Universidade de Santiago de Compostela
Departamento de Electrónica e Computación

PhD THESIS

MODELING EARLY VISUAL CODING AND SALIENCY THROUGH ADAPTIVE WHITENING: PLAUSIBILITY, ASSESSMENT AND APPLICATIONS

Presentada por:
Antón García Díaz

Dirixida por:
Xosé Ramón Fernández Vidal
Xosé Manuel Pardo López

Santiago de Compostela, Decembro de 2010

Dr. **Xosé Ramón Fernández Vidal**,
Profesor Titular de Universidade da
Área de Física Aplicada da
Universidade de Santiago de
Compostela

Dr. **Xosé Manuel Pardo López**,
Profesor Titular de Universidade da
Área de Linguaxes e Sistemas
Informáticos da Universidade de
Santiago de Compostela

FAN CONSTAR:

Que a memoria titulada **Modeling early visual coding and saliency through adaptive whitening: plausibility, assessment and applications** foi realizada por D. **Antón García Díaz** baixo a nosa dirección no Departamento de Electrónica e Computación da Universidade de Santiago de Compostela, e constitúe a Tese que presenta para optar ao grao de Doutor.

Santiago de Compostela, Decembro de 2010

Asdo: **Xosé Ramón Fernández Vidal**
Codirector da tese de doutoramento

Asdo: **Xosé Manuel Pardo López**
Codirector da tese de doutoramento

Asdo: **Francisco Fernández Rivera**
Director do Departamento de
Electrónica e Computación

Asdo: **Antón García Díaz**
Autor da Tese

Na lembranza de meu pai,
artífice da curiosidade

Agradecementos

Esta tese teria sido imposible sen Xosé Ramón Fernández Vidal e Xosé Manuel Pardo López, os magníficos e xenerosos directores dos que tanto aprendín nestes anos. Foi certamente unha grande sorte poder contar en todo momento coa súa implicación e axuda, sen que os quilómetros e os horarios que mediaron se convertisen nunca nun impedimento.

Tamén a paciencia e a axuda –en tempo real– de Raquel Dosil en moitos aspectos técnicos e conceptuais e na loita co meu inglés!, tiveron un papel moi importante para poder realizar a miña laboura durante este tempo.

A Víctor Leborán, a Pilar Garcia, a David Garcia, moi especialmente por todas as xornadas e manteis compartidos, pero tamén aos demais compañeiros cos que coincidín no grupo de visión artificial da USC, teño que agradecerlles un excelente clima de traballo, así como a súa axuda, tan útil, en moitas pequenas cousas. A Fernando López quero agradecerlle a súa colaboración indispensable no traballo de recoñecemento de esceas, así como a David Luna a súa participación nos experimentos. Aos compañeiros de coche a Compostela, grazas por facer tan divertidas e acolledoras tantas horas de ir e vir.

Manter o optimismo e o ánimo despois desas case oito horas que se adican a outra cousa é esencial para aproveitar as poucas horas que aínda restan. Grazas aos compañeiros de traballo en Ferrol, especialmente a Maria Jesús Rodríguez e a Regino Varela por un clima de traballo ben máis valioso do que a fermosa vista do CIS sobre a ria, e polos seus consellos que enseguida fixeron dunha bisbarra allea a miña casa. A Isabel Garcia quero agradecerlle o seu inestimable apoio para tantas cousas. Foi un privilexio no profesional e no persoal traballar con Marisol Torres e compartir as cousas boas que xorden nos momentos complicados. A Higinio González grazas polas oportunidades brindadas.

Sen amigos e familia non se vive, e ben tiven ocasión de comprobalo durante estes anos nos que o tempo non queria ser libre nunca, e porén os ánimos e apoios empuxaron tanto cara adiante. O mellor de escribir isto é que parece anunciar as areladas ocasións para repoñer as ausencias repetidas. Ao Brais e á Maria en primeiro lugar mil grazas por estaren sempre aí, liberando tempos imposibles e abrindo as rutas máis inesperadas e marabillosas. A Xurxo e a Carmen un exemplo en tantas cousas e uns superavós, grazas por iso e moito máis como as abondosas

e catárticas risas.

Á miña tia Maria como agradecerlle a súa incondicional presenza nos momentos máis importantes e necesarios? e esa habilidade súa para soster todo un mundo do que tanto me enorgullece facer parte.

Aos amigos, bos e xenerosos, teño que lles agradecer as findes, as conversas, os imprescindibles bos momentos. Aos amigos alaricanos grazas tamén pola cálida acollida, estes últimos dous anos, facendo de Allariz unha vila aínda máis fermosa do que xa é.

A Caamaño, meu pai, como se che bota en falta!, e á Marisa, a miña mai, para quen esta tese supuxo inmerecidas soidades, que non lles vou agradecer?. Grazas polas trabes do mundo, que me gardan de que o ceo rompa e caia en mil anacos sobre a miña cabeza. E grazas por tanta causa desa incurable saudade que sempre me acompaña. Ao meu irmán Roi, que se supera como tío, grazas por todo.

E o máis importante, aínda eu non merecéndoo, é a arela de que a Uxia, como tamén duns meses a esta parte a nosa filla Xoaniña, sigan por sempre iluminando a meniña dos meus ollos, poñendo a cor das cousas todas e traéndome ao maxín unha e outra vez a mesma idea: avante toda.

Decembro de 2010

[...] fiquei pasmado... Aquel ollo de vidro que de nada me servira na vida sírveme agora pra mirar. Tolo de contento quitei o ollo, dinlle catro bicos e volvino a pór no seu sitio.

Un ollo de vidro. Memorias dun esquelete.
Castelao

Contents

List of Figures

List of Tables

Abbreviations

- a. [opponent color component in a Lab color model]
- AIM. Attention from Information Maximization
- AUC. Area Under the Curve
- AWS. Adaptive Whitening Saliency
- b. [opponent color component ina Lab color model]
- BOLD. Blood Oxygenation Level Dependent
- BY. Blue-Yellow [opponent color component]
- DoG. Difference of Gaussians
- DoOG. Difference of Oriented Gaussians
- EEG. Electroencephalography
- FFT. Fast Fourier Transform
- FIT. Feature Integration Theory
- fMRI. functional Magnetic Resonance Imaging
- GBVS. Graph Based Visual Saliency
- HMAX. Hierarchical Model And X
- HVS. Human Visual System
- ICA. Independent Components Analysis
- IOR. Inhibition Of Return
- IR. Infrared

- L. Long [wavelength cone response] or Luminance [component in a Lab color model], depending on the context.
- Lab. (L,a,b) color representation.
- LGN. Lateral Geniculate Nucleus
- LMS. (L,M,S) [color representation]
- log Gabor. logarithmic Gabor [filter]
- M. Medium [wavelength cone response]
- NIMBLE. Natural Input Memory with Bayesian Likelihood Estimation
- NN. Neural Network
- n-NN. n Nearest Neighbors [rule or classifier]
- Op-art. Optical art.
- PET. Positron Emission Tomography
- PCA. Principal Components Analysis
- RG. Red-Green [opponent color component]
- RGB. (Red,Green,Blue) [color representation]
- ROC. Receiver Operating Characteristic
- ROI. Region Of Interest
- S. Short [wavelength cone response]
- SIFT. Scale Invariant Features Transform
- SUN. Saliency Using Natural [image statistics]
- SURF. Speeded-Up Robust Features
- TMS. Transcranial Magnetic Stimulation
- V1. Prymary Visual Cortex
- VOCUS. Visual Object detection with CompUtational [attention system]

- VSD. Voltage Sensitive Dye
- VSOTF. Visual Strehl ratio based on the Optical Transfer Function.
- WTA. Winner Take All

Resumo

A visión biolóxica esteblece un amplo abano de metas non acadadas por ningún sistema artificial en termos de eficiencia, robusteza e en xeral de funcionamento en tarefas visuais activas. A pesar da complexidade e variabilidade das imaxes naturais, os sistemas visuais dos mamíferos son sorprendentemente capaces de recoñecer obxectos e contextos nunha primeira fitada e dirixir de forma eficiente unhas poucas fixacións cara as partes máis salientes dunha escea descoñecida.

Estas capacidades requiren dunha selección da información dramática e realizada de forma activa, fornecendo unha causa de primeira orde para a atención visual. Isto semella razoable se considerarmos o enorme fluxo de información que recibe o sistema visual humano (SVH) a través dos fotoreceptores retinianos, estimada en máis de $10^{10} bits/s$ [AvEO05]. O procesamento adaptativo ascendente (*bottom-up*) e a percepción da saliencia considérase que residen na base deste comportamento temperán con tan notable eficiencia. Estes mecanismos parecen xogar un rol esencial no control da atención visual humana –en cooperación co control descendente (*top-down*) – como mostran multitude de resultados procedentes dunha ampla variadade de experimentos.

O termo saliencia visual é usualmente empregado para referir medidas que pretenden cuantificar o carácter conspicuo e distintivo dun estímulo visual. Isto é, tentan cuantificar canto sobresae un estímulo do contexto, a partir das súas propiedades físicas. A representación máis común da saliencia adoita ser mediante un mapa retinotópico (o mapa de saliencia). Unha fonte de información principal –aínda que de ningún xeito a única– para entender o funcionamento da atención visual é a distribución espacial de fixacións oculares obtidas en experimentos de seguemento ocular. Os movementos oculares producen fixacións que determinan as pequenas rexións dunha imaxe dada que son proxectadas sobre a fóvea. En condicións de boa iluminación (de visión fotópica), estas pequenas rexións reciben unha resolución espacial moito maior debido á densidade moito maior de fotoreceptores presentes na fóvea. Por este motivo, os movementos oculares representan unha

Figura 1: Exemplos dun mapa de saliencia (centro-superior) e un mapa de densidade de fixacións (arriba dereita) para unha imaxe típica (esquerda). Debaixo de cada mapa amósase o resultado de o superpoñer sobre a imaxe orixinal.

primeira forma de forte selección espacial da información visual. Na figura 1 amósase un exemplo de mapa de saliencia así como o correspondente mapa de densidade de fixacións oculares, para unha imaxe típica. Dá unha idea das implicacións reais da selección espacial dirixida pola saliencia. Porén, convén sinalar que tamén a visión periférica se ve afectada pola selección debida á atención, sen necesidade de que medie un movemento dos ollos. Esta cuestión será retomada de novo ao longo desta tese.

Doutra banda, os modelos mecanísticos de procesamento visual temperán céntranse na explicación dos campos receptivos visuais e no seu comportamento adaptativo, tanto fronte a características locais como globais. Un obxectivo primordial destes modelos é a formulación de estratexias de codificación temperá que sexan bioloxicamente plausibles e que sexan capaces de explicar determinados fenómenos relacionados coa visión temperá, e en particular coa adaptación contextual do comportamento neuronal e perceptual.

O problema de medir a saliencia ou distintividade nunha imaxe ten tamén unha grande relevancia na visión por computadora e en xeral no desenvolvemento de sistemas de visión, moi especialmente de sistemas activos. De feito, a atención espacial emerxente ten demotrado ser moi útil en importantes funcións visuais como a aprendizaxe e o recoñecemento e moitas aplicacións de visión como se amosa no primeiro capítulo desta tese. Asemade, a extracción de características de baixo nivel adecuadas ten unha enorme importancia na análise de imaxes e na visión por computadora. Ambos os dous problemas –representación de baixo nivel e saliencia– adoitan aparacer estreitamente ligados en moi variadas solucións. Un exemplo destacado atópase nos detectores de puntos de interese máis extendidos, mais tamén en moitos outros modelos de visión por computadora.

Ambas as preocupacións tanto en torno á comprensión do SVH como

ao desenvolvemento de sistemas de visión activos, promoveu e promove un importante esforzo interdiciplinar para proporcionar medidas melloradas de saliencia. En particular, nos últimos anos asistimos a un extraordinario e crecente esforzo investigador no modelado bioinspirado da saliencia e as súas aplicacións.

Porén, hai claramente unha carencia de modelos que aborden a relación entre a adaptación contextual dirixida polos datos observada na codificación visual temperá e a percepción da saliencia. Comprender esta relación é esencial para o desenvolvemento dun cadro computacional para a codificación visual temperá que sexa bioloxicamente plausible. Semellante cadro debe formular representacións retinotópicas intermedias plausibles e adaptadas á imaxe. Estas representacións intermedias deben ser capaces de termar dunha medida adecuada de saliencia, mais tamén de axustarse a características observadas propias da visión temperá. As aproximacións a este problema son tamén moi interesantes para a visión por computadora, na medida en que poder brindar modelos mellorados tanto de características de baixo nivel como de saliencia.

Ademais, a maioria de modelos de saliencia están fundamentados en principios da teoría da información, sen unha especificación das fontes físicas involucradas e, aínda máis importante, dos diferentes xeitos no que estas contribúen á saliencia visual. Esta especificación, de ser posible, é moi importante xa que ofreceria unha ligadura adicional para comprender a función visual en termos das súas raíces físicas. Así mesmo poderia proporcionar excelentes pistas para o desenvolvemento de aproximacións de visión activa e en xeral para o procesamento e análise adaptativos de imaxes.

Coa intención de cubrir estes ocos, esta tese proporciona unha aproximación funcional e coherente tanto á codificación visual temperá como á saliencia, dun xeito bioloxicamente plausible. Así mesmo, o cadro proposto enraiza nunha interpretación física que involucra unhas poucas magnitudes ópticas. Demóstrase como o modelo resultante explica unha serie de ilusións visuais e supera claramente os modelos de saliencia do estado da arte existentes usando as probas de avaliación máis extendidas, incluíndo a predición de fixacións oculares e a reprodución de resultados psicofísicos.

