PRÁCTICA 5: Modelos de salencia espectrais e a súa avaliación

Xosé R. Fdez-Vidal
xose.vidal@usc.es

Grao de Robótica. EPSE de Lugo — 24 de outubro do 2022

1. Introdución

A *relevancia visual ou saliencia* é un termo técnico da psicoloxía cognitiva que trata de describir a capacidade de determinados obxectos ou puntos de atraer a nosa atención visual inmediatamente. O cerebro constantemente dirixe a nosa mirada cara as rexións importantes da escena visual e as mantenos rastrexadas ao longo do tempo, o que nos permite explorar rapidamente o noso entorno en busca de obxectos e acontecementos de interese, sen ter en conta as partes menos importantes. Un exemplo dunha imaxe RGB normal e a súa conversión nun mapa de relevancia, onde o as rexións relevantes estatisticamente aparecen brillantes e as outras escuras (Fig. 1).

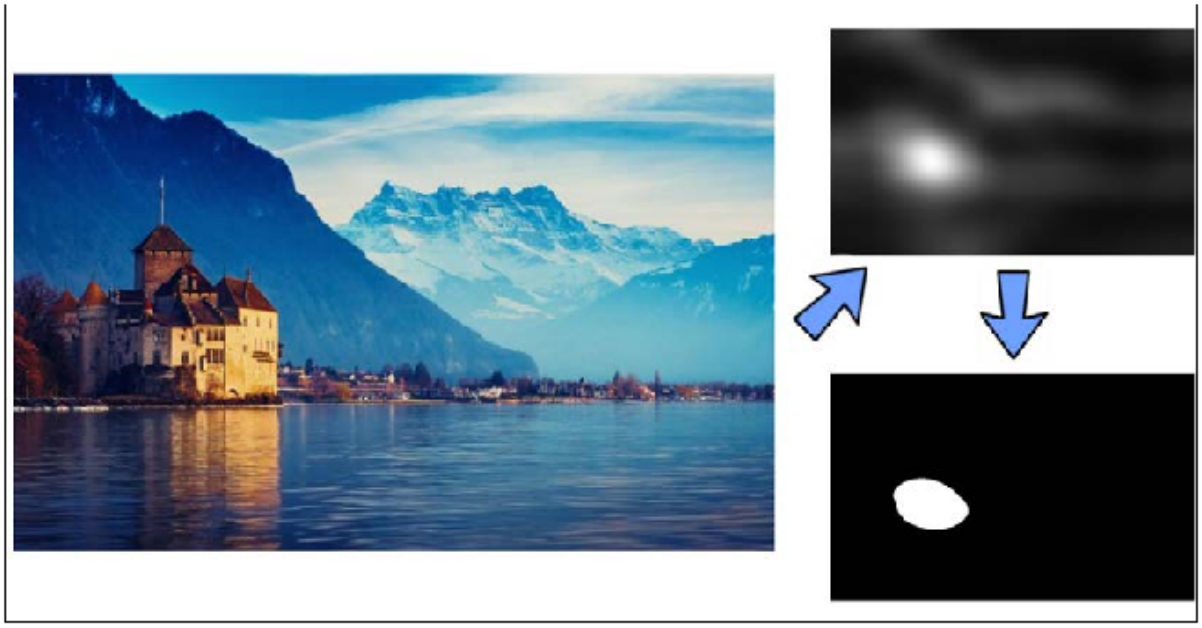

Figura 1: Imaxe xunto co seu mapa de saliencia visual e a segmentación obtida a partir do citado mapa.

Pénsase que esta é unha estratexia evolutiva para tratar a información que constante chega aos nosos ollos, desbotando a non relevante e evitando unha saturación do noso sistema visual. Por exemplo, se por casualidade decides ir a dar un paseo pola selva, queres asegurarte de que no entorno non hai un puma disposto a atacarte antes de adicar toda a túa atención a admirar unha bonita bolboreta que se pousou nunha folla diante de ti. Como resultado, os obxectos visualmente salientables teñen a capacidade de destacar do seu entorno, ao igual que as barras xiradas ou de distinto cor destacan entre os distractores do seu entorno (ver Fig. 2):

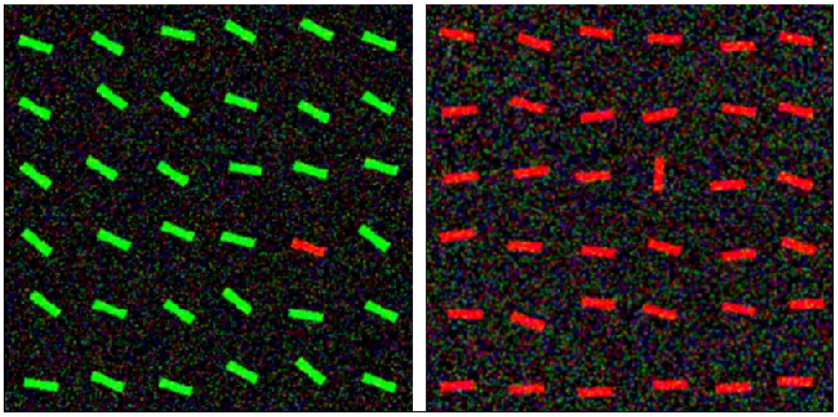

Figura 2: Experimento pop-out onde o estímulo obxectivo destaca fortemente entre os distractores circundantes.

As características visuais que fan que determinados eventos destaquen case nunca son triviais en experimentos máis complexos. Se miras a imaxe da esquerda en cor da Fig. 2, podes percibir inmediatamente a barra vermella na imaxe posto que é moi distinta dos estímulos circundantes. Non obstante, se miras a mesma imaxe en escala de grises (Fig. 3), a barra "obxectivo" será difícil de atopar (é a cuarta barra desde a parte superior, a quinta barra desde a esquerda).

