D$. Los parámetros de este modelo se ajustaron de acuerdo a lo que establece la literatura [5], [10].

Los pesos de conexión lateral entre las poblaciones neuronales dentro del mapa de prominencia se calculan según (17).

Donde w_E=5.0; w_I=250.0; σ_{E1}=0.05*X_T; σ_{E2}=0.05*Y_T σ_{I1}=0.75*X_T; σ_{I2}=0.75*Y_T, donde X_T y Y_T son el número total de filas y columnas del mapa de prominencia.

Los parámetros libres son constantes que definen la forma de las funciones gaussianas respecto a las distancias x e y. Este esquema garantiza la existencia de interacciones competitivas centro-alrededores entre las características dentro del mapa de prominencia, acentúa la diferencia de actividad entre la prominencia del objetivo y los distractores y media en la sobrecompetencia que potencialmente presentarían las poblaciones neuronales.

Resultados

El modelo de atención visual *bottom-up* detallado en este trabajo fue implementado en MatLab™ ver. 7.6.0. Asimismo los resultados de esta implementación se muestran en la figura 2. Las imágenes de escenas naturales utilizadas para poner a prueba el modelo fueron tomadas de la *MSRA Salient Object Database* [22], que contiene alrededor de 4500 imágenes en las que confluyen objetos sobresalientes en un rango diverso de complejidad de detección, cada imagen contiene 300 x 400 píxeles aproximadamente.

Desempeño general del modelo y comparación entre sistemas de color doble oponente

Para medir el desempeño del modelo, se midió la consistencia del etiquetado de las imágenes de la base de datos siguiendo el procedimiento detallado en [22]. Se escogió una muestra de las imágenes que presentaron alta consistencia en el etiquetado (de 80.0% a 100.0%) y finalmente se tomó una muestra aleatoria de 700 de estas imágenes. Este procedimiento fue realizado con el fin de obtener una medida del desempeño objetiva para escenarios reales, dado que en otros estudios esto ha sido catalogado como un proceso altamente subjetivo [1], [6], [10].