A primeira das carencias descritas enriba pode ser doadamente aprezada nas dúas estratexias habituais de representación de baixo nivel adoptadas polos modelos de saliencia existentes. Moitos deles comezan cunha descomposición multiresolución de tres compoñentes de cor predefinidas, nun determinado modelo de cor. Isto fano proxectando as compoñentes da imaxe de cor sobre filtros lineares semellantes a campos receptivos de células do córtex visual primario (V1), que adoitan ser modelados mediantes bancos de filtros Gabor ou gaussianos desde que o modelo estándar de V1 foi inicialmente

proposto por Hubel e Wiesel [HW59, HW68]. Os seguintes pasos xeralmente involucran procesos de competición e integración que rematan nunha medida final de saliencia, un esquema presente xa nos primeiros modelos baseados na arquitectura de atención de Koch e Ullman [KU85]. A outra aproximación típica aborda a descomposición mediante a proxección da imaxe sobre compoñentes independentes de parches de imaxes naturais, eludindo deste xeito a parametrización das canles de cor e dos filtros e alén do tamaño de parche. Esta proposta está baseada na interpretación estatística do modelo estándar de V1 como resultado da evolución e o desenvolvemento neuronal para axustarse ás estatísticas que caracterizan as imaxes naturais [O$^+$96, BS97].

Ambos os dous esquemas brevemente descritos, tanto o baseado en bancos de filtros como na análise de compoñentes independentes de imaxes naturais, comparten unha importante propiedade: sempre empregan as mesmas porcións do espazo de características para representar unha imaxe calquera. As aproximacións baseadas en bancos de filtros proxectan un conxunto fixo de compoñentes de cor sobre unha partición fixa do dominio espectral. Asemade, as compoñentes independentes son determinadas a partir dun conxunto de adestramento de imaxes naturais e non son modificadas posteriormente.

Estas aproximacións á codificación temperá adoptadas pola inmensa maioria dos modelos de saliencia son estáticas e non se axustan por tanto ao comportamento do SVH. De feito, o SVH adapta as súas respostas ás características globais e locais de cada imaxe específica. Exhibe unha clara adaptación de curto prazo e contextual ao contraste, á cor e á estrutura espacial. Esta adaptación ten lugar desde os fotoreceptores e as células G ata ás células corticais e tense observado que produce en conxunto unha representación decorrelacionada [BF89, RR09, Koh07, CWS$^+$07, SHD07]. Deste xeito a decorrelación adaptativa semella ser un mecanismo neuronal plausible. Non debe sorprender pois, que moitos modelos mecanísticos recentes de redes neuronais corticais así como modelos de computación mediante poboacións de neuronas producen en conxunto unha representación decorrelacionada e branqueada da entrada.

Desde unha perspectiva computacional, existen tamén razóns a favor dun modelo de adaptación contextual. Aproximacións que non presenten tal adaptación son máis susceptibles de sofrer de caracterizacións nesgadas, limitando a aplicabilidade da correspondente medida de saliencia.

Así, o problema da saliencia parece estar estreitamente relacionado co problema da selección dunha representación de baixo nivel do mesmo xeito que a súa adaptación. No contexto da visión biolóxica, a codificación visual temperá preséntase como un problema ineludible se se pretende defender a plausibilidade biolóxica dun modelo. Por outra banda, unha perspectiva adecuada sobre a codificación visual temperá pode fornecer directrices no

deseño de representacións de baixo nivel de imaxes, que sexan adecuadas para funcións visuais activas que poidan ser de utilidade en aplicacións de visión por computadora e de sistemas de visión. De feito, e de xeito similar a outros traballos no campo, a motivación orixinal desta tese naceu froito dun proxecto a longo prazo de desenvolvemento dun cadro xenérico e bioloxicamente inspirado para aproximar e estudar problemas de visión activa.

Hipóteses e obxectivos

A hipótese de traballo asumida nesta tese é que a adaptación contextual que xorde do branqueado adaptativo é o factor clave involucrado na determinación da saliencia visual. Así, do mesmo xeito que hai unha adaptación a longo prazo da codificación neuronal dirixida polas estatísticas das *imaxes naturais*, hai tamén unha adaptación contextual a curto prazo da codificación temperá dirixida polas estatísticas de cada imaxe particular. Unha asunción implícita nesta hipótese é que os principais mecanismos computacionais subxacentes á adaptación contextual son a decorrelación das respostas e a normalización ao contraste.

Repousando nestas ideas, esta tese céntrase na investigación, en termos de magnitudes sinxelas, tanto da adaptación contextual da representación de baixo nivel como dunha definición coherente da saliencia visual.

Polo tanto, podemos dicir que se procuran tres obxectivos principais, a saber:

- A proposta dun cadro teórico capaz de explicar dun xeito coherente unha variedade de fenómenos relacionados coa adaptación contextual e a saliencia visual. Este cadro debe ser bioloxicamente plausible, e daquela debe cumprir cunha serie de limitacións impostas polo comportamento coñecido do SVH.

- A implementación dun modelo de saliencia computacional que supere aproximacións previas, en termos de reprodución de resultados en experimentos visuais con observadores humanos. Estes experimentos deben incluír exemplos representativos daqueles adicados á atención visual, tanto dos que involucran movementos oculares como dos que non. Entre eles, unha referencia primordial é a predición de fixacións humanas en observación libre de imaxes usando bases de datos de seguemento ocular de libre acceso.

- A demostración da utilidade do modelo de saliencia nunha variedade de aplicacións, posto que a aplicabilidade pode tamén ser vista como unha condición de validez para calquera novo modelo. Dada a grande

cantidade de aplicacións xa existentes, fíxase unha selección de tres aplicacións obxectivo. A primeira é amosar a utilidade da medida de saliencia proposta como a base dunha segmentación obxectos-fondo xenérica, un problema de primeira orde na análise de imaxes. O segundo obxectivo consiste na mellora da selección de puntos de referencia e de interese, unha cuestión central nos problemas de navegación de robots. O terceiro obxectivo pretende ampliar o campo de aplicacións da saliencia a un ámbito novo, á análise de representacións espaciais non visuais, posto que a aproximación aquí proposta fundaméntase nunha aproximación física –xeralizable– á visión temperá.

Contribucións desta tese

A seguir, resúmense as principais contribucións desta tese

- Acádase unha nova perspectiva sobre a adaptación contextual e a condificación temperá –dirixida polos datos– no SVH, a través dun sinxelo cadro de branqueado progresivo de compoñentes de cor e de escala. Xa que logo, proponse unha representación adaptada ás estatísticas da imaxe nunha forma computacional sinxela capaz de explicar unha variedade de ilusións visuais.

- Unha definición da correspondente medida de saliencia derívase como o módulo na representación branqueada obtida, que se propón como estimación dun invariante no SVH. O modelo resultante é así nomeado como saliencia por branqueado adaptativo (AWS nas siglas en inglés). Así mesmo, esta medida de saliencia demóstrase como está directamente relacionada cunha definición equiparable de *variabilidade óptica* en termos de lonxitudes de onda espectrais e de frecuencias espaciais a partir da descrición típica dunha imaxe en óptica de Fourier. Asemade, esta ligazón fornece un obxectivo explicativo para a catástrofe de codificación dentro da hipótese de codificación eficiente, en termos de invarianza do SVH na representación da variabilidade óptica presente na imaxe, dentro dunha *ventá óptica visual* consecuentemente definida.

- Proponse o emprego da capacidade preditiva de fixacións humanas amosada polos propios humanos como referencia para mellorar unha extendida medida de avaliación baseada na análise ROC. Deste xeito, a valoración do funcionamento do modelo fronte á variabilidade entre esceas é mellorada, e obtense unha valiosa información adicional sobre a robusteza de modelos ou a fortaleza da saliencia. O modelo AWS exhibe

un funcionamento equivalente aos humanos, superando claramente outros modelos do estado da arte que semellan sufrir de diferentes nesgos de deseño que limitan a súa xeralidade. Por outra parte, o modelo AWS demóstrase capaz de reproducir un conxunto representativo de resultados psicofísicos, que até onde nós coñecemos non foron conxuntamente reproducidos por ningún outro modelo anterior.

- Demóstrase a aplicabilidade do modelo AWS en problemas de visión por computadora e sistemas de visión. En particular a aplicación directa do AWS a imaxes multi e hiperespectrais, até onde sabemos non proposta anteriormente con ningún outro modelo bioinspirado de saliencia. É importante tamén a demostración de resultados equiparables proporcionados polo modelo empregando tanto unha representación comprimida extraída a partir de moitos sensores espectrais de anchura de banda estreita, como unha representación tricromática clásica de detectores de banda larga. Asemade, realízase unha proposta de medida de avaliación para comprobar a correcta proxección da variabilidade física en técnicas de fusión de sensores para a visualización espacial, baseada no emprego do modelo AWS tanto sobre os datos orixinais como sobre os datos visualizados.

Esquema seguido nesta tese

Esta tese organízase como segue.

- No capítulo 1, o concepto de saliencia, o seu rol e funcionamento na visión humana, os modelos computacionais existentes e os seus campos de aplicación son revisados con certo detalle.

- O capítulo 2 adícase á investigación da codificación visual temperá. Proponse un sinxelo cadro funcional que posibilita a adaptación contextual. A adaptación lógrase mediante o branqueado adaptativo de características de cor e escala. O branqueado implica a adaptación do rango dinámico da correspondente dimensión de características aos datos específicos observados nunha escea dada. Aplícase primeiramente ás compoñentes de cor seguido dunha descomposición multiescala das compoñentes de cor branqueadas. Isto faise para varias orientacións. A seguir, as características de escala resultantes son branqueadas tamén. Tal representación é analisada á luz de varios fenómenos psicofísicos relacionados con ilusións visuais usando imaxes sintéticas, representacións artísticas e imaxes naturais. Aliás, a plausibilidade biolóxica avalíase á luz de propiedades coñecidas do SVH.

- O capítulo 3 investiga as ligazóns da proposta de branqueado adaptativo cunha simple descrición óptica de imaxes. Como resultado, a aproximación de branqueado adaptativo demóstrase que está directamente relacionada cunha simples definición de *variabilidade óptica* en función de lonxitudes de onda espectrais e de frecuencias espaciais, cando se computa nos límites sensoriais do SVH, que definen a *ventá óptica visual*. Proponse ademais unha definición coherente de saliencia, así como o correspondente modelo computacional. Demóstrase como este modelo se deriva de xeito natural do cadro proposto no capítulo anterior. Así, está baseado na adaptación contextual a curto prazo da representación do espazo de características ao contido dunha escea específica. Para cada orientación e compoñente de cor, a saliencia compútase como o cadrado do módulo no espazo de características multiescala branqueadas. A saliencia final resulta da suma destas saliencias parciais para cada canle de cor e orientación. Como se verá, os resultados do modelo son practicamente independentes do método empregado para branquear, estea este método baseado na análise de compoñentes principais ou na análise de compoñentes independentes. Unha implementación concreta do modelo descríbese polo miúdo.

- No capítulo 4 avalíase a capacidade do modelo para predicir fixacións oculares. Empregando un procedemento de uso extendido, demóstrase que o AWS supera outros modelos do estado da arte na predición de fixacións humanas, tanto en termos de funcionamento como de robusteza. Isto faise sobre dúas bases de datos de seguemento ocular de libre acceso. Mais, como se probará, as incertezas proporcionadas por este procedemento non reflicten a verdadeira variabilidade do resultados entre esceas. Esta observación lévanos a propor unha comparación coa capacidade preditiva dos propios humanos. O AWS revela ter un funcionamento equiparable ao humano promedio e semella estar libre de nesgos de deseño, a diferenza doutros modelos que teñen problemas evidentes ante simetrias e características salientes de altas frecuencias espaciais.

- No capítulo 5 demóstrase como o modelo logra reproducir unha selección representativa de resultados psicofísicos descritos en experimentos con humanos. A saber: a non-linearidade fronte ao contraste de orientacións; a linearidade fronte ao ángulo de esquinas; a asimetria de presenza-ausencia e a lei de Weber; a influencia do fondo sobre as asimetrias de cor; os efectos de emerxencia (*pop-out*) de orientación, cor e tamaño; unha variedade de exemplos de procura eficiente (paralela)

e ineficiente (en serie); así como o comportamento de humanos baixo diferentes arranxos de similaridade branco-distractor e de heteroxeneidade de distractores.

- No capítulo 6 exponse unha selección de aplicacións do modelo e as implicacións das mesmas. O AWS posibilita a separación de proto-obxectos do seu contexto mediante unha simple segmentación do mapa de saliencia. Asemade, demóstrase que o uso do modelo mellora considerablemente a eficiencia nunha aproximación ao recoñecemento de esceas na navegación de robot. A estratexia de branqueado adaptativo proposta permite tamén o uso sen modificación ningunha do código con imaxes multiespectrais e hiperespectrais, mediante a simples substitución dos sensores (R,G,B) por outros calquera con diferentes propiedades espectrais, abrindo a posibilidade de segmentar e analisar imaxes multi e hiperespectrais. Isto poderíase aplicar a problemas de imaxes de satélite e a análise multi e hiperespectral de curto alcance. Até onde nós sabemos, este é o primeiro modelo bioinspirado de saliencia en aplicarse neste ámbito dos sistemas de visión. Tamén se apunta á posible aplicación do modelo sobre outros tipos de sensores físicos, capaces de producir unha representación do espazo.

- Finalmente, trázanse as conclusións e sinálanse os camiños abertos para o traballo futuro.

Introduction

Biological vision establishes a wide variety of unrivaled benchmarks in terms of efficiency, robustness, and general performance in active visual tasks. Despite the complexity and variability of natural images, visual systems of mammals are surprisingly skilful in recognizing objects and contexts at a first glance and to efficiently drive few fixations to the most salient parts of a new unknown scene.

These capabilities demand an active and dramatic selection of information that poses a main cause for visual attention. It seams reasonable considering the huge flow of information entering the human visual system (HVS) through the retinian photoreceptors, estimated to be over $10^{10} bits/s$ [AvEO05]. Bottom-up adaptive processing and perception of saliency, are thought to lie at the basis of this early behavior with such a remarkable efficiency. They appear to play an essential role in the control of human visual attention –working in cooperation with top-down control– as a number of results from a wide variety of experiments have shown.