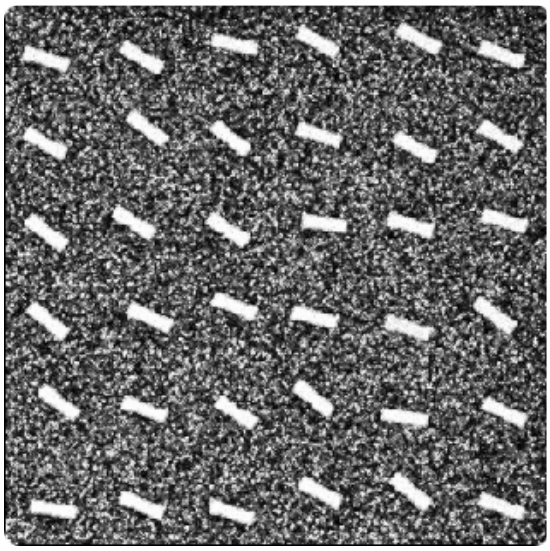

Figura 3: Experimento pop-out en escala de grises. O efecto salientable (a cor) desaparece ao eliminar esa característica visual da imaxe.

Semellante á cor relevante, hai unha barra perceptualmente salientable na imaxe da dereita da Fig. 2. A cor agora non dirixe a atención senón que é a orientación o que a diferencia do entorno (distractores). Pero ningunha destas características mencionadas funciona tan rápido, se miras a Fig. 4. Ao mesturar as dúas características (precisamos integrar cor e orientación) o efecto pop-out desaparece e precisamos integrar en cada posición cor e orientación (enlentecendo notablemente o proceso) para atopar o estímulo diferente:

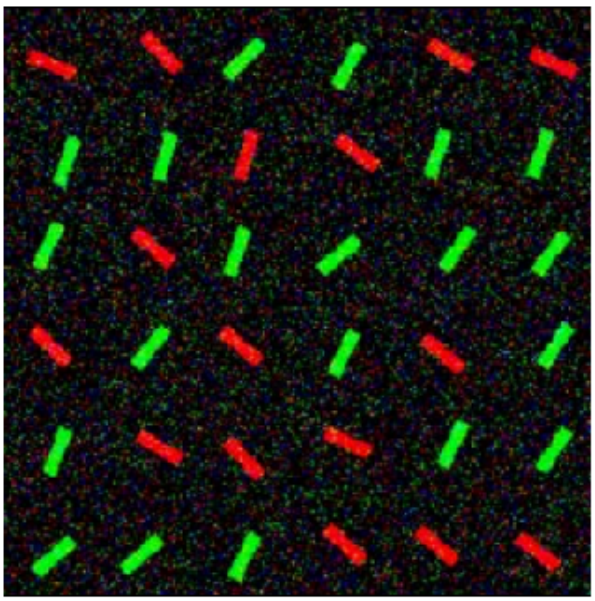

Figura 4: Influencia da contorna define a relevancia dun obxecto na imaxe (pop-out).

Na Fig. 4, hai de novo unha barra que é distinta e diferente de todas as outras. Non obstante, debido á forma en que se deseñaron os elementos distractores, hai pouca saliencia para guiar a atención cara á barra de obxectivo. Probablemente, estarás inspeccionando a imaxe, aparentemente ao azar, buscando algo interesante [1].

Que teñen que ver estes experimentos coa visión artificial? moito, en realidade. Os sistemas de visión artificial sofren unha sobrecarga de información ao igual o sistema visual humano, coa agravante de que non teñen experiencia "vital" do entorno ao que se enfrontan. Polo tanto, a pregunta é: poderíamos extraer algunhas ideas da bioloxía e utilízalas para programar os nosos algoritmos? Imaxina unha cámara no cadro de mandos do teu coche, que enfoca automaticamente os sinais de tráfico relevantes en cada instante ou unha cámara de vixilancia dunha estación de observación da vida salvaxe que detectará e rastrexará automaticamente o avistamento dos tímidos ornitorrincos ignorando todo o demais. Como podemos ensinarlle ao algoritmo o que é importante e o que non?

2. Formulación do problema

Como xa dixemos en teoría, hai moitas aproximacións ao problema da saliencia visual. A grandes liñas, poderíamos clasificalos en tres grandes grupos: bio-inspirados, estatísticos e frecuenciais. Neste curso, centrámonos nos frecuenciais que son os que teñen tempos de execución moi baixos e poden ter utilidade no campo da robótica; aínda que os seus rendementos sexan inferiores a outros enfoques computacionalmente moito máis custosos.

O cerebro humano descubriu, debido á evolución, como concentrarse en obxectos visualmente salientables. As escenas naturais teñen unhas regularidades estatísticas singulares, que as diferencian de outras máis artificiosas como un patrón de taboleiro de xadrez. Esta regularidade estatística é coñecida como a lei do $1/f$ e afirma que a amplitude, en promedio para conxunto amplo de imaxes naturais, obedece a unha distribución $1/f$ (ver Fig. 5). Polo tanto, podemos asumir que a distribución de enerxía a distintas frecuencias permanece case invariante. Ademais, esta lei potencial é unha manifestación da *natureza invariante a escala* das imaxes, que indica que a estatísticas da imaxe permanece invariante ante cambios de escala.