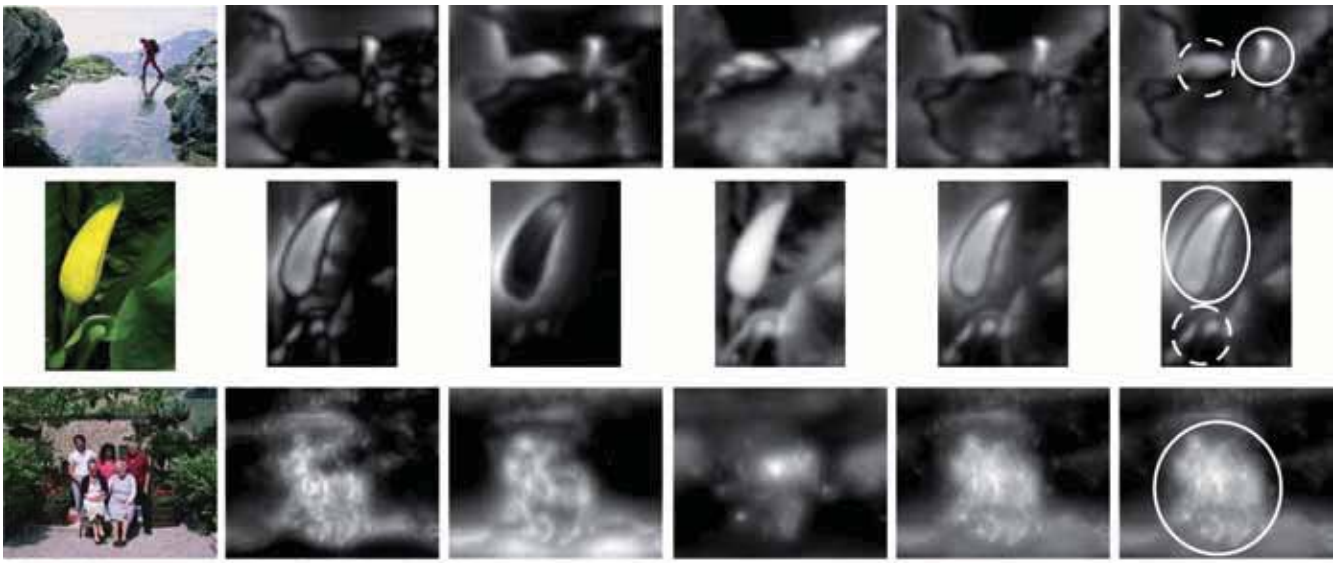

Figura 2. Resultados del modelo de atención visual *bottom-up*: (a) Imagen de entrada; (b) Mapa de intensidad; (c) Mapa de orientación; (d) Mapa de color; (e) Mapa de prominencia; (f) Locaciones atendidas por el modelo en alrededor de 100 ms (línea sólida) y en un rango de 200 a 400 ms (línea punteada)

Se tuvieron en cuenta las cantidades de verdaderos positivos y falsos positivos arrojadas en la prueba de desempeño para calcular el índice de exactitud (Ac), que representa la probabilidad de que el sistema entregue una respuesta adecuada ante un estímulo determinado. En términos generales el modelo presenta un desempeño satisfactorio respecto a las fijaciones esperadas por un sistema visual humano (para el caso de escenas naturales, Ac: 76.00% y para el caso de imágenes sintéticas, aunque no se midió formalmente en este trabajo, Ac: ~100.00% [1], [6]).

Un aspecto importante que se evidencia en nuestro modelo es la implementación del sistema de color doble oponente complementado. La influencia de este sistema sobre la prominencia de los objetos en escenas naturales es ilustrada por la figura 3. Nótese que existe una mejora significativa en los promedios de prominencia en la diferenciación entre el objetivo y los distractores, esto muestra, además de plausibilidad neurobiológica, que el modelo implementado robustece los mapas de color en la búsqueda de objetos sobresalientes.

Por otra parte, la red neuronal WTA implementada en este modelo mejora los resultados de la red realizada en [5], amén de las interacciones locales entre las poblaciones de neuronas, cuyo efecto neto es el de colaboración entre centros cercanos y competencia entre centros lejanos. Igualmente, este esquema es coherente con los pesos relativos (respecto al procesamiento visual) del centro de una escena y sus alrededores [8].

Figura 3. Imágenes para la comparación entre sistemas de color doble oponente: (a) Imagen de entrada; Sistema de color doble oponente: (b) clásico; (c) complementado

Conclusiones

El control de la atención asociado al procesamiento *bottom-up* en primates ha sido modelado utilizando representaciones primarias sobre características básicas inherentes a los objetos del campo visual. Estas representaciones fueron calculadas de acuerdo a otros modelos establecidos en la literatura [1], [5]-[6], [10]-[13]. Sin embargo, en esta aproximación se integra un sistema de color doble oponente modificado, que contribuye a una representación completa de los mapas de color operando sobre los planos básicos R, G y B sugerida por la literatura [15]. Este sistema de color doble oponente probó mejorar con suficiencia los mapas de color estándar hasta ahora conocidos en la literatura e implementados computacionalmente. Adicionalmente, el control de cambio de foco de atención ha sido modelado utilizando una red neuronal WTA, cuya dinámica integra interacciones locales y competencia entre las unidades encargadas de dar la representación que reúne todas las características básicas de una escena. El esquema propuesto presentó una exactitud alta en imágenes de escenas naturales (76.00%), asimismo es neurobiológicamente plausible, pues engloba los hallazgos experimentales descritos hasta el momento en la literatura. Finalmente, este modelo constituye una fuerte base para diversas aplicaciones en visión artificial.

Referencias bibliográficas

[1] L. Itti, C. Koch. “A saliency-based search mechanism for overt and covert shifts of visual attention”. *Vision Research*, Vol. 40, 2000, pp. 1489–1506.

[2] J.E. Hoffman. “Search through a sequentially presented visual display”. *Perception & Psychophysics*, Vol. 23, 1978, pp. 1-11.