Visual saliency is usually employed to refer measures that aim to quantify the conspicuity or distinctiveness of a visual stimulus. That is, it intends to quantify how much a stimulus stands out from the context, given its physical properties. The common representation of saliency is given in the form of a retinotopic map (the saliency map). A main –though in no way the only– source of information to understand the functioning of visual attention is the spatial distribution of human eye fixations obtained in eye-tracking experiments. Eye movements result in fixations that determine the small regions of a given image that are sensed by the fovea. Under good illumination conditions (i.e. for photopic vision), these small regions receive a much higher spatial resolution due to the much higher density of photoreceptors present in the fovea. Consequently, eye movements represent a first form of strong spatial selection of visual information. In figure 2, an example of a saliency map as well as the corresponding density map of eye fixations for a typical image is shown. It gives an idea of the actual implications of a saliency-driven spatial selection. It must be noticed however that also peripheral vision is

Figure 2: Examples of a saliency map (top center) and a map of density of fixations (top right) for a typical image (left). Below each map, the result to superimpose it to the image is shown.

affected by attentional selection, without the need of eye-movements. This issue will be considered further along this dissertation.

Otherwise, mechanistic models of early visual processing with a biological concern are focused on the explanation of the visual receptive fields and their adaptive behavior, to local and contextual features. A main goal of these models is the formulation of early coding strategies that are biologically plausible and that are able to explain observed visual phenomena related to early vision, and particularly to contextual adaptation of perceptual and neural behavior.

The problem of measuring the saliency or distinctiveness in an image has also a great relevance in computer and machine vision, specially in the development of active systems. Indeed, bottom-up spatial attention has shown to be very useful in important visual functions like learning and recognition and many vision applications as shown in the first chapter of this thesis. Besides, the extraction of suitable low level features is of enormous importance in image analysis and computer vision. Both of these problems –low level representation and saliency– use to appear closely related in a variety of solutions. A remarkable example can be found in the most popular interest point detectors, but also in many other computer vision models.

Both concerns on the understanding of the HVS and on the development of active vision systems have fostered an important and crossdisciplinary research effort to provide improved measures of saliency. Particularly, the bioinspired modelling of saliency and its applications have seen an extraordinary and increasing amount of research efforts in the last years.

However, there is clearly a lack of models that address the relationship between the contextual data-driven adaptation observed in early visual coding and the perception of saliency. Understanding this relation is essential for the development of a computational framework of early visual coding with

biological plausibility. Such a framework should formulate plausible intermediate retinotopic representations adapted to the image. These intermediate representations must be able to maintain a suitable measure of saliency, but also to match observed characteristics of early vision. Approaches to this problem are very interesting for computer vision too, as far as they may yield improved models of both adaptive low level features and saliency.

Furthermore, most models of saliency are grounded on an information theoretic foundation, without an specification of the physical sources involved, and more importantly, of the different ways in which they contribute to visual saliency. This specification, if possible, is very important since it would offer an additional constraint to understand the visual function in terms of its physical roots. As well it could yield excellent cues for the development of active vision approaches and in general for the adaptive processing and analysis of images.

With the aim of filling these gaps, this thesis provides a coherent functional approach to both early visual coding and saliency, in a biologically plausible manner. Likewise, the framework proposed is rooted in a physical interpretation involving few simple optical magnitudes. The resulting model is shown to explain a variety of visual illusions and to clearly outperform the existing state-of-the-art models of saliency using the most popular evaluation tests, including the prediction of eye fixations and the reproduction of psychophysical results.

The first pointed lack can be easily appreciated in the two typical strategies of low level representation adopted by existing models of saliency. Many of them start with multiresolution decomposition of three predefined color components, in a given color model. This is done by projecting the image color components on linear filters resembling receptive fields of cells in V1, which are usually modeled by Gabor-like and Gaussian-like functions ever since the standard model of V1 was first proposed by Hubel and Wiesel [HW59, HW68]. The following steps generally involve a competition and integration process that delivers a final measure of saliency, a scheme already found in early models based on the Koch and Ullman architecture of attention [KU85]. Otherwise, the other typical approach involves decomposition through the projection of the image on independent components of natural image patches, avoiding color components and filter parameterization beyond patch size. This proposal is based on the statistical interpretation of the standard model of V1 as the result of evolution and neural development to match the statistics of natural images [O$^+$96, BS97].

Both of these schemes, either based on filter banks or on independent components analysis, share an important property: they always use the same portions of the feature space to represent any image. Filter bank approaches

project a fixed set of color components on a fixed partition of the spectral domain. Independent components are determined from a set of training natural images and are not modified subsequently.

The described static approaches to early coding underlying most of current models of saliency do not match the behavior of the HVS. Indeed, it adapts its responses to the global and local features of each specific image. It shows short-term and contextual adaptation to contrast, to color content and to spatial structure. This adaptation takes place from photoreceptors and G cells to cortical cells and has been shown to produce overall a decorrelated representation [BF89, RR09, Koh07, CWS$^+$07, SHD07]. Adaptive decorrelation seems thus to be a plausible neural mechanism. Not surprinsingly, many recent mechanistic models of neural cortical networks as well as models of computation by populations of neurons produce an overall decorrelated and whitened representation of the input.

From a computational point of view, there are also reasons in favor of a contextual adaptation model. Approaches that do not present such adaptation are more likely to be affected by feature biases, reducing the applicability of the corresponding measure of saliency.

Therefore, the problem of saliency appears to be closely related to the problem of selection of a low level representation as well as its adaptation. In the context of biological vision, early visual coding appears to be an unavoidable problem to tackle wether biological plausibility is claimed. Otherwise, a proper insight in early visual coding can deliver guidelines to design low level representations of images, suitable for active visual functions that migth be useful for computer and machine vision applications. Indeed, and similarly to other works in the field, the original motivation of this dissertation was born within a long-term project of developing a generic and biologically inspired framework to approach and study active vision problems.

Hypothesis and objectives

The working hypothesis assumed in this thesis is that contextual adaptation arising from the adaptive whitening of low level features is the key factor involved in the determination of visual saliency. Thus, as well as there is a long-term adaptation of neural coding driven by *natural images* statistics, there is also a short-term contextual adaptation of early coding driven by particular image statistics. An implicit assumption in such a hypothesis is that the main underlying computational mechanisms of contextual adaptation are decorrelation of responses and contrast normalization.

Relying on these ideas, this dissertation is focused on the investigation, in terms of simple magnitudes, of both contextual adaptation of the low level

representation and a coherent definition of visual saliency.

Consequently, the following three major objectives are pursued:

- The proposal of a theoretical framework able to explain in a coherent manner a number of phenomena related to contextual adaptation and visual saliency. This framework must be biologically plausible and hence it must accomplish with a number of constraints impossed by the known behavior of the HVS.

- The implementation of a computational model of saliency that outperforms previous approaches, in terms of reproduction of results in visual experiments with human observers. These experiments must include representative examples of those devoted to overt and covert attention, that means with and without involvement of eye movements. Among them, a main benchmark is the prediction of human fixations in free surveillance of images using open access eye-tracking datasets.

- The demonstration of the usefulness of the model of saliency in a variety of applications, since applicability can also be seen as a condition of validity for any new model. Given the huge amount of applications, a selection of three application goals has been done. The first is to show the usefulness of the proposed measure of saliency as the basis for generic figure-ground segmentation, a main problem of image analysis. The second goal is the improvement of landmark and interest points selection, a main issue in robot navigation problems. The third goal is to extend the field of applications of saliency, particularly to the analysis of non-visual spatial representations, since the approach adopted here is theoretically grounded on a physical –generalizable– aproximation to early vision.

Contributions of this thesis

The main contributions of this dissertation can be hence summarized as follows:

- A new insight is achieved in contextual adaptation and early –data driven– visual coding in the HVS, through a simple framework of forward whitening of color and scale components. An image representation adapted to image statistics is thereby proposed in a simple computational form that is able to explain a variety of visual illusions.

- A definition of the corresponding measure of saliency is derived as the modulus in the obtained whitened representation, which is proposed

to estimate an invariant in the HVS. The resulting model is hence named as adaptive whitening saliency (AWS). Likewise, this measure of saliency is shown to be directly related to an equivalent definition of *optical variability* in terms of spectral wavelengths and spatial frequencies from a typical description of an image in Fourier optics. Besides, this link yields an explanatory goal to the coding catastrophe within the efficient coding hypothesis, in terms of invariance of the HVS to cope with optical variability in the image, inside a defined *optical visual window*.

- The use of the predictive capability of human fixations shown by humans themselves is proposed as a reference to improve a popular measure based on ROC analysis. This way, the assessment of model performance against inter-scene variability is improved, and valuable information about robustness of models or saliency strength is obtained. The AWS model exhibits a performance equivalent to humans, clearly outperforming other state-of-the-art models that appear to suffer from different feature biases. Otherwise, the AWS model is shown to reproduce a representative ensemble of psychophysical results, to our knowledge not reproduced together by any other model before.

- The applicability of the AWS model in problems of computer and machine vision is demonstrated. Particularly, a straightforward application of AWS to multispectral and hyperspectral images, to our known not proposed with any other bioinspired model of saliency before. It is also important the demonstration of equivalent results yielded by the model both using a compressed representation extracted from many narrow spectral sensors, and using a classic trichromatic representation from broadband detectors. As well, a proposal to check the correct projection of physical variability in techniques of sensor fussion for spatial visualization is pointed, by applying the AWS model on both the original and the displayed data.

Thesis outline

This thesis is organized as follows.

- In chapter 1, the concept of saliency, its role and functioning in human vision, the existing computational models, and its fields of application are reviewed in some detail.

- Chapter 2 is devoted to the investigation of early visual coding. A simple functional framework that enables contextual adaptation is pro-

posed. Adaptation is accomplished by the whitening on color and scale features. Whitening implies the adaptation of the dynamic range of each feature dimension to the specific data observed in a given scene. It is first applied to color components and it is followed by a multiscale decomposition of the whitened color components. This is done for a number of orientations. Next, scale features are further whitened. Such a representation is analyzed in the light of several psychophysical phenomena related to visual illusions using synthetic images, artistic pictures and natural images. As well, its biological plausiblity is assessed in the light of known properties of the HVS.

- Chapter 3 investigates the links of the adaptive whitening proposal with a simple optical description of images. As a consequence the adaptive whitening approach is shown to be directly related to a simple definition of *optical variability* in function of spectral wavelengths and spatial frequencies, when computed in the sensorial limits of the HVS denoted as the *optical visual window*. A coherent definition of saliency in proposed, as well as the corresponding computational model. This model is shown to be naturally derived within the framework proposed in the previous chapter. Therefore the model is based on the short-term contextual adaptation of the feature space representation to the contents of a specific scene. For each orientation and color component, saliency is computed as the squared modulus on the whitened multiscale feature space. The overall saliency is the result of simple summation of the conspicuities for each color channel and orientation. As we will see, the results of the model are practically independent of the method used to whiten, being this method based on principal components analysis (PCA), or on independent components analysis (ICA). A concrete implementation of the model for experimental evaluation is described in detail.

- In chapter 4 the capability of the model of predicting human fixations is evaluated. Through a widely used assessment procedure, the AWS will show to outperform other models of the state of the art in predicting human fixations, in terms of both performance and robustness. This is done on two different open access eye-tracking data sets. But as it will be also shown, the uncertainties provided by that procedure do not reflect the actual inter-scene variability. This observation leads us to propose a comparison with the predictive capability of humans themselves. The AWS reveals to have equivalent performance to the average human and seems to be free of design biases, unlike other models

that have evident problems with symmetries or high frequency salient features.

- In chapter 5 the model is shown to be able to reproduce a representative selection of psychophysical results decribed in experiments with humans. Namely: the non linearity against orientation contrast; the linearity against corner angle; the presence-absence asymmetry and the Weber's law; the influence of background on color asymmetries; the orientation, color and size pop-out; a variety of examples of efficient (parallel) and inefficient (serial) search; as well as human behavior under different target-distractor similarity and distractor heterogeneity arrangements.

- In chapter 6 selected applications of the model and their implications are shown. The AWS allows the separation of proto-objects from context by means of simple segmentation of the saliency map. As well, the model is shown to improve efficiency in an approach to scene recognition in robot navigation. Besides, the proposed adaptive whitening strategy can be used with multispectral and hyperspectral images, through a simple replacement of (R,G,B) sensors by any others with different spectral properties, offering a way to segment and analyze multispectral proto-objects. This could be applied to satellite imagery and to close range multispectral and hyperspectral analysis. To our knowledge, this is the first bio-inspired model of saliency to be applied in that field of machine vision. Its further applicability to manage other kind of physical sensors able to produce a representation of the space is also pointed.

- Finally, conclusions are drawn and open paths for future work are pointed.

Chapter 1

Saliency: Concept, Models and Applications

It is worth remarking from the beginning that in this dissertation the term saliency will be used as arising from only bottom-up, data-driven processes. Therefore, saliency will be broadly conceived like a measure or estimation of the spatial conspicuity or distinctiveness of a point or region in an image. This approach to the concept of saliency is very frequent in the use that traditionally receives in computer vision, as a measure that provides a front end to select landmarks, interest points or regions of interest, in general purpose descriptors used for learning and recognition, for segmentation, and in general for any task requiring an unsupervised selective pre-processing. A point or region that stands out from the context by its physical properties (e.g. color, size, structure) use to still do it after moderate variations on the illumination or the point of view. Thus, a measure of saliency is expected to provide a high degree of invariance and robustness under perspective, orientation, scale or illumination transformations. This makes such a kind of measures very interesting and potentially useful.

Besides, the stated use of the term saliency also agrees with a terminology used in the context of neuroscience, which differentiates between three types of measures: saliency, relevance and priority. As explained by Fecteau and colleagues, the aim of such a differentiation is to provide a clear ground to manage with neural phenomena that drive human attention and arise respectively from only bottom-up stimuli (saliency), from only top-down choices (relevance), or from the mechanisms that efficiently combine both of them (priority) [FM06].

1.1. The computational and information theoretic concept of saliency

As pointed above, saliency has been intuitively or subjectively modelled in many computer vision approaches. It is generally used as a front end to detect interest points or regions, without a main concern on any kind of justification beyond the good performance of the model for the corresponding purpose. For instance, the most popular schemes of interest point detectors and descriptors like SIFT or SURF, use an interest (saliency) map to facilitate further stable and distinctive point selection in an efficient manner [Low04, BETG08].