[1] spoiler: o obxectivo é a única barra vermella e case vertical da imaxe, a segunda fila dende a arriba, terceira columna desde a esquerda

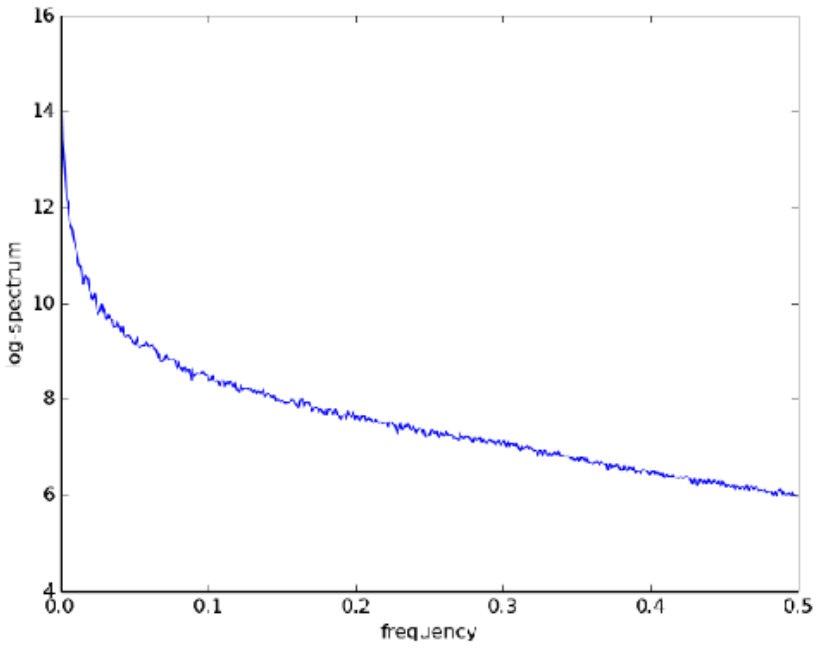

Figura 5: Lei de $1/f$ á que obedecen todas as imaxes naturais.

A pregunta agora é a seguinte: como podemos utilizar o noso coñecemento da estatística das imaxes naturais para dicirlle a un algoritmo que nunha imaxe, como a da Fig. 6, non mire para a árbore da esquerda, senón que se centre no barco que atrae a nosa atención?

Figura 6: Imaxe natural dunha paisaxe fluvial.

As cousas que merecen a nosa atención nunha imaxe non son as zonas que cumpren a lei $1/f$, senón

aquelas que se separan desa tendencia xeral: as **anomalías estatísticas**. Este peculiaridade é a que denominaremos **residuo espectral** [1], e corresponden aos parches potencialmente interesantes dunha imaxe (ou protoobxectos). Un mapa que mostra estas anomalías estatísticas como puntos brillantes chámase mapa de relevancia ou de saliencia (Fig. 7).

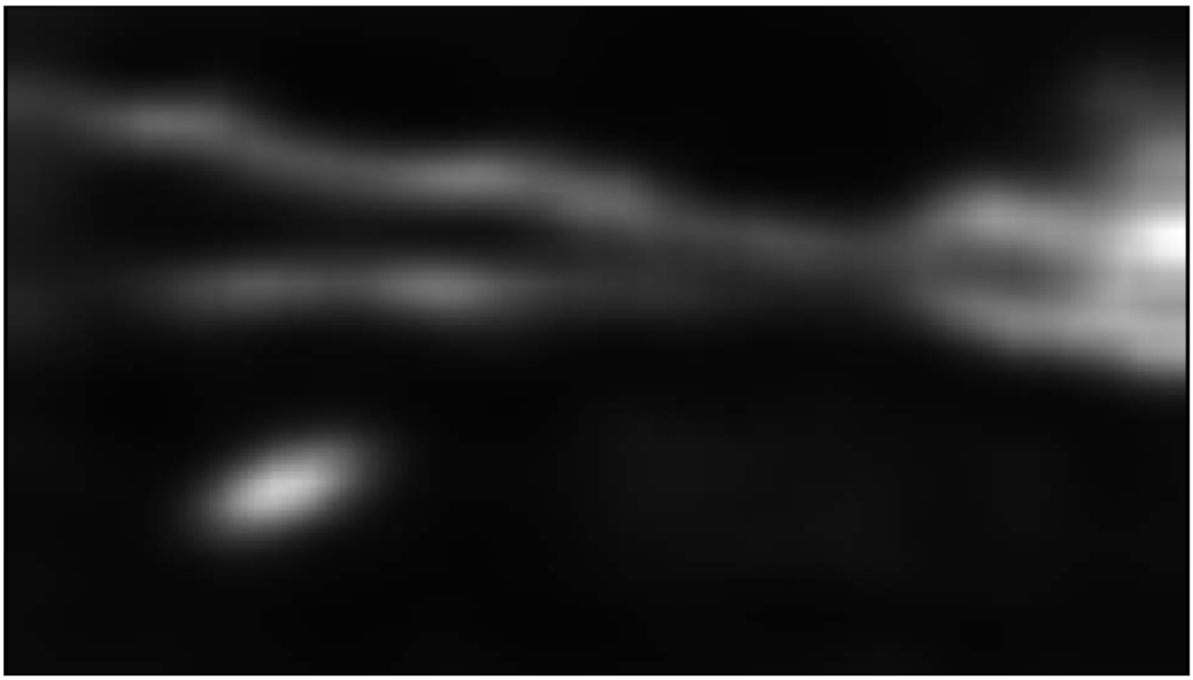

Figura 7: Mapa de relevancia ou saliencia da imaxe da Fig. 6.

3. Algoritmos de saliencia a partir irregularidades globais

Saliencia Estática: estes algoritmos se centran en buscar irregularidades globais a partir da información espectral e sen ter en conta a clave visual do movemento.