[3] A. Treisman, M. Sykes, G. Gelade. “Selective attention stimulus integration”. In *Attention and performance VI*. S. Dornie (ed.). Eds. N. J. Hilldale: Lawrence Erlbaum, 1977, pp. 333-361.

[4] P. Verghese, K. Nakayama. “Stimulus discriminability in visual search”. *Vision Research*, Vol. 34, 1994, pp. 2453-2467.

[5] H.R. Wilson. “Computation by excitatory and inhibitory networks”. In *Spikes, Decisions and Actions: The Dynamical Foundations of Neuroscience*. H.R. Wilson (ed.). Oxford University Press, 2004, pp. 89-115.

[6] L. Itti, C. Koch, E. Niebur. “A Model of Saliency-Based Visual Attention for Rapid Scene Analysis”. *IEEE Trans. Patt. Anal. Mach. Intel*, Vol. 20, 1998, pp. 1254-1259.

[7] R.J. Peters, A. Iyer, L. Itti, C. Koch. “Components of bottom-up gaze allocation in natural images”. *Vision Research*, Vol. 45, 2005, pp. 2397-2416.

[8] P.J. Burt. "Attention Mechanisms for vision in a dynamic world". *Proceedings of 9th International Conference on Pattern Recognition*, 1988, pp. 977-987.

[9] Z. Li. "A saliency map in primary visual cortex". *Trends in Cognitive Science*, Vol. 6, 2002, pp. 9-16.

[10] M. DeBrecht, J. Saiki. "A neural network implementation of a saliency map model". *Neural Networks*, Vol. 19, 2006, pp. 1467–1474.

[11] D. Gao, V. Mahadevan, N. Vasconcelos. "The discriminant center-surround hypothesis for bottom-up saliency". *Proceedings of Neural Information Processing Systems*, Vancouver, Canada, 2007.

[12] J.H. Reynolds, D.J. Heeger. "The normalization model of attention". *Neuron*, Vol. 61, 2009, pp. 168-185.

[13] S.J. Park, K.H. An, M. Lee. "Saliency map model with adaptive masking based on independent component analysis". *Neurocomputing*, Vol. 49, 2002, pp. 417-422.

[14] R. Desimone, J. Duncan. "Neural mechanisms of selective visual attention". *Annu. Rev. Neurosci*., Vol. 18, 1995, pp. 193-222.

[15] B.R. Conway. "Spatial structure of cone inputs to color cells in alert macaque primary visual cortex (V-1)". *The Journal of Neuroscience*, Vol. 21, 2004, pp. 2768-2783.

[16] K.I. Naka, W.A. Rushton. "S-potentials from colour units in the retina of fish". *J. Physiol*, Vol. 185, 1966, pp. 584-599.

[17] P.J. Burt, E. H. Adelson. "The Laplacian pyramid as a compact image code". *IEEE Trans. Com*., Vol. 31, 1983, pp. 532-540.

[18] H. Greenspan, S. Belongie, R. Goodman, P. Perona, S. Rakshit, C. H. Anderson. "Overcomplete Steerable Pyramid Filters and Rotation Invariance". *Proc. IEEE Computer Vision and Pattern Recognition*, 1994, pp. 222-228.

[19] S. Engel, X. Zhang, B. Wandell. "Colour Tuning in Human Visual Cortex Measured With Functional Magnetic Resonance Imaging". *Nature*, Vol. 388, 1997, pp. 68–71.

[20] R.M. Klein. "Inhibition of return". *Trends Cogn. Sci.*, Vol. 4, 2000, pp. 138–147

[21] S.L. Macknik, S. Martinez-Conde. "The role of feedback in visual attention and awareness". *Cognitive Neurosciences*, Ed. Gazzinga, MIT Press, 2009.

[22] T. Liu, J. Sun, N.N. Zheng, X. Tang, H.Y. Shum. "Learning to detect a salient object". In: *Proceedings of IEEE Computer Society Conference on Computer and Vision Pattern Recognition*, 2007.