However there are also a variety of generic approaches to the concept, trying to derive it in a principled manner. The underlying goal is to provide a generic framework for an unsupervised task-independent and efficient computation of saliency. Currently, the most accepted view justifies saliency in terms of probability. The most improbable or unpredictable a local low level feature is, the most salient it is. Hence the different models of saliency are justified as a suitable and efficient approach to compute the inverse of the probability of the local features.

Several state-of-the-art bioinspired models that will be referred along this dissertation, are grounded on this conception. They define a low level representation, and subsequently they propose a measure that approaches the computation of the inverse of the probability density. Some of them also claim for the biological plausibility of the proposed measure in terms of neural computations.

An interesting and popular approach has been done by Kadir and Brady [KB01] that already points in the direction of adaptive interaction between scales. They proposed that there are salient scales that define saliency in a given point. To find them, they measured the entropy on the neighbourhood of each point at different scales, and they took the scales that present a peak of entropy. The most interesting aspect of this approach is the proposal of a local selective interaction between scales to determine local saliency, that avoids rigid schemes like the highly frequent center-surround differences modelled through differences of gaussian filters.

Much more recently, in a formal approach, Loog and Lauze showed that the interest map of the Harris detector, is inversely proportional to the probability density of a local *uncommitted* low level feature [LL10]. They claim that this straight relation with the Harris interest map gives it a strong support in the computation of saliency over other approaches. Since most bioinspired approaches are principled in an estimation of the local probabil-

ity density, they even suggest a probable superiority of Harris detector in the context of modelling of human visual attention. However they do not provide experimental evidences in this regard.

This discussion around the theoretical foundation of visual saliency will be taken up again in the chapter 3, but in terms of the optical variability existing in the perceived image. Saliency will be formulated as a measure of the local spatial distinctiveness existing across the physical magnitudes that an image sensor is able to sense.

1.2. Computational models of human attention

Biologically plausible models of saliency have arisen in the context of the research in human visual attention. The first approaches to attention were motivated mainly by the need of a systematic and comprehensive explanation for a variety of psychophysical observations. There were also proposals arisen in the context of neurophysiological observations and neural networks theories. These initial concerns, mainly focused on particular aspects (visual search phenomena, models of neural networks) from concrete disciplines, have converged in a huge crossdisciplinary effort to explain human attention.

The feature integration theory (FIT) by Treisman and Gelade [TG80] marked the starting point for the development of computational models of visual attention. Its main contribution lies on the proposal of an early parallel processing of simple, *integral*, features able to capture attention, in opposite to the serial -sequential- process of attention needed to detect *conjunctions* of features. As a remarkable result from this parallel processing of few features proposed and maintained by Treisman in several works, arises the explanation of both pop-out effects observed in visual search experiments with humans for certain features and the serial search observed for the conjunction of those features. These experiments pointed that stimuli clearly different in one unique feature from an almost homogeneous surrounding rapidly attract our glance without the need of examining the scene, regardless of the number of nearby objects acting as distractors. In contrast, when distractors were clearly heterogeneous, or when the target differed from all of them in a combination of features rather than in only one, subjects seemed to need to examine the scene object by object, checking for a match with the target. So that the time spent in searching grew linearly with the number of distractors. Treisman held that this could be understood if parallel processing of features exhibiting pop-out effects was assumed. Once saliency was determined from

this parallel processing of a number of features, search was serial on the basis of this final measure of saliency. Thus only the feature map corresponding to the unique different feature in the case of a singleton target, would strongly fire in the location of the target, thus conveying this high value to a final map of activation, and directing attention to it. On the other hand, in the heterogeneous and conjunctive cases none -or several maps in different locations- would fire, without providing for a clear salient location, thus explaining the need for a serial search. This theory fostered the search for simple features responsible for pop-out in the HVS, but also provided a suitable ground for the computational modeling of visual attention.

These ideas were gathered by Koch and Ullman, to conceive a saliency-based computational architecture of attention [KU85]. They proposed, in agreement with Treisman, the parallel computation of retinotopic feature maps and the integration of the activity of these feature maps at each position in the corresponding position of a unique map of saliency. This map of saliency would guide attention, directing it to points with highest values of saliency. They also introduced a winner take all (WTA) network to determine the next most salient region, combined with a mechanism of inhibition of return (IOR) acting on the saliency map, to allow for a dynamic selection of different regions of a scene in the course of time. This architecture is essentially bottom-up, although they pointed the possibility of introducing top-down knowledge through biases of the feature maps. Besides, this proposal have had a great influence in the development of computational models of saliency. The figure 1.1 shows a scheme representing this model, adapted from [KU85].

An important subsequent model of attention trying to explain more results on visual search experiments is the Guided Search Model hold by Wolfe [Wol94]. In this model, feature dimensions (color, orientation, size) rather than features (vertical, green, small, etc.) are assumed to be processed in parallel and, therefore, to have an independent map of *activation* (saliency) extracted from the *input categorical channels*. Besides, top-down influences are considered by means of top-down *activation* (relevance) maps for each feature dimension. Top-down maps are extracted directly as well from the *input categorical channels* through rules specially designed for a given task. All of these *activation* maps are further combined through weights that are task-dependent. Interestingly, while weights of top-down maps are allowed to be zero, weights of bottom-up maps are only allowed to be reduced up to a minimum (non-zero) amount.

There are many other models of attention that were conceived mainly from psychophysical and neurophysiological observations -not only related to visual search-. Many of them claim for a biological plausibility by pro-

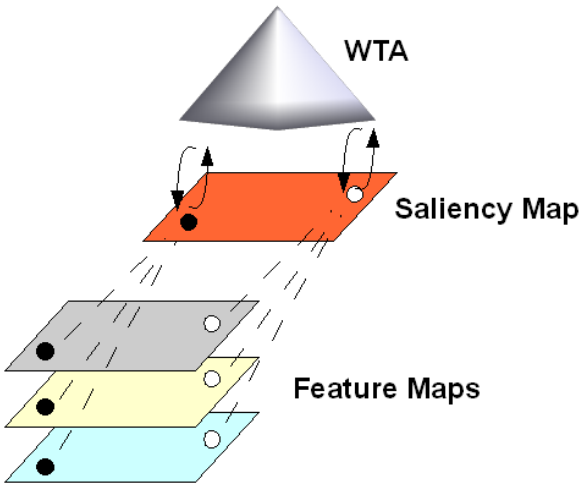

Figure 1.1: Koch and Ullman architecture of visual attention

viding a detailed description of a neural circuitry linked to known data from neurophysiology. This is the case of the adaptive resonance theory to model attention proposed by Grossberg [Gro76,CG03], the neural model of dynamic routing of information by Olshausen et al. [OAE93], the FeatureGate model by Cave [Cav99], the neurodynamical approaches hold by Deco and coworkers [DZ01, DR04], or the model of bottom-up saliency coded in V1 cells by Zhaoping [Zha02].

Meanwhile, other models are motivated by the study of attention taking advantage of the information theory, trying to catch and describe the strategy of information processing of the HVS in terms of formal principles or statistical descriptors. Therefore Tsotsos [TCW+95] proposed the Selective Tuning Model, exploiting the complexity analysis of the problem of viewing, and achieving in this way several predictions on the real behavior of the HVS. Rajashekhar et al. [RvdLBC08], have studied the statistical structure of the points that attract the eye fixations of human observers in natural images in surveillance and search tasks. In this way, they have modeled a set of low level gaze attractors, in the form of filter kernels.

Finally, other models of attention have focused more on top-down than in bottom-up aspects. An outstanding example was provided by the extensive work conducted by Oliva and Torralba on modelling contextual influences on attention. They proposed a simple scheme to introduce the reasoning on the gist of a scene and its layout as driving scene recognition at very early stages [OT01, TOCH06].

1.2.1. Sources of information about visual attention

In the development of the previous theories, the initial sources of knowledge about visual attention have been largely broadened, to feed, support and refuse a variety of proposals. Here are briefly mentioned some of these research techniques that will provide arguments for the development of this dissertation. The main sources of information can be gathered in three groups according to its origin: psychophysical experiments, neurophysiological data and the statistical and computational analysis of images.

Psychophysical experiments have a long history. They have shed light on key aspects of visual attention from the beggining of the research in this field. Pop-out effects, evidences for efficient and inefficient search, asymmetric behavior, influence of distractors, of heterogeneity, of similarity, of different kinds of feature contrasts, of binocular rivalry, and a long etcetera have provided invaluable information and constraints to build models. In many cases these constraints affected to the covert attention, that is to say, to attention that does not involve eye movements. Besides, overt attention that is related to eye movements has been largely studied through eye-traking experiments. As we will see, the spatial distribution of eye fixations is a quite straigthforward and clear reference of priority. It allows, hence, for the quantitave assessment of measures of saliency and relevance not only on synthetic images, but also on natural images.

Neurophysiological data from single cell recordings have described much about receptive fields, contrast, color, scale and orientation sensitivities, as well as neural organization. Multielectrode recording techniques like EEG or visualization techniques like fMRI, PET, or VSD have provided information about neural activity, and the response of different regions of the brain to different natural and synthetic stimuli. But also other techniques from neurophysiology like TMS have delivered relevant observations in the functioning of visual attention.

Finally, image analysis through statistical and computational tools have also provided powerful concepts, measures and algorithms that have supported the construction of theories and models to describe and reproduce the attentional function, and remarkably the computation of saliency.

Currently, we can indeed find several of the previous techniques involved in a single experiment. Therefore, there are studies of fMRI imaging of brain combined with recording of eye fixations with an eye-tracker, which next analyse the results in the light of a given computational model of neural behavior.

Therefore, a model of saliency that claims for biological plausibility should find support in observations from this *tripod* of sources, and fit the best with

the constraints imposed by the results delivered by them.

1.3. The interplay between saliency and relevance to determine priority in human vision

This problem is often addressed in different ways and from different views. It poses a number of questions like: what is stronger saliency or relevance?, how their relative strength varies with time?; how it varies with the type of scene?; are there separate neural mechanisms coding each of them in the brain?. There is a wide variety of works that have tackled these questions that still remain open. But a deal of worthy knowledge has been aquired in the aims of finding answers. Since the performance of the HVS excceds that shown by computer systems in any minimally complex visual task, the known constraints on its functioning in relation to these questions are also valuable in the design of computer vision systems. Moreover, they are essential when considering the biological plausibility of a given computational model.

1.3.1. Relative strength of saliency versus relevance

Many simple computational models of saliency operating on natural and synthetic images, have shown a remarkable ability to predict human eye fixations as well as a variety of psychophysical results [MCBT06, BT09, GMV08, ZTM+08, SM09]. However many works argue the actual strength of saliency in governing human attention. This question is briefly revised in the following, in the light of the recent works that deal with this issue.

It is worth to start clarifying that we are not trying to refute results that point to a dominant influence of knowledge and top-down mechanisms on the control of gaze during the realization of complex tasks, like for instance driving. The dominance of top-down influences on eye movements when doing strong goal-oriented tasks has been observed early by Yarbus [Yar67], and is still being studied in depth like for instance in the remarkable works of Hayhoe and Ballard [HB05].

However, from a view found in many papers, top-down processes develope a role stronger than saliency in driving attention, even in early free surveillance. An interesting example is found in the work by Einhauser et al. [ESP08] holding that objects predict fixations better than saliency. They support this observation from results in experiments involving eye-tracking data, human segmentation of the objects after surveillance, and comparison

with the model by Itti and Koch [IK00]. But this work presents two main weaknesses. First, there are available models of saliency of the state of the art able to predict fixations in natural images clearly better than the employed by them. Second, the design of the eye-tracking experiment, where observers where asked for remember objects following each image, could have introduced a top-down bias towards *objects*. Also recently, Birmingham et al. [BBK09] reported that saliency does not account for fixations to eyes in social scenes, but again they have used the same model that suffers from poor performance as well as strong design biases. Again, this fact could explain at least in part, the reported results as arising from a poor computation of saliency.

Taking into account the influence of contextual cues observed in psychophysical experiments as well as the ability to recognize scenes even with displaying times as low as 30ms, Oliva and Torralba [OT01, Oli05, OT06, TOCH06] proposed that attention was guided by the gist of the scene. They modelled fast scene recognition through feedforward processing of low spatial frequencies. Since this recognition is supposed to subsequently drive attention, this model can be seen as supporting for a stronger role of top-down mechanisms to determine priority, even for early fixations. However the initial feedforward scheme of processing to characterize the scene is clearly stimulus-driven. Thus, it could be interpreted as evidence of a fast feedforward stimulus-driven representation that serves as ground for scene recognition, and even object recognition. This question will be examined further at the end of the chapter 3.

A deal of classical and recent works have shown evidences of the strong role of stimulus-driven mechanisms in human behavior. In a popular review, Wolfe and Horowitz [WH04] went over the features that does guide and does not guide attention on the basis of psychophysical evidences. They provided a list classifying a variety of features, from the lowest level, like contrast, color or orientation, to highest level, like words or faces. They made the classification as a function of the evidence and the probability for each feature of driving pop-out or not. Interestingly, they find several low level features that *undoubtedly* guide the deployment of attention, as well as a number of probable features. Moreover, high level features (e.g. faces, name, semantic category) are clasified under doubtful or probable non guiding features. Besides, the analisys of the low level information content of fixated locations, as opposite to non fixated ones, points to a strong influence of low level features at least at the beginning of subject observation [TBG05, BT06a, FU08]. This has been recently reinforced with the analysis of microsaccades [OTM+08].