- **Residuo espectral Hou e Zhang, [1]** Inspirados polo descubrimento sobre a estadísticas das imaxes naturais, Hou e Zhang, [1] propuxeron obter o mapa de saliencia visual en tres pasos: 1) transformar sinais espaciais ao dominio frecuencial para obter o espectro de amplitude logarítmica; 2) eliminación da varianza no espectro de amplitude logarítmica para obter o residuo espectral; e 3) transformar o residuo espectral de volta ao dominio espacial para localizar o subconxuntos de rexións visualmente salientables. Os detalles dos tres pasos explicámolos a continuación. Sexa I a canle de intensidade da imaxe de entrada e $\mathcal{F}[\cdot]$ a Fourier. A imaxe transformase a unha resolución de 64×64 segundo o natureza invariante de escala das imaxes[2] e a pasamos ao dominio transformado:

$$\mathcal{A}(\omega) = |\mathcal{F}[I]|\,, \qquad \mathcal{P}(\omega) = \varphi(\mathcal{F}[I]). \tag{1}$$

onde $\mathcal{A}(\omega)$ e $\mathcal{P}(\omega)$ son os espectros de amplitude e fase da imaxe, respectivamente. Neste algoritmo, $\mathcal{A}(\omega)$ transfórmase no espectro logarítmico $\mathcal{L}(\omega) = \log(\mathcal{A}(\omega))$. Dado o espectro logarítmico, pódese derivar o residuo espectral eliminando as compoñentes invariantes a partir del. Para obter a invarianza no dominio da frecuencia, Hou e Zhang propuxeron empregar o espectro logarítmico das amplitudes e restarlle unha versión suavizada no entorno cun filtro da media, $h(\omega)$, de tamaño 3×3 ,:

$$\mathcal{R}(\omega) = \mathcal{L}(\omega) - h(\omega) * \mathcal{L}(\omega). \tag{2}$$

[2]A estadística dunha imaxe natural é independente da escala

Fíxate que o espectro de fase permanece inalterado. Finalmente, o mapa de saliencia no dominio espacial podes ser obtido como:

$$S = G(0,\sigma) * \left(\mathcal{F}^{-1}[\mathcal{R}(\omega)\exp(j\mathcal{P}(\omega))]\right)^2. \quad (3)$$

onde $G(0,\sigma)$ é un filtro gaussiano con $\sigma = 8$ para suavizar o mapa de saliencia. Na Fig. 8, podemos ver que este algoritmo funciona ben con zonas salientes pequenas (altas frecuencias) e relevantes en intensidade (na súa concepción inicial, non traballa con imaxes de cor).

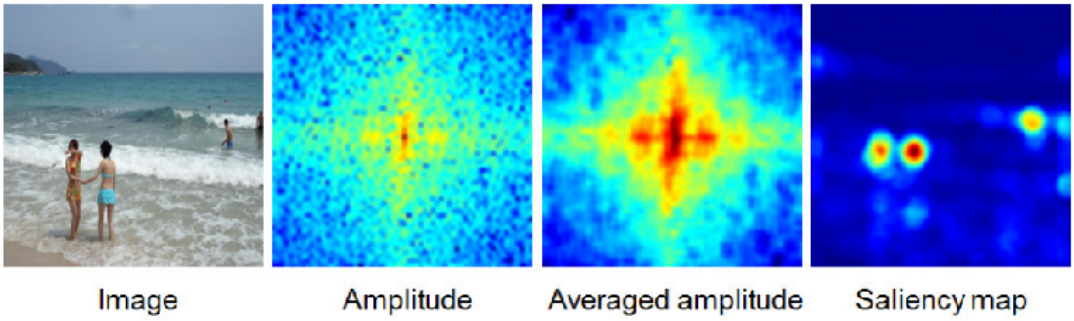

Figura 8: Mapa de saliencia producido polo algoritmo do residuo espectral.

Este algoritmo é subministrado co material desta práctica ou tamén pode utilizarse a versión de OpenCV.

- **Fase global (Guo et al, [2])**

  Inspirados na idea de Hou e Zhang, [1], Guo et al, [2] propuxeron detectar as irregularidades espazo-temporais mediante o *espectro de fases* da transformada de Fourier dun cuaternión. Observaron que ao reconstruír a imaxe con só esta información, as localizacións con menos periodicidade[3] aparecerán destacadas. Ademais, propoñen representar a imaxe de entrada de cor como unha imaxe cuaternión onde as catro compoñentes son: intensidade, oposición vermello-verde, oposición azul-amarelo e movemento (en caso de vídeo). Non obstante, podemos implementar esta idea sen empregar cuaternións se renunciamos a incluír nun todo a información de cor e movemento. Se collemos a mesma aproximación que no residuo espectral, obteremos facilmente a saliencia dunha imaxe de intensidade a partir da fase global a partir do espectro de Fourier. Para incluír a cor, podemos achar a saliencia para cada banda por separado e combinalas ao final mediante un promedio ou o máximo en cada punto.

- **Modelo bio-inspirado de Itti et al, [3]**:

  Un modelo clásico para detectar saliencia foi proposto en (Itti et al,[3]). Obviamente, non encaixa no grupo de modelos frecuencias pero se introduce neste traballo como modelo para comparar cos anteriores e ter unha comparativa razoable (dentro da limitación dun traballo de prácticas). Neste enfoque, a importancia visual dunha localización cuantificase como a diferenza dela coas localizacións veciñas en diversas características e escalas. O marco principal deste enfoque pódese resumir en tres módulos principais, incluíndo: 1) extracción de características pre-atentivas (intensidade, cor e orientación); 2) cálculo do contrastes das características citada mediante filtrado centro-envolvente multiescala; e 3) integración dos mapas contraste nun único mapa de saliencia (suma normalizada). Na Fig. 9 podemos ver un esquema do modelo e na Fig. 10 resultados do mapas de saliencia deste modelos sobre diferentes imaxes.