Regarding the relative influence of saliency related to the time of observation of a scene, there is an extended view supporting for a decrease with

time. An early observation in this sense was provided by Nakayama and Mackeben [NM89], showing that transient and sustained attention are different components. They hypothesized that transient attention *is operative at an earlier stage of visual cortical processing.* This view has recently received additional support from psychophysical results pointing in the transient effect of saliency in guiding eye fixations [PLN02, vZD06, DS10]. Indeed there is consensus in that consistency in fixation locations between subjects drops under prolonged viewing [TBG05]. However we find different interpretations to this observation. In a revealing work, Tatler and colleagues [TBG05] showed that while consistency between subjects decreases over time, even without forcing a common starting location, there is no evidence for variation in the discrimination between the saliency at fixated and non-fixated locations. They used a number of specifically modelled low level features to account for saliency. Recent results by Foulsham and Underwood agree with this observation [FU08]. In the light of this finding Tatler an colleagues assess four different hypothesis for the involvement of saliency in the course of time: i) saliency divergence with a relative drop of bottom-up influence in comparison to top-down one as proposed by Parkhurst and coworkers [PLN02], ii) saliency rank, that would mean the selection of locations with basis only in saliency like in the model of attention of Itti and colleagues [IKN98], iii) random selection with distance weighting independent of bottom-up and top-down processes as proposed by Melcher and Kowler [MK01], and iv) strategic divergence, which as proposed by the authors means that top-down strategies chosen by observers are different, while the bottom-up frame of reference remains the same. This last possibility is the only compatible with both a decrease in the consistency between observers, even with free starting locations, and the constancy of low level content of fixations over time, both reported in the study. From comparison of eye fixations on natural images between patients with visual agnosia and healthy subjects, Mannan et al. showed that consistency between observers in the very first fixations was equivalent for healthy and unhealthy subjects. However for subsequent fixations only unhealthy subjects (impaired to understand the image) maintained the consistency between fixation patterns [MKH09]. This also points to a constant influence of saliency and an increasing and divergent influence of relevance in the spatial distribution of fixations in healthy subjects.

Just to have an idea of the difficulties involved in this effort to asses the relative strength between saliency and relevance, it is illustrative the recent study by Verma and McOwan, showing that a number of results claiming for top-down influences in change detection were saliency biased. What was supposed to arise from top-down behavior of subjects was however easily explained by a simple measure of saliency [VM10].

In sum, there is a strong support for physical saliency as driving human attention at least in the beginning of free surveillance of images. This is very important, since it would mean that saliency, or in general a stimulus-driven representation, is also a key factor in unsupervised learning of new objects and scenes, and even in spontaneous object recognition. In the following chapters, the capability of predicting human fixations as well as a variety of psychophysical results of a new simple model of low level, feedforward representation of images, and the corresponding measure of saliency, will be shown. These results will reinforce the support for a strong role of saliency in the determination of early priority, but also the support for a strong role of bottom-up mechanisms in the adaptive capability of the HVS to different images. Besides, we will claim that AWS is a more robust and accurate measure of saliency, and thus it is more suitable to be used in studies on the assessment of the relative strength of bottom-up versus top-down processes in human behavior.

1.3.2. Coding of saliency in the brain. How and where

The coding and location of saliency in the brain is also one of the main open questions tackled by literature related to attention, and it has seen an increasing research effort in recent years. The way in wich the bottom-up and top-down attentional functions are deployed remains unclear, and in particular, the existence of some kind of an image-based saliency map in the brain is still under discussion. In a recent review, Fecteau et al. have hold that the concept of a priority map rather than a saliency one, is more probable to find a neural correlate [FM06]. They remark the fact that the term salience or saliency is frequently used with the meaning of priority in many neurophysiological studies that claim to identify the location or coding of its neural correlate. They ground their analysis on four main properties that a neural saliency map must have: i) it should encode spatial visual information in a featureless manner; ii) lessions of its neural substrate should produce defficits in attention; iii) electrical stimulation of part of its neurons should facilitate attention to the corresponding region of the visual field ; iv) it should receive information from the ventral visual pathway to sum the relative saliency of an object. They point that the oculomotor network is known to meet these properties and that relevance is also known to influence its behavior. Besides, they remark that the temporal spiking profiles of these neurons, recorded simultaneously to a singleton pop-out, only allow to discriminate between target and distractor in the recurrent epoch but not in the feedforward one. This is important because this argumentation is supported mainly in observations of the temporal profile of single cells from

the frontal eye field.

According to Zhaoping and coworkers, saliency could be computed at V1, a lower visual area than usually thought. They propose that *V1's neural responses can be used as universal currency to bid for attentional selection, despite the feature tuning of the V1 neurons* [Zha08, Zha02]. This hypothesis is supported by a combination of different psychophysical and neurophysiological observations as well as the predictions of a computational model of V1 [Zha98, ZM07, ZS06]. It implies a dissociation between bottom-up attention and awareness, consistent with additional findings on binocular rivalry [SVP10] and monocular attention [SR10] . It also would remove the need for a master saliency map for the bottom-up saliency. However they point the possibility that other cortical areas could be reponsible for integration with top-down attentional factors. All of this seem to overcome in part the objections posed by Fecteau et al., mentioned above. However it clearly contradicts their assertion about the saliency map as requiring input from later visual areas, which refuses the idea of saliency as a summary of early visual processing.

Indeed, visual areas from the parietal cortex are usually proposed to encode saliency in different forms. Parietal cortex is thought to play a crucial role in saccade updating and in general in attention deployment and analysis of space. Many neurophysiological studies have hold that this area maintains a neural representation of visual priority, with basis in the analysis of recordings of single cells in monkeys as well as human brain imaging by different methods [GKG98, TM04]. In much more recent studies, the right anterior intraparietal cortex is proposed to be the neural substrate for maintaining a priority map across saccades with basis on results under TMS [vKGS$^+$10], and also the posterior superior parietal cortex is proposed to host a priority map from fMRI observations [RFV10]. Saalmann et al. reported that, under a visual matching task, the posterior parietal cortex and the medial temporal area become synchronized. They suggest that this points to posterior parietal cortex as driving a selective modulation of activity in earlier sensory areas to enable focused spatial attention [SPV07]. But these studies do not tackle the analysis of saliency versus relevance.

In an exhaustive review of neurophysiological literature, Corbetta and Shulman find enough evidence for the existence ot two segregated networks devoted respectively to goal-directed selection and to stimulus-driven selection [CS02]. They propose that there exists a bottom-up system that involves the temporoparietal cortex and inferior frontal cortex and that is largely lateralized to the right hemisphere. It would work as a *circuit breaker* for the dorsal system, directing attention to salient stimuli. The top-down system would include parts of the intraparietal cortex and superior frontal cortex.

They also propose that while these systems interact during normal vision, both are disrupted in unilateral neglect. More recently, Buschman and Miller recorded simultaneously cells from prefrontal and posterior parietal cortex, finding that while bottom-up signals arise from the sensory cortex, the top-down signals arise from the frontal. Moreover, they also proposed that bottom-up mechanisms are associated to high frequencies, while top-down signals are associated to low frequencies, and thus emphasizing synchrony at different bands [BM07]. Besides, in an EEG study in humans involved in a face discrimination task, Landau et al. observed different effects of voluntary and involuntary attention on EEG activity in high frequencies [LER+07]. Very recently, Mavritsaki and colleagues have compared fMRI images from subjects involved in visual search tasks with the results of BOLD predictions of a computational model of attention for the same tasks. The obtained results make them to propose that a saliency map is coded in the right temporoparietal junction in agreement with the proposal of Corbetta and Shulman. They also identified separate networks of areas in parietal and occipital cortex, linked to top-down mechanisms of facilitation and suppresion [MAH10].

1.4. Computational models of saliency

The main concern of the models mentioned in section 1.2 was the understanding of the attentional function in the human visual system. Some of them have also been employed in technical applications of computer vision with remarkable achievements. Therefore it is sometimes difficult to establish a clear separating line between biological and technical models. However, we will make this classification here and we will tackle in this section the description of the last group. These models are characterized by either a main concern on technical performance, a principled approach to the concept of saliency, and/or a remarkable contribution to the development of applications in computer vision. Thus, the driving goal underlying them is to deliver an efficient and generic way to select information, to reduce the high complexity of a variety of visual tasks requiring image analysis. In most cases we also find claims of either biological plausibility or a contribution to the understanding of the HVS.

In the 90's we find two particular implementations of the Koch and Ullman architecture being of special interest. The first was made by Milanese and was initially only bottom-up [Mil93], employing as low level features gaussians of opponent color components, oriented gaussian first derivatives of intensity to measure orientation and edge magnitude, and divergence of the

gradient of intensity to measure curvature. These initial maps were further filtered by a conspicuity center-surround operator, involving the difference of oriented gaussians. Within each feature the maps were integrated in a unique conspicuity map by taking the maximum response through scales and orientations. Finally, the feature maps were integrated in a final measure of saliency by means of a relaxation rule. Further on, in subsequent works [MWG+94], a top-down component conspicuity and a motion alerting system were incorporated. The top-down component had the form of a system for object recognition, which applied to a few small regions of interest provided by the bottom-up component, delivered a top-down map favoring regions of recognized objects, and was integrated in the same relaxation process with the bottom-up conspicuity maps to determine the final saliency, highlighting known objects against unknown ones. The alerting system based on motion detection was used to drive attention instead of saliency through switching dependent on an alertness parameter. This model is one of the first efficient approaches to bottom-up saliency computation on natural images, and many of its details have been incorporated in subsequent models.

The second implementation of the Koch and Ullman architecture was hold by Itti et al. [IKN98], who similarly made use of contrast, color and orientation as separate features, in a center-surround approach, but introducing a more simple integration process of weighting and addition of maps at first and of iterative spatial competition and addition in a subsequent work. It starts, like Milanese, decomposing the image in intensity and two color opponent components (RG and BY). These components are further decomposed through filtering with Gaussian pyramids, and with a bank of real valued Gabor filters. In the original version, which is the most efficient and simple, normalization is performed using an operator that favors a low number of local maxima with a value close to the global maximum. In a later version, this operator was replaced by a non-linear and iterative filtering with difference of Gaussian (DoG) filters [IK00], followed by normalization and integration by the summation of the resulting maps. This filtering increases the computational cost and it is too selective. What is more, its performance in predicting human fixations is lower. Anyway, these two approaches to integration were significantly faster than the relaxation rule proposed by Milanese. In a later work, Navalpakkam et al. introduced a top-down component in the model based on the learning of the feature values of a target from training images, yielding a feature vector which is used afterward to bias the feature maps of the bottom-up component. In this way they are able to speed up the detection of a known target in relation to the use of the bottom-up model alone [NI05]. This model has been modified and extended with a variety of features and additional functions. It has been compared

with human performance as shown in section 1.3, and tested in a variety of applications, some of which will be referred in the next section. It is usually a reference for comparisons in the most recent works related to saliency. In sum, it has yielded a number of successful results and applications, being the basis of a wide research activity in the field.

In recent years, more and more new models are emerging with improved results. In the following, we briefly mention a selection of the most remarkable approaches. Le Meur et al. built a model based on a sequence of bioinspired preattentive modules [MCBT06]. They used the Krauskopf color space and also performed a multioriented and multiscale decomposition. Likewise, the competition and integration process involved several bioinspired operations: contrast sensitivity, visual masking, center-surround competition and perceptual grouping. They achieved with this aproach results that improved the model by Itti et al. in the prediction of eye fixations on a number of images. In a subsequent version they added a dynamic component of saliency, enabling the model to be applied in video sequences [MCB07]. Gao et al. studied the hypothesis that saliency-based decisions are optimal in a decision theoretical sense [GMV08]. With this aim, they validated a discriminant center-surround saliency measurement. In this approach, they used the same color space than Itti et al. and a similar filtering process, involving DoG and differences of oriented Gaussians (DoOG). Saliency was obtained through a center-surround discriminant process through the use of mutual information. They studied in depth the biological plausibility of their approach, and also obtained a series of psychophysical results that the model by Itti et al. was unable to reproduce [GV09]. Harel et al. held a markovian approach to the extraction of saliency by means of a graph algorithm from the same feature maps used by Itti et al., and denoted by graph-based visual saliency (GBVS), showing a high performance in predicting eye fixations on a dataset of gray scale natural images [HKP07].

Several recent models propose decomposition through the projection of the image on independent components of natural image patches, avoiding filter parametrization. This proposal is based on the statistical interpretation of the standard model of V1 as the result of evolution to match the statistics of natural images [BS97, O$^+$96]. From these responses, Bruce and Tsotsos [BT09] proposed an approach to visual attention through information maximization (AIM), where saliency is modeled using a self-information measure. It provided remarkable results in the prediction of eye fixations and the reproduction of several psychophysical phenomena. Similarly, Zhang et al. [ZTM$^+$08] proposed a model that computes saliency on independent component patches but through a Bayesian approach. This model was denoted as saliency using natural (SUN) image statistics. Seo and Milanfar [SM09], us-

ing also a statistical-based decomposition, have proposed a self-resemblance measure to obtain saliency, again with remarkable results.

Other recent approach hold by Hou and Zhang relies on the processing of the image in the frequency domain, through a very simple computation of spectral residues [HZ07a], achieving a high performance in terms of processing speed, while keeping state-of-the-art performance reproducing psychophysical results. This approach has given place to other models also computing saliency in the frequency domain. For instance the static and dynamic models of saliency proposed by Guo et al., that also work in the frequency domain, but instead of using the amplitude, they rely on the use of the phase spectrum to compute saliency [GMZ08].

All of these models have been assessed through the capability to reproduce psychophysical observations, the capability to predict human fixations, and also through the comparison with -at least- the model of Itti et al.

1.5. Applications in computer vision and image processing and analysis

Most of the mentioned models have also shown their suitability to provide a generic solution in a wide variety of applications of computer vision. Again the leading position in the number of applications is clearly hold by the model of Itti et al. (and a deal of modified versions). In this section, a brief description of the fields of application of bioinspired models of saliency is provided, along with a selection of illustrative examples. We do not tackle here the use of models of saliency in the study of the HVS, since it has already done in the previous sections. Likewise, we have left appart measures of saliency that have been extensively used in computer vision for interest points or ROI selection, but without a biological concern either in its formulation or in its evaluation procedure.

1.5.1. Visualization

A usual application of saliency is related to improve visualization under a variety of technical constraints, such as spatial modulation of compression rates, or the need of resizing found for instance in thumbnailing or in visualization in mobile devices. Therefore we can find examples of image [OBH+01] and video [Itt04] compression based on saliency-based foveation. Likewise, thumbnailing application has been used as an example presented together with the formulation of a powerful model of saliency by Le Meur et al. [MCBT06], showing its suitability to drive cropping of images for selection

of thumbnail views. This is also the case of Hou and Zhang [HZ07b], who have applied a model of bottom-up saliency to thumbnail generation.