[3] que se separan da tendencia do resto

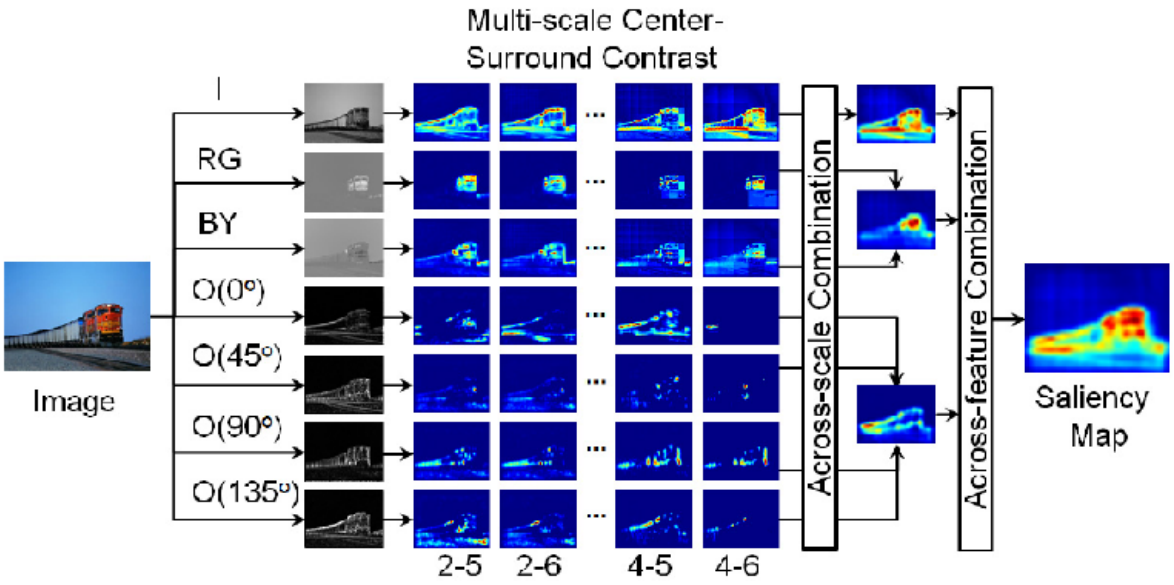

Figura 9: Esquema do modelo de Itti.

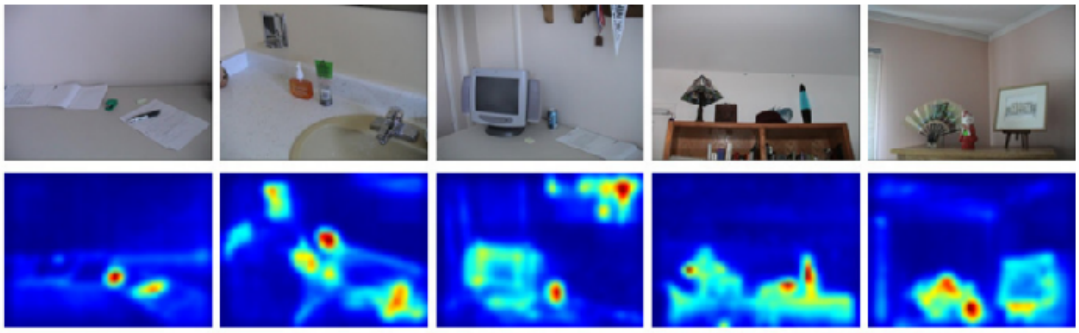

Figura 10: Mapas de saliencia producidos polo algoritmo de Itti.

Saliencia Dinámica estes algoritmos se centran en buscar irregularidades globais no espectro espazo-temporal dun vídeo. Adicionalmente, se pode fusionar nun único mapa a información saliente dinámica e estática con promedios o máximos en cada punto do vídeo. Este tipo de relevancias está estreitamente relacionado con unha variedade de aplicacións críticas na robótica como o seguimento de obxectos importantes e a identificación de escanas.

- **Discrepancia na fase (Zhou et al, [4])** Os autores, para extraer obxectos en movemento de fondos dinámicos, o seu modelo segue a idea de codificación preditiva (o movemento supón unha alteración na fase do vídeo). En primeiro lugar, predicimos o seguinte frame só considerando movementos do fondo. Logo, comparando a predición coa observación real, os píxeles que representan o primeiro plano emerxen debido á gran reconstrución erro. Os autores, demostran que cunha aproximación de MATLAB de 9 liñas son capaces de recuperar o movemento da cámara cun erro limitado. Ler a cita [4] para detalles.

Pasos comúns a todos os algoritmos de saliencia: Hai dúas cousas comúns importantes, a todos os algoritmos de saliencia, para obter o mapa final de saliencia: potenciación do contrates (elevamos ao cadrado directamente ou facemos a normalización iterativa e logo elevamos ao cadrado) e finalizamos cun suavizado Gaussiano (a anchura depende autor).

4. Bases de datos a utilizar

A maioría dos "benchmarks" existentes están baseados en imaxes debido a cantidade de datos dispoñibles neste formato fronte a os de vídeo e as dificultades engadidas que presenta este último tipo de datos: movemento da cámara, tipo de escenas, duración, etc. Con independencia de se a base de datos é de imaxes ou de vídeo, todas cumpren dúas premisas: unha selección de estímulos acorde a algún criterio (naturais, urbanos, sintéticos, etc.) e un mapa de fixacións para cada imaxe ou frame que nos indica o "ground-truth"[4] conformado dun conxuntos variado de humanos.