An interesting trend in visualization is found in the application of saliency to boost video and image resizing and retargeting techniques, which illustrates well the benefits of combining saliency with other image processing approaches. Wang et al. have proposed to combine gradient and saliency from the model of Itti et al. in an importance map to drive image resizing with very good results [WTSL08]. Hua et al. have employed saliency tracking for distortion-free video retargeting [HZL+09]. Hwang and Chien have applied the model of Itti et al. in image retargeting through adaptive seam carving [HC08]. Liu et al. have recently shown improved results with a similar approach with different modifications, among others the use of an improved measure of saliency [LYS+10].

1.5.2. Segmentation

Saliency has been assessed and applied in the solution of segmentation problems, comparing the results of saliency based segmentation against handmade segmentation by humans as ground truth. Therefore, Hou and Zhang and Seo and Milanfar have shown the suitability of their models in applications of generic saliency-based object segmentation [HZ07a, SM09]. Achanta et al. have proposed a model of frequency-tuned salient region detection for object segmentation, that achieves remarkable results on a large dataset of 1000 natural images segmented by humans, outperforming a selection of models of saliency based on both pure computational and bioinspired approaches [AHES09]. We also find approaches that use saliency for boosting models that address more especific segmentation problems. For instance, in a recent model of human figure segmentation Soceanu et al. have used modified saliency maps [SBR+10]. Also recently, Sun et al. have proposed a combination of edge detection with GBVS for automatic pre-segmentation of high resolution remote sensing images [SWZW10].

1.5.3. Detection and recognition

Models of attention, and specially models of saliency are showing more and more their suitability for the design of generic models for learning and recognition of objects. Frintrop et al. [FBR05] have dealt object recognition using a version of VOCUS incorporating a depth map from a laser scaner, itself a modified version of the model of Itti et al. Walther et al. have shown that the use of saliency can considerably boost learning and recognition performance, enabling simultaneous learning of multiple objects from

one image [WRKP05]. Gao et al. have applied discriminant saliency in the localization of objects and the classification of images [GHV09]. Han and Vasconcelos have applied measures of saliency and top-down relevance with a HMAX network to provide a biologically plausible approach able to achieve state-of-the-art performance in object recognition [HV10]. Barrington et al. proposed NIMBLE, a model of visual memory based on fixations extracted from saliency [BMwHC08]. They showed its capability to recognize faces from only one fixation. Kanan and Cotrell have recently improved this approach by using natural image statistics in the extraction of features as well as in the computation of saliency [KC10]. Their model outperformed other state-of-the-art approaches clasiffying objects and recognizing faces in popular datasets. Seo and Milanfar have used salient features as descriptors to implement an effective algorithm for training-free generic object detection, that was able to learn objects from one unique example, subsequently achieving a high recognition performance [SM10]. Very recently, a powerful generic object detector has been proposed that uses a measure of *objectness* of a given window, mainly based on the measure of saliency [ADF10]. This objectness measure is shown suitable to learn unknown object categories from surveillance of unknown images, but also to improve performance in detection and localization of known objects and categories.

Several works have theoretically studied the suitability and optimality of spatial attention in object learning and recognition [HK09]. In any case, these recent approaches to generic object recognition, pushing forward state-of-the-art performance, exemplify quite well the benefits of using saliency for driving learning of novel objects from free surveillance, as well as for efficient scene checking for recognition of already known objects.

1.5.4. Robot vision

Saliency is being extensively applied in robot vision as a method to select spatial information. It seams to be quite reasonable since robotic approaches usually aim to mimic human behavior in a given task. This statement particularly holds in the context of humanoid robotics, where there are examples of application of bottom-up saliency, like the approach by Ruesch et al. that fuses visual and auditory saliency to drive the exploratory behavior of the iCub robot [RLB+08]. Frintrop et al. used VOCUS, a modified version of the model of Itti et al., in the basis of an active system for mapping and robot localization [FJC07]. Siagian and Itti have applied a version of their model of attention that incorporated a gist module to rapid scene classification for robot navigation [SI07]. Meger et al. have used spectral residual saliency in a robot vision system able to locate objects and to build an

spatial-semantic map of a region [MFL+08]. Santana et al. have recently employed the saliency maps of Itti et al. using only intensity and colour channels to boost real-time trail detection [SACB10]. Likewise, Montabone and Soto have used saliency-based features inspired in a real-time implementation of VOCUS [FKR07] to boost human detection in a mobile platform, outperforming other state-of-the-art approaches [MS10].

1.5.5. Other applications

Since it is a generic, low level and data-driven tool for the analysis of natural images, saliency has been used in a variety of aplications needing a front end for selection of interest points or regions. Michalke et al. have proposed a driver assistance system combining saliency, relevance, tracking and recognition to provide warnings in dangerous situations [MGS+07]. The system was mainly based in bioinspired models of attention [IKN98, FBR05, NI06]. Tian and Yue have proposed an adaptation of the model of Itti et al. for change detection in remote sensing images, showing a high performance [TWY07]. Mahadevan et al. have used saliency for anomaly detection in crowded scenes [MLBV10]. Huang et al. have proposed a method for image search re-ranking based on saliency, that distinguishes cluttered from uncluttered images using the distribution of saliency inside them [HYZF10]. In a recent work, Parikh et al. have proposed a real time version of the model of Itti et al. for the use of saliency-based image processing jointly with retinal prostheses to allow identification of important objects, in spite of the low resolution of these implants [PIW10]. One such application points in a very interesting direction: the development of improved prostheses able to mimic biological processing of images. It seems reasonable to expect a notable role for a measure of saliency, and in general for human-like bottom-up processing in such an approach.

Chapter 2

Whitening in Early Visual Coding

As described in the previous chapter, existing models of saliency that have incorporated a biologically plausible scheme of image decomposition, have done it in a rigid and non adaptive manner. Other models, like the spectral residual approach have avoided the question of image decomposition, through a direct measure of spectral dissimilarity on the Fourier transform of the image.

Therefore, both of the two widely used schemes to decompose images, either based on filter banks or on independent components analysis, share an important property: they always use the same basis of vectors in the feature space to represent any image. Filter bank approaches project a fixed set of color components on a fixed partition of the spectral domain. Independent components are determined from a set of training natural images and are not modified subsequently. In sum, existing models of saliency rely all the adaptive work in a rigid process of normalization and weighted summation of the initial responses or directly in a subsequent measure of dissimilarity or improbability.

This does not match the behavior of the HVS, which adapts its responses to the global and local features of each specific image. It shows short-term adaptation to contrast, to color content and to spatial structure. A wide variety of neural mechanisms of adaptation have been described all across the visual pathway, from retinal photoreceptors and G cells to striate and even extrastriate cortical cells [RR09, Koh07, CWS+07]. One of the main functional benefits of this adaptation is thought to be the decorrelation of cell responses, in order to improve representational efficiency [Koh07]. Indeed, neural adaptation under natural stimulation has been shown to produce overall a decorrelation of neural responses [VG00, EBK+10, ALR93].

It has been pointed that the undertanding of contextual influences will have important implications for understanding human vision, but also for the development of applications such as adaptive visual aid [SHD07].

Short-term decorrelation of neural responses seems thus to be a plausible adaptation mechanism in populations of neurons. Consequently, in this chapter contextual adaptation is approached in early coding, from an ecological perspective. The underlying hypothesis in such an approach is that adaptation manages to match the response properties of our visual system to the statistics of the images that we see. Whitening from response decorrelation and variance normalization in populations of neurons is studied as a possible overall effect of short-term contextual adaptation. Its biological plausibility will be examined and discussed, and different results on its impact on perceptual performance will be shown.

2.1. Temporal scales and types of adaptation

The homogeneity of objects in natural scenes makes images highly redundant, allowing for instance the prediction with high confidence of luminance and color values of an unknown part of an image from the known values of part of the image, this is a classical observation already pointed by Attneave in the 50's [Att54]. Thereby, this redundancy of information present in natural images that arises from their statistical characteristics, has been long seen as a powerful stimulus for sensory adaptation. Particularly, it has motivated the proposals by Barlow of efficient sensory coding as well as neural decorrelation as a powerful adaptation mechanism [Bar61, BF89]. This is usually referred as the *efficient coding hypothesis*.

In an enlightening review on the relation between natural image statistics and neural representation, Simoncelli and Olshausen [SO01] point out that *an important issue for the efficient coding hypothesis is the timescale over which environmental statistics influence a sensory system. This can range from millenia (evolution), to months (neural development), to minutes or seconds (short-term adaptation).* Indeed, adaptation has been observed to occur in temporal scales as short as a few tens of milliseconds [RR09, Koh07, CWS$^+$07, SHD07].

In a more recent review on the neural mechanisms and models of short-term and mid-term adaptation, Kohn states that *to a first approximation, adaptation effects appear qualitatively similar on a wide range of time scales with more prolonged adaptation resulting in stronger effects* [Koh07]. On the other hand short-term adaptation mechanisms and the related perceptual effects use to be tackled under two different paradigms, one is contextual

adaptation -adaptation to the spatial context- that seems to be mainly related with intra-areal neural connections, and the other is temporal adaptation -adaptation to the temporal context- that implies a certain involvement of memory. Despite this differentiation, these two kinds of adaptation seem to be closely related, both functionally and in terms of their perceptual consequences [SHD07].

The previous observations motivate our approach, aiming to extrapolate -in a first approximation- the observed long-term adaptation of the visual system to shortest time scales, that might explain contextual adaptation. This agrees with the fact that existing *functional characterizations of cortical visual areas suggest that their processing either implicitly or explicitly reflects the statistical structure of the visual inputs* [SHD07]. Therefore, the match of cell receptive fields to the statistics of the set of natural images is briefly reviewed in the next section. Next the match of the adaptation of these receptive fields to the statistics of particular images is formulated and modeled under similar terms, with a concern on biological plausibility.

2.2. Long term adaptation

Arisen from the seminal works of Hubel and Wiesel [HW59,HW68], cortex cells exhibit receptive fields that are selective to scale and orientation while being well localized in space, and that can be approximated by a bank of Gabor-like filters.

This model is also interpreted as the result of the adaptation of the HVS to the statistics of natural images. It has been shown how this kind of receptive fields naturally emerge when imposing sparse coding constraints or when computing principal or independent components from patches of natural gray-scale images [O$^+$96, BS97].

Besides, natural scenes exhibit a limited range of chromatic distributions, giving rise to a limited range of adaptation states, similarly to what happens with spatial structure [WM97]. Consonant with this fact, Hoyer and Hyvarinnen extended the statistical interpretation of cell responses to color natural images [HH00]. This work provided a coherent interpretation of spatial and color receptive fields, as an over-representation of the image being watched, coding it through independent components of the set of natural images. Furthermore, these extracted components are in fair agreement with color coding using luminance and two double-opponent components, RG and BY, as found in psychophysical and physiological observations. With a different approach, Lee et al. have shown from the analysis of hyperspectral images that color opponency emerges as an efficient representation of spectral

properties in natural scenes [LWS02].

Since these interpretations support adaptation to the set of natural images, they find an explanation in a long-term adaptation of the HVS to chromatic and spatial features.

2.3. Short term contextual adaptation

As pointed above there is a wide variety of evidences supporting the fact that neurons adapt their responses to different changes in images in small temporal scales. As small as tens of milliseconds. Numerous evidences, both psychophysical and physiological, suggest that this short-term adaptation acts to reduce dependencies between neurons, driving an active recalibration of neural sensitivity based on experience [Koh07,SO01]. It seems to alter neural responsiveness to take advantage of the available dynamic range, changing cell tunning slopes, and hence stretching or compressing the range of stimuli that influence neural receptive fields.

Besides, at the basis of the efficient coding hypothesis decorrelation is a mechanism of adaptation that allows the HVS to use fully the limited dynamic range of V1 neurons [BF89]. Likewise, gain control and variance (contrast) normalization are two widely accepted adaptation mechanisms that provide a good support for the decorrelation and whitening of responses [Koh07]. Starting from the standard receptive fields resulting from long term adaptation of the visual system, the proposal here formulated is that the adaptive whitening of the corresponding responses, explains contextual adaptation to particular images. Since color and spatial coding are related to differentiated neural mechanisms, we will study separately each of them, but the formal approach to adaptation to both of them will converge in a unique adaptive paradigm based on whitening.

As it will be shown in the subsequent chapters, short-term whitening in populations of neurons will also allow us to explain a variety of phenomena related to bottom-up saliency.

2.3.1. Coding of color

Webster and Mollon pointed that: *highly restricted color distributions characterizing natural images may provide a potent stimulus for adaptation, inducing strongly selective changes in color appearance.* Moreover, *the variability is large enough so that very different contrast adaptation effects will occur for individual scenes, and observers in specific contexts may thus encode colors differently* [WM97]. These statements convey an explanation for

the many evidences already found of such short-term adaptation in both psychophysical and physiological experiments. These evidences had also led to Atick et al. to propose a decorrelating and gain controlled neural network for adaptive color coding in the early visual pathway [ALR93].

In an illustrative study combining both physiological and perceptual data, Wachtler et al. observed the responses of V1 neurons from awake monkeys with a strong color selectivity [WSA03]. Stimulation was made through relatively large color patches, covering all the classic receptive field of the recorded cells. Taken the responses of these neurons as representatives of populations with color sensitivity they found that the opponent component representation observed in LGN, is already recoded in V1 through a rather complex transformation. Besides, they observed changes in tunning properties of cells driven by changes in the chromaticity of the background, that is by the chromatic contrast of the stimulus to the background. Likewise, they also found a correspondence between physiological observations in monkeys and perceptual results of color appearance for humans using exactly the same stimuli, indicating that coding of color in V1 might already contribute to color appearance. Hence color coding admits schemes different to raw RG and BY opponencies already in early visual areas, and its contextual adaptation seems to be closely related to the corresponding perceptual changes.

From psychophysical experiments contextual adaptation has been observed to produce sensitivity changes that alter color appearance by reducing the perceived contrast of colors that are similar to the adapting axis, and by biasing the perceived color of other stimuli away from the adapting axis [WMBW02]. Rosenholtz et al. reported reversal and alteration of color search asymmetries depending on the color of the background. They explained the observed behavior through a simple mahalanobis distance in a McLeod and Boynton color space [RNB04]. Since they did not propose how to combine such a measure with spatial saliency, it is difficult to test it in natural images. But, interestingly, it points in the direction adopted by us, since a mahalanobis distance is in fact a statistical distance, and hence it can be taken as the euclidean distance in a whitened representation of a given original space.