Debido á limitación dos dispositivos de seguimento ocular, só as coordenadas de cada fixación son gravadas. É dicir, cada fixación está ligada a un só píxel na pantalla e sen ningunha información de escala/rexión. Dado que as fixacións revelan a distribución dos subconxuntos visuais salientables, o mapa de relevancia "ground-truth" pódese aproximar mediante o mapa de densidade de fixacións que se obtén convolucionando o mapa de fixacións (cada fixación é un valor de 255 no píxel indicado) cunha gaussiana de desviación típica de un grao visual (campo visual da fóvea). Para obter a distancia que subtende un grao debemos coñecer a distancia da fóvea (observador) a pantalla de observación ($\sigma_G = \arctan(\theta/2) \cdot dist_{obs}$). Non obstante, o normal é que os autores da base de datos aporten os estímulos visuais (imaxes), os mapas de fixacións e os mapas de densidades (no seu defecto aportan os datos precisos para obter a sigma da gaussiana para obter os mapas de densidade).

Nesta práctica, imos a empregar a base de datos de Bruce e Tsotsos[5]. Estes autores propuxeron un banco de probas con 120 imaxes de cor con escenarios diversos tanto de interiores como de exteriores e cunha resolución de 681×511 (ver Fig. 11).

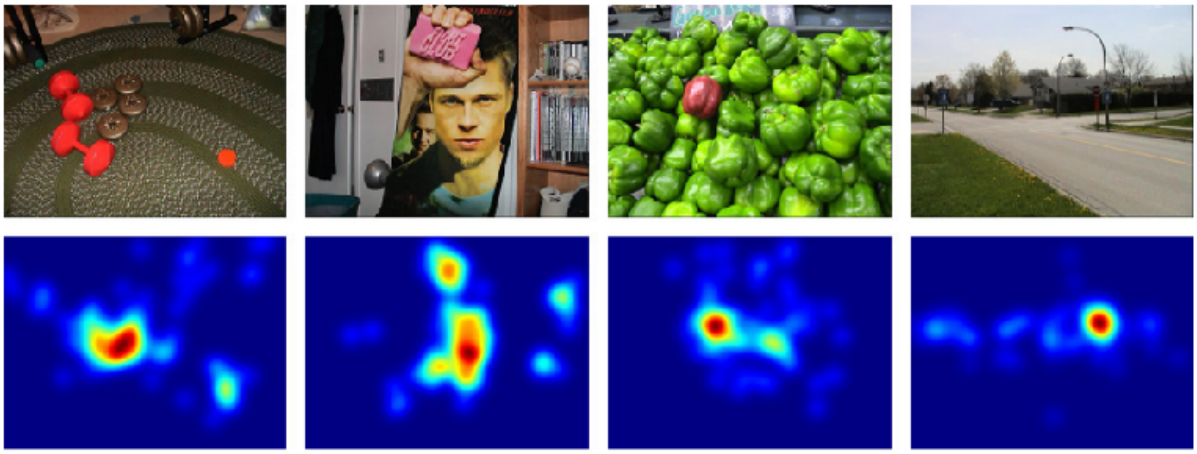

Figura 11: Imaxes representativas da base de datos de Bruce e Tsotsos[5] xunto cos seus mapas de densidade de fixacións.

Algunhas imaxes conteñen obxectos obvios, en primeiro plano, mentres que outras non presentan rexións de interese especial. Durante o experimento con humanos, este conxunto de 120 imaxes mostráronse aleatoriamente a un conxunto de 20 suxeitos (con distintas idades, visión normal ou corrixidas por lentes). En cada quenda, o observador estaba situado a $dist_{obs} = 0{,}75\,\mathrm{m}$ dun monitor de 21 polgadas para visualizar a imaxe test durante $4\,\mathrm{s}$, en condicións de observación libre (sen instrucións concretas). Esta

[4] indica o mapa verdadeiro de saliencia co cal comparase e se obtivo a partir das fixacións dos humanos

base de datos, converteuse nunha das máis empregadas para avaliar algoritmos de saliencia en artigos de investigación neste campo.

Para vídeo tamén existen bases de datos que podedes baixar libremente. Por exemplo, na USC elaboramos a CITIUS Video Database que inclúe 72 vídeos descargados de Internet e algúns vídeos sintéticos xerados no laboratorio. Os vídeos pódense clasificar en catro categorías, naturais e sintéticos, con cámara fixa ou de movemento. Inclúe 27 vídeos sintéticos con efectos emerxentes dinámicos. Dende esta URL pódense descargar os estímulos, as fixacións e as utilidades software para manexar os datos experimentais. Ao final da mesma, hai referencias a outras bases de datos existentes en Internet.

5. Medidas de avaliación da saliencia

Das discusións anteriores, podemos ver que xa existen moitas bases de datos de imaxes e vídeo para a avaliación da relevancia dos algoritmos de saliencia. Así outro problema é, que tipo de metodoloxías de avaliación debemos empregar para comparar cuantitativamente os mapas de fixacións dos humanos e os producidos polo algoritmos? Nesta subsección, presentaremos as tres métricas de avaliación comunmente utilizadas.

- **Receiver Operating Characteristic (ROC)** é a métrica comparativa máis popular para medir o rendemento dos modelos de relevancia visual. No proceso de avaliación, a ROC considera o mapa de saliencia do modelo como un clasificador binario e avalía o seu rendemento baixo diferentes criterios (limares). Supón que os mapas de relevancia previstos están normalizados ao rango dinámico de $[0, 255]$, a avaliación será canalizada usando todos os limiares probables entre $\{0, 1, ..., 255\}$. En cada limiar, os mapas de relevancia se binarizarán en dúas rexións: primeiro plano (PP) e fondo (BK) e a partir delas, acharemos o número de verdadeiros positivos ($\#TP$) e verdadeiros negativos ($\#TN$):

  $$TP = \#(PP \,\&\, \text{Fixación}), \qquad FP = \#(BK \,\&\, \text{Fixación})$$

  Como se mostra na Fig. 12, pódese xerar unha curva ROC usando todos os pares (TP, FP) obtidos para todos os limiares fixados. Podemos ver que a curva ROC pode revelar o rendemento dun modelo de saliencia en diferentes condicións. Por exemplo, un pequeno limiar deixará pasar máis rexión fixadas pero tamén permitirá pasar mais zonas que non foron fixadas polo humano facendo que o rendemento do clasificador binario diminúa. Para medir o rendemento global, emprégase a área baixo a curva ROC, denotado como **AUC**. Un modelo de saliencia perfecto (plena coincidencia do mapa de saliencia e as fixacións do humano) corresponde a un AUC de 1.0, mentres que un modelo aleatorio terá un AUC de 0.5. Revisa o código aportado para ver como se implementa esta medida a partir do mapa de saliencia do modelo computacional e as fixación dos humanos para cada imaxe da base de datos.

  Cando usamos a ROC e AUC, normalmente hai dúas formas de avaliar o rendemento global para o conxunto de imaxes da base de datos: 1) calcular o valor da AUC en cada unha imaxe e despois calcular a media e a desviación estándar de todos os valores da AUC; e 2) sumar o número de verdadeiros positivos e falsos positivos en todas as imaxes do benchmark e xerar unha curva ROC única para achar unha única AUC. Os dúas metodoloxías teñen sentido pero, segundo os últimos artigos no campo, a primeira forma é a preferible.

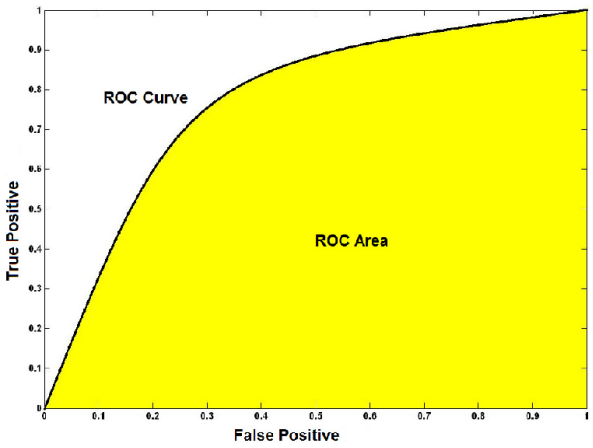

Figura 12: Exemplo de curva ROC.

- **Normalized Scanpath Saliency (NSS)** é outra forma de medir o rendemento dos modelos de saliencia mediante fixacións dos humanos. A idea básica é normalizar os valores de saliencia de todas as rexións dunha imaxe para ter media cero e desviación estándar unitaria. A continuación, mediremos os valores de saliencia sobre o mapa do modelo computacional nos lugares fixados polos humanos[6]. Vemos que a NSS é unha especie *z-score*, que cuantifica a discrepancia dos resultados experimentais con respecto da media expresándoo como número de desviacións estándar que se separa da media. Canto maior sexa o valor de NSS, menos probable será o resultado experimental se deba ao azar e máis se parece ao humano. A NSS pódese calcular como:

$$\mu = \frac{1}{N}\sum_{n=1}^{N} s_n,$$

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(s_n - \mu)^2},$$

$$NSS = \sum_{f=1}^{N_{fix}} \frac{s_e - \mu}{\sigma}.$$

  onde s_n e s_e son os valores no mapa de saliencia do modelos na n-enésima posición e os valor do mapa de saliencia do modelo na posición onde hai unha fixación do humano, respectivamente. N é o número total de localizacións candidatas na escena. Se $NSS = 1$ significa que as fixacións dos suxeitos caen nunha rexión cuxa relevancia prevista é unha desviación estándar por encima da media.

- **Linear correlation coefficient (CC)** é unha medida que compara directamente o mapa de saliencia do modelo computacional (S) e o mapa de relevancia da verdadeiro (G) que ven dado pola función de densidade de fixacións dos humanos. Tal relación pódese medir como:

$$CC = \frac{cov(S,G)}{std(S)\cdot std(G)}$$

Por que empregar varias medidas? Cando avaliamos modelos de saliencia computacional empréganse varias medidas (nós utilizaremos 3). A razón e non existe ningunha que recolla un comportamento fiable ao $100\,\%$, polo tanto, os modelos ganadores serán aqueles, que en media, exhiban un mellor comportamento no número máis elevado de medidas empregado.

6. Tarefas

As tarefas que teñen que realizar os estudantes van a ser divididas en dous bloques: implementación e avaliación de modelos de saliencia estática, por un lado, e polo outro para saliencia dinámica. No primeiro caso, empregaremos a base de imaxes de Bruce e Tsotsos[5] e para vídeo, empregaremos uns 10-15 vídeos da base CITIUS Video Database que o estudante libremente elixa. As tarefas concretas son:

Tarefa 1 : Saliencia estática

As tarefas que os estudantes teñen que levar a cabo esta parte son:

(a) Lete o artigo de Hou e Zhang [1] e o código do residuo espectral. Con isto na cabeza, idea un método para incluír a información de cor (espazo de cor é unha decisión relevante) no algoritmo e unha posible forma de obter saliencia invariante a escala (o mellor que logres facer!).