To date, most color-based saliency models, do not consider this contextual adaptation of color representation. This also holds for the decomposition of the image through its projection on the independent components of a large number of patches extracted from natural images, without a specific color model. These approaches to color coding only reflect long-term adaptation, and are not altered by the chromatic composition of a particular image.

In a revealing work, in the context of computer vision without any particular concern on biological plausibility, van de Weijer et al. presented a

method for boosting saliency to take advantage of color in interest point detectors [vdWGB06]. In essence, they proposed to adapt the color representation to the set of images involved. To do this, several axes of the color jet in a particular color model were rotated up to get a decorrelated representation, and next the axes were rescaled up to have a spherical color distribution in the new space. Hence, the norm in such a space serves as a measure of distinctiveness. From this representation, the necessary transformations were applied to obtain saliency with a given interest point detector, improving its performance. However, this approach still does not provide adaptation to the image, but only to the reduced set of images to be further used. It is interesting to note that with different random selections of images from the same dataset, adaptation parameters were obtained that although being close, were still clearly different.

According to the previous argumentation, the use of a color model that is whitened for each different scene is proposed here to account for contextual influences and to efficiently use the available dynamic range. We propose that the L,M,S signals from photoreceptors are involved in a process of adaptation that overall can be modeled by the effects that produces to use other trichromatic whitened representation. This ideally would provide a representation of the chromatic space with points spherically distributed. Actually, for a single natural image points most probably will not be spherically distributed, since there use to be strong higher order correlations in the color distribution. Other important benefit of this approach is that in most of images it improves discriminability of different chromatic regions respect to a rigid representation. On the other hand, in such a representation the modulus is a suitable measure of distinctiveness, able to explain psychophysical results as shown by Rosenholtz et al [RNB04]. But before calculate the modulus we can use this whitened representation to analyse the spatial structure of each of the adapted components. We can hence perform a spatial decomposition of this already adapted representation of color, to further involve it under contextual spatial adaptation.

It can be objected that this is a rather coarse approach to color decorrelation, since it is globally done between three unaltered components, before any kind of spatial filtering. This observation seems to detract from the biological plausibility of our approach because different mechanisms of spatial pooling occur from the beginning in the retina [RR09]. Moreover it does not match the continuum of color selectivities observed in V1 [WSA03].

However, the HVS seems to have specific functional subsystems devoted to color and form processing. Therefore color and spatial selectivity are thought to arise from different neural mechanisms. An important number of results in psychophysics, neuroanatomy and neurophysiology give support

to such a coarse aproach of an amount of independent processing of color and spatial structure [LH87, GM04] . Indeed color components follow very early different visual pathways, while most of spatial selectivity is thought to occur in the cortex. Therefore, in a first approximation color decorrelation can be modeled without taking into account spatial interactions. Moreover, it remains a possibility that in fact decorrelation of color information be produced independently of decorrelation of spatial information. Hence, it is reasonable to suppose that there is a decorrelation between luminance and the two color opponent components coded early in the brain. That said, the simplest option adopted here of modeling color and spatial decorrelation appart seems legitimate and even advisable for the sake of clarity. This holds even more since the aim is to model an overall effect and test the explanatory capability of a simple approach to recoding for representational efficiency, without going down to a mechanistic level involving real neural networks in the brain. That is, the approach here proposed is a functional approach rather than a mechanistic one.

To sum up, whitening of color components constitutes in our view a plausible contextual adaptation mechanism, which boosts and clarifies the further computation of saliency, but also provides an adaptive representation of the image that contributes to an efficient use of the available dynamic range.

2.3.2. Coding of spatial structure

There are many examples of visual illusions produced by the spatial context in an image in which perceptual estimations become sistematically erroneous, accounting for contextual biases in visual processing –see for instance the review by Schwartz et al. [SHD07]. Correspondingly, contextual adaptation of neural responses to spatial frequencies and orientation that rely outside of the classical receptive fields have been observed in a number of physiological studies [Koh07, SHD07]. We find hence many examples that show the existence of strong contextual influences beyond the classical receptive fields [VG00, SGJ+95, SLF03, CBM02]. Besides, neural responses show strong deviations from the standard model predictions under observation of stimuli from natural images, with a contextual content very different from synthetic stimuli [VG02].

One already classic example of contextual influences are illusory contours, which are perceived as sharp boundaries between two regions that do not differ in mean luminance or chromaticity. Cortical cells with orientation selectivity have been observed to respond to these contours. Thus cells without any oriented stimulus in their classical receptive field respond to a contour

that arises from the contextual texture. Grossberg et al interpreted this fact as the result of a context sensitive perceptual grouping taking place in V1 and even in a larger scale also in V2 [GMR97]. Montaser-Kouhsari et al have used fMRI imaging to measure short-term adaptation to illusory contours. They have found that early (V1 and V2) and a number of higher visual areas exhibit adaptation to illusory contours, with increasing orientation selectivity from early to higher visual areas [MLHL07].

Consequently, there is a considerable number of experimental observations that have already pointed to the insufficiency of the standard model to explain the V1 functioning. Olshausen and Field consider that these failures are closely related to experimental setup biases, which affect the standard model, and that must be overcome [OF05]. The description of responses of single cells and the use of synthetic stimuli as some of the most relevant causes of these biases are pointed. Their analysis allows them to assess the fraction of V1 functioning understood in a maximum of a 15%.

It is worth remarking the need of analyzing and modeling the behavior of groups of neurons, rather than single neurons. In a neurophysiological experiment, Chen et al. observed that small visual targets elicit widespread responses in V1 [CGS06]. Their study on the optimal decoding of correlated population responses indicates that *decorrelation at the decoding stage has to do with noise rejection.* Recently, decorrelation has also been shown even in nearby neurons with similar orientation tuning [EBK$^+$10]. Other works also show the need of modeling neural behavior in terms of population codes [PDZ00, ALP06]. It was noticed that this approach can be used to explain temporal adaptation –long, mid or short-term– as a consequence of neural plasticity. We find a remarkable example of such an adaptation in the observation of improvement of orientation coding in V1 neurons by means of practice [SVQO01]. Therefore population codes might help to explain the close links between adaptation to the spatial context and adaptation to the temporal context.

In agreement with these theories, lateral connectivity, present in V1, supports the idea of a collective organised behavior on the visual cortex. Olshausen and Field also drew attention to the numerous neurophysiological and psychophysical evidences indicating that a stimulus with a given orientation produces facilitation for other stimuli in the visual field with similar orientation [OF05]. However, it causes an inhibition, or at least a lesser facilitation, for stimuli that are orthogonal to it. This suggests that V1 neurons have an orientation specific connectivity structure, beyond what is considered by the standard model. Hence, the lateral connections within cortical layers might be responsible, at least in part, for adaptation to spatial features of stimuli and for the deployment of contextual influences. This kind of lateral

connections resembles common implementations of Hebian NN for forward whitening.

The above argumentation led us to propose the adaptive whitening of local scales with orientation specificity. Again, in such a representation of scale composition, distinctiveness can easily be computed as the modulus. It must be noticed that this whitening, although based on a local representation of scales, is spatially global. Thus, the resulting short-term adaptation is matched to the global structure of the image. This provides a suitable ground for the important contextual influences already mentioned. It also agrees with the recent observation that the number of salient stimuli in a display reduces their saliency, regardless of their proximity [WPPF10]. The adaptive whitening strategy for spatial features fits the global nature of the competition for saliency, and its close relation to the capacity limits of perceptual encoding.

2.4. Adaptive whitening, a functional framework approach to early visual processing

Following the line introduced in the previous sections, concrete schemes of decorrelation founded in known characteristics of the HVS have been implemented to study their explanatory capabilities. The four main schemes considered consisted of decorrelation of RGB or LMS color components followed by multioriented multiscale decomposition and: 1) whitening of scales within each orientation; 2) whitening of orientations within each scale followed by whitening of scales within each whitened orientation; 3) whitening of color components using as samples all the scales and orientations followed by the steps of 1); and 4) joint whitening of band-pass responses, mixing orientations and scales.

The first scheme is the simplest one and is justified by the discussion in the previous sections. The second scheme aimed to test the effects of decorrelating not only scales but also orientation, doing it separately. The third scheme was motivated by the impact of spatial structure in color perception and distinctiveness, and the need to check in a simple and general manner possible effects of this interaction. Finally, the fourth scheme was conceived to check possible differences when decorrelating responses without the assumption that decorrelation is constrained by the orientation specific connections observed in the visual cortex.

Since computational complexity of whitening is highly dependent on the number of components (original coordinates) rather than the number of sam-

ples, an efficient coding approach based on whitening will benefit from independently whitening of different kinds of magnitudes. Therefore, if no difference appears between the four tested schemes, it is clear that the lightest and most efficient approach is given by the first scheme.

To implement spatial decomposition, a measure of local energy at several bands of orientation o_i and scale s_j is used. The details of the filters employed are given in the next chapter. Three different methods of whitening have been also considered: z-scores from PCA, and independent components from Jade and FastICA algorithms [CSP93, HO97].

2.4.1. Color statistics and whitened color components

Color statistics in a given natural image present a high degree of correlation in a typical RGB space. This also holds for a LMS space related to the retinian cones responses. In this section this fact is qualitatively examined, as well as the effect of both whitening of color components and their nonlinear transformation to get a Lab space representation.

To this end, a large dataset of calibrated color images [PVV09] has been used. This dataset provides a representation of the XYZ and LMS components of a large number of images. The XYZ components have been used to obtain the corresponding sRGB image and the components in a Lab color model. Besides, the LMS components obtained through the classical characterisation of Smith and Pokorny [SP75] have been used as initial color components for further whitening. This representation of color is related to the spectral sensitivities of cone photopigments. It can hence be taken as a representation of the image in the trichromatic color space determined by the typical retinian sensors of a healthy subject. All the images of the *urban*, *snow and seaside*, and the *natural objects 01* groups of the dataset were processed after a cut from the left side to remove the calibrating sphere present in all of them, with the aim of avoiding a possible bias in the results.

The figures 2.1 to 2.8 present some of the results obtained for eight images that are representative of the typical statistics found all across the dataset. They show for each color representation –among a selection of four different representations –one 3D density plot as well as three 2D density plots with the posible pairs of components. The 2D plots show a typical colormap scale from blue to red that varies linearly with density. For the sake of a good visualization, the 3D plot shows a colormap scale of five steps with a transparency value of 0.9 that ranges from light blue to dark red, and varies linearly with the logarithm of the density. That is, each step represents an increase of an order of magnitude in the density value. Therefore, while the 3D plots catch well the whole distribution also showing the low density regions,

the 2D plots show the high density regions and they are suitable to estimate the real strength of correlations between variables. The color four representations selected are: LMS retinian responses, Lab color model, z-scores from LMS responses, and ICA components using the jade algorithm [CSP93] on LMS responses. In the study, a sRGB and a ICA representation using the FastICA algorithm [HO97] on LMS responses were also used. However, they did not introduce any remarkable additional element for the following discussion. The behavior of the RGB components was very similar to the LMS components, appearing to be only slightly less correlated. As well the behavior of FastICA components did not led to different appraciations than those derived from the comparison between z-scores and jade ICA components. For the sake of clarity, they were not used in the figures that give support to following discussion.

A clear and outstanding fact is that, as expected, LMS components always show a very high degree of correlation. Transformation to the Lab color model removes an important amount of the first order decorrelation present in them. Whitening through z-scores or independent components computation completely removes the first order correlation and makes apparent the higher order correlations present in the color distributions of each image. In principle, the main difference between these two procedures of whitening relies in the fact that independent components use to provide more discriminative directions that match better directions of higher order correlations as they also deal with the reduction of higher order correlations. Otherwise, since these coordinate systems (PCA and most important ICA) simply differ by a rotation, distinctiveness of a given trichromatic color point will be the same in both of them.

Otherwise, images that appear to have a higher degree of clutter tend to show less higher order correlations in their color distribution. For instance the urban scene with bycicles in the image of the figure 2.3 shows much stronger higher order correlations in the whitened components than images in figures 2.5 and 2.6. As well natural landscapes of snow or of a sandy beach and sea, with a low degree of clutter appear to have strong higher order correlations as can be seen in figures 2.7 and 2.8. In general, since scenes dominated by the presence of man made objects present a lower degree of clutter than natural scenes dominated by vegetation, there is a trend of increasing higher order correlation from natural to man made scenes. It is however far to be a valid general rule, and nature scenes can also show higher order correlations as strong as man-made scenes whether clutter is low like in the pointed landscapes.