(b) Inspirado na forma de implementar o residuo espectral (sen empregar cuaternións), programa un algoritmo que obteña a saliencia empregando a fase global (Guo et al [2]). Completa este algoritmo inicial para incluír a información de cor e a invariancia a escala.

(c) Lete o artigo de Itti et al, [3] e o código aportado para saber como obter os mapas de saliencia deste modelo sobre unha imaxe de entrada.

(d) Lete o artigo aportado de [7] para ter unha idea clara de como avaliar o rendemento dos modelos de saliencia sobre unha base de datos (nós empregaremos [5]). Aquí os modelos que avaliaremos será o residuo espectral (punto 1), fase global (punto 2) e o modelo de Itti et al, [3]. Para avaliar este modelos, empregaremos as tres medidas estudadas neste documento: AUC, NSS e CC. Analiza e comenta os resultados (presentaos correctamente tabulados).

(e) Visualiza os 3 mellores casos (onde a saliencia do modelo e o humano teñen a mellor coincidencia) e os tres peores para cada modelo e cada métrica. Que características teñen en común?

(f) Se constrúo unha gaussiana situada no centro cunha dimensión iguais ás das imaxes da base de datos e cun sigma amplo e a emprego como se fose un mapa de saliencia dun modelo, que resultados obteño coas medidas aportadas sobre a base de datos?. Que efectos observas? Cal cres que pode ser a razón? Que medida consideras que ten un mellor rendemento e porque?

(g) Se collo o mapa de todas as fixación da base de datos e o convoluciono cunha gaussiana cunha sigma que subtenda un grao visual (sei a distancia do suxeito á pantalla de observación $dist_{obs} = 0{,}75\,\mathrm{m}$ a unha pantalla de 21 polgadas) e o emprego como mapa de saliencia dun superhumano. Que resultados obteño coas medidas? É razoable o resultado obtido? Para que cres que pode servir esta información?

Material aportado: código do algoritmo do residuo espectral, base de datos de imaxes coas fixacións e mapas de densidade, así como un script para ler as fixacións, e superpoñelas sobre a imaxe orixinal a efectos de visualización. Por último, apórtase o código das medidas de avaliación necesarias e un exemplo de como utilizalas cos mapas de saliencia do algoritmos do residuo espectral. Se escrutas a estrutura de directorio, atoparas os artigos nos que se basea cada algoritmo para que os leas e te enteres dos pormenores.

Tarefa 2 : Saliencia dinámica

As tarefas que os estudantes teñen que levar a cabo esta parte son:

(a) Xeneraliza o algoritmo de saliencia baseado na fase global (punto 2 Tarefa 1) para un vídeo. Pensa unha maneira de incluír cor e movemento sen empregar cuaternións.

(b) Implementa o algoritmo de discrepancia na fase (Zhou et al, [4]). Pensa unha forma de incluír a información de cor.

(c) Para a base de vídeos CITIUS Video Database (ou outra que ti desexes), elixe un conxunto de 10 ou 15 vídeos representativos. Baseado nas utilidades que aportan os autores, implementa en Python o script que lea e visualice as fixacións sobre os vídeos seleccionados.

(d) Para os dous modelos de saliencia dinámica implementados e sobre a selección de vídeos que realizaches no punto anterior, avalía o rendemento coas tres medidas aportadas. Dado que as métricas están pensadas para aplicarse sobre unha imaxe (ou frame no caso dun vídeo), pensa unha forma de xeneralizadas ao noso problema de vídeo. Explica a estratexia que adoptes coas medidas e analiza a comparativas entres os dous modelos, a gaussina centrada na imaxe e o mapa do superhumano.

(e) Se collo os vídeos seleccionados e visualizo o mapa de densidades de fixacións dos humanos para os primeiros frames (5-10). Que observo? Elixe dous casos onde atopes o fenómeno máis acentuado e aporta hipóteses sobre a súa orixe?

7. Entrega

Puntos que debe cumprir a entrega do documento que subas ao Campus Virtual:

1. A práctica debe ser autocontida en dous cadernos de Jupyter: un para saliencia estática e outro para a dinámica.
2. Tabulación dos datos en formato entendible por un humano.
3. Explicación das túas achegas e dos análises dos resultados obtidos.
4. A base de datos de vídeo (so os seleccionados por ti) e utilidades software que desenvolviches para ler e visualizar as fixación en cada frame do vídeo.

8. Rúbrica da práctica

- Implementación dos modelos 2D e 3D $\rightarrow$ **25 pts**
- Adaptación das medidas a vídeo e ferramentas para ler/visualizar as fixación da base de datos indicada $\rightarrow$ **25 pts**
- Análise e presentación dos resultados, comentarios pedidos e acerto na explicación dos fenómenos requiridos $\rightarrow$ **50 pts**

Referencias

[1] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[2] a. M. Q. Guo, C. and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," *Preceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.

[3] L. Itti, C. Koch, and E.Ñiebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[4] B. Zhou, X. Hou, and L. Zhang, "A phase discrepancy analysis of object motion," in *Computer Vision – ACCV 2010* (R. Kimmel, R. Klette, and A. Sugimoto, eds.), (Berlin, Heidelberg), pp. 225–238, Springer Berlin Heidelberg, 2011.

[5] N. Bruce and J. Tsotsos, "Saliency based on information maximization," *Neural Information Processing Systems (NIPS)*, pp. 155–162, 2005.

[6] R. Peters and L. Itti, "Applying computational tools to predict gaze direction in interactive visual environments," *ACM Transactions on Applied Perception*, vol. 5, no. 2, pp. 1–19, 2008.

[7] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2019.