In sum, first order as well as higher order correlations are very strong in any of the images for a LMS representation. A rigid (non-adapted) represen-

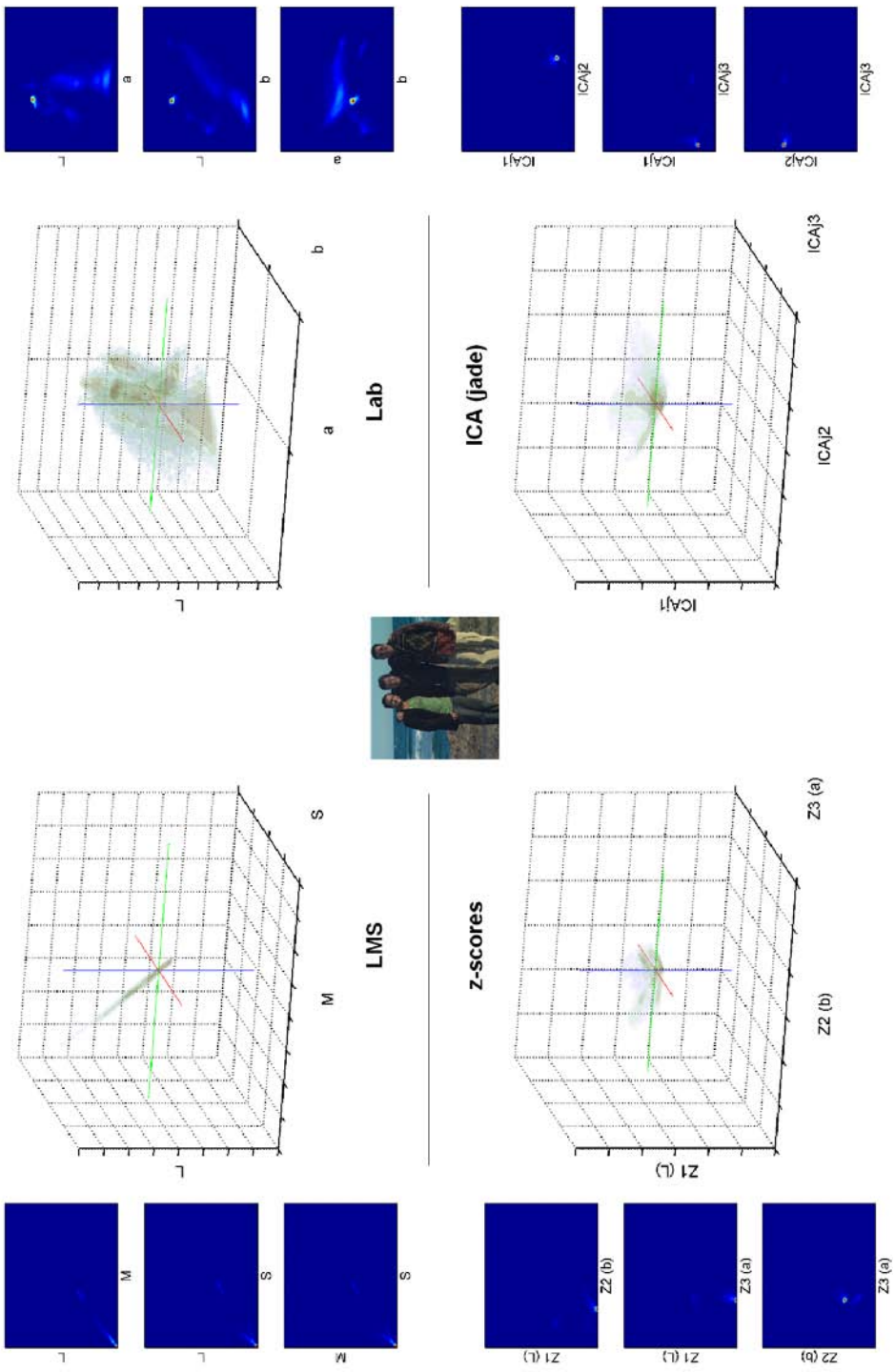

Figure 2.1: Example of color statistics in different representations (I).

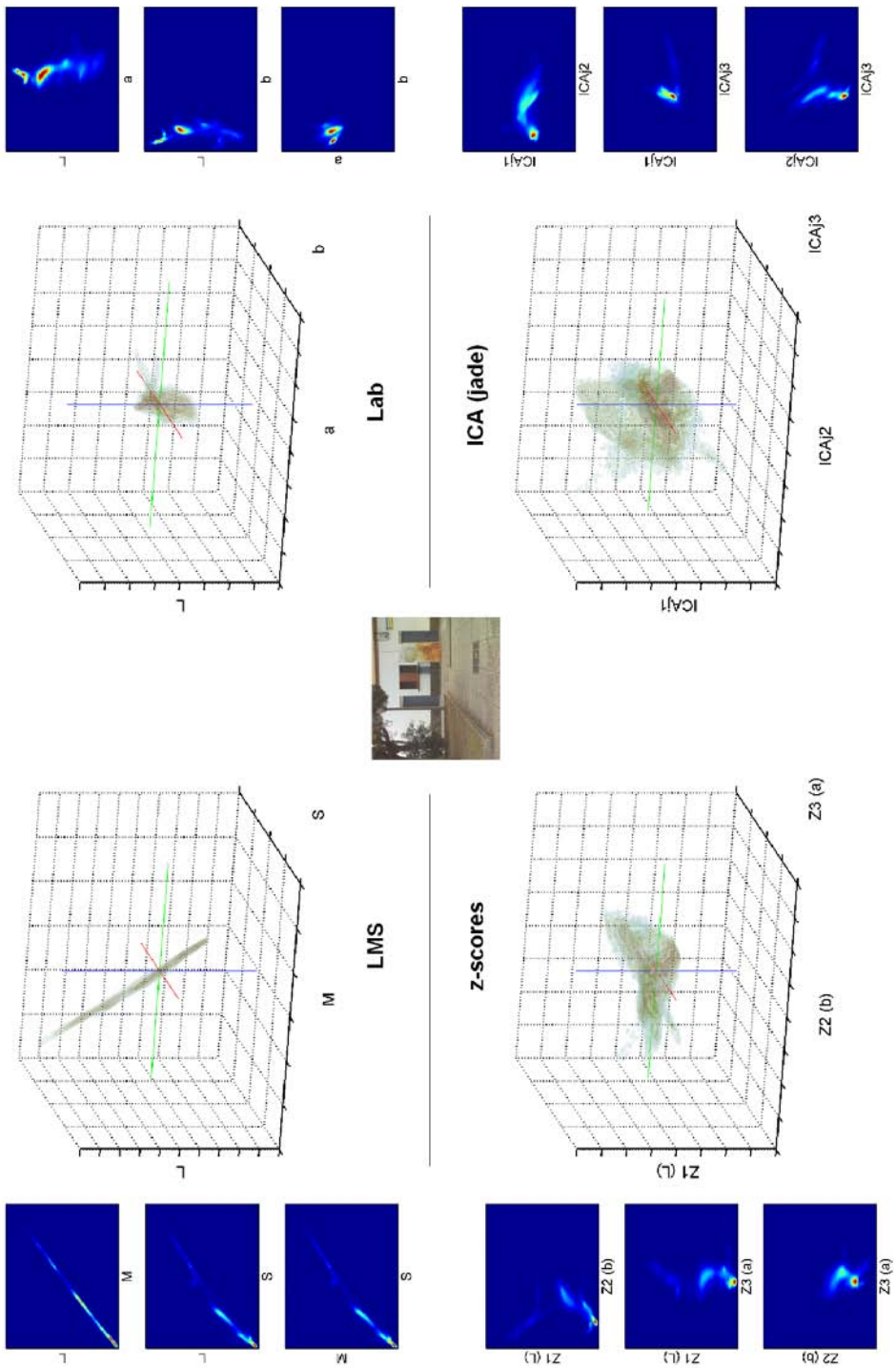

Figure 2.2: Example of color statistics in different representations (II).

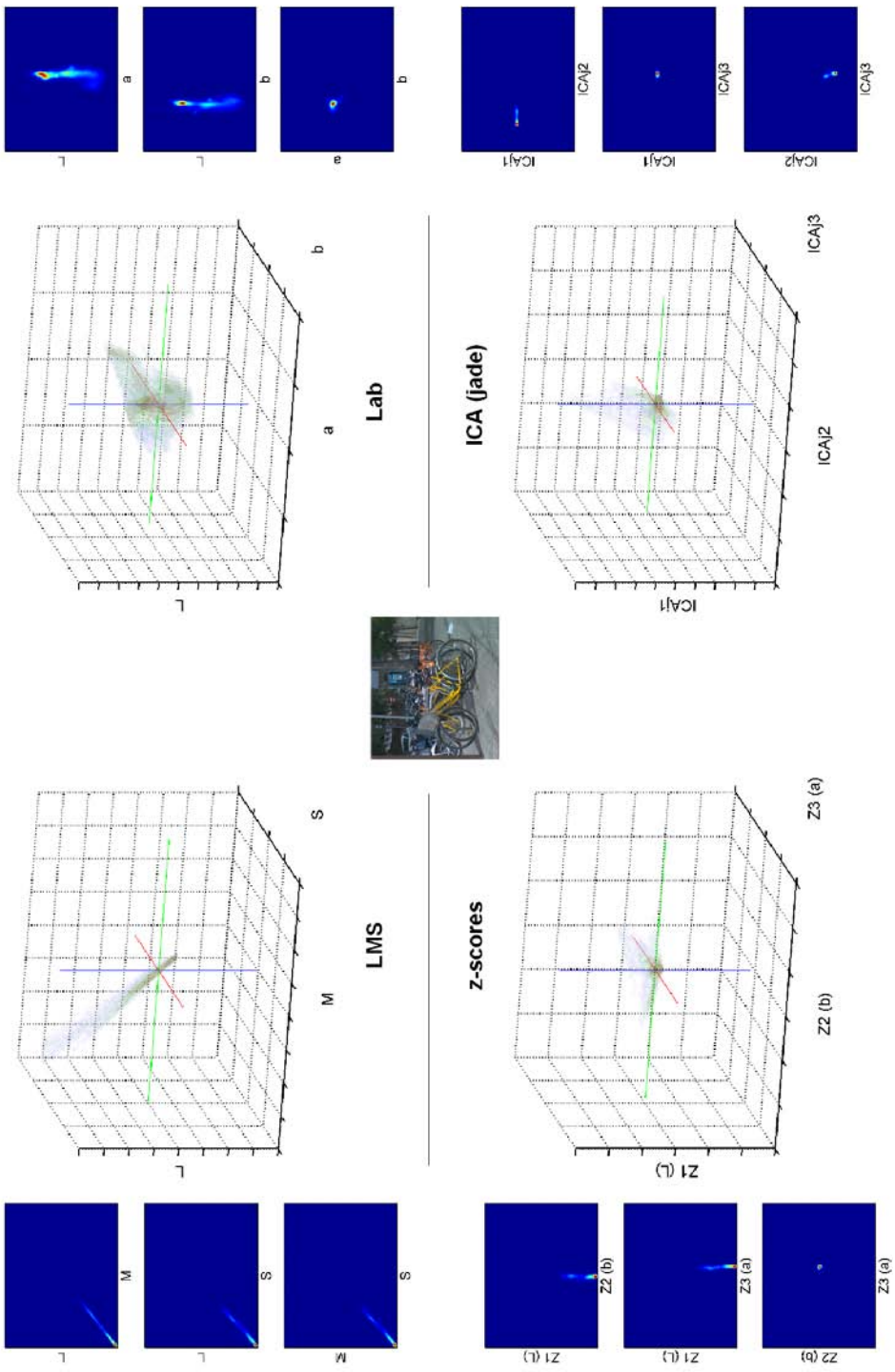

Figure 2.3: Example of color statistics in different representations (III).

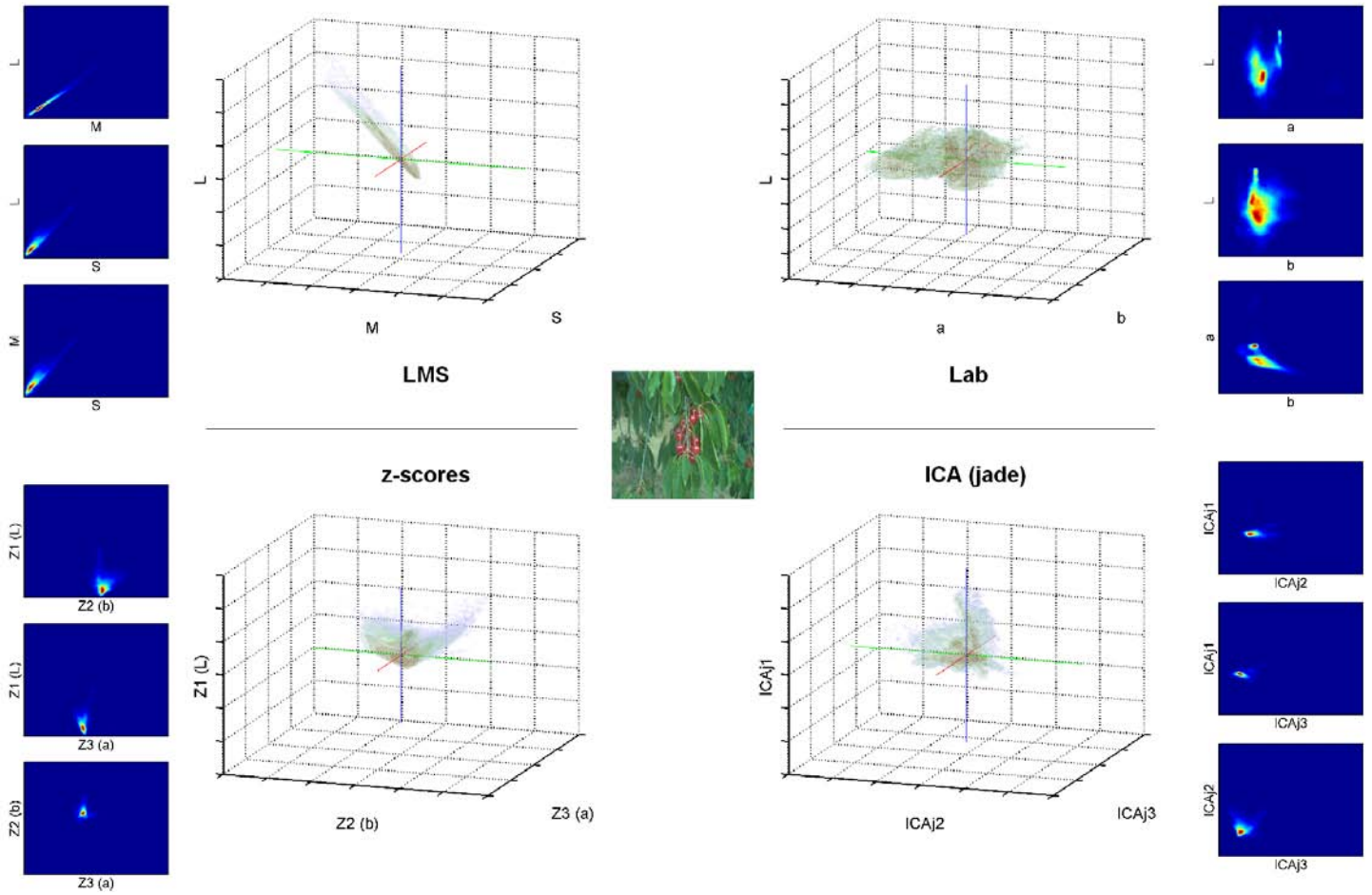

Figure 2.4: Example of color statistics in different representations (IV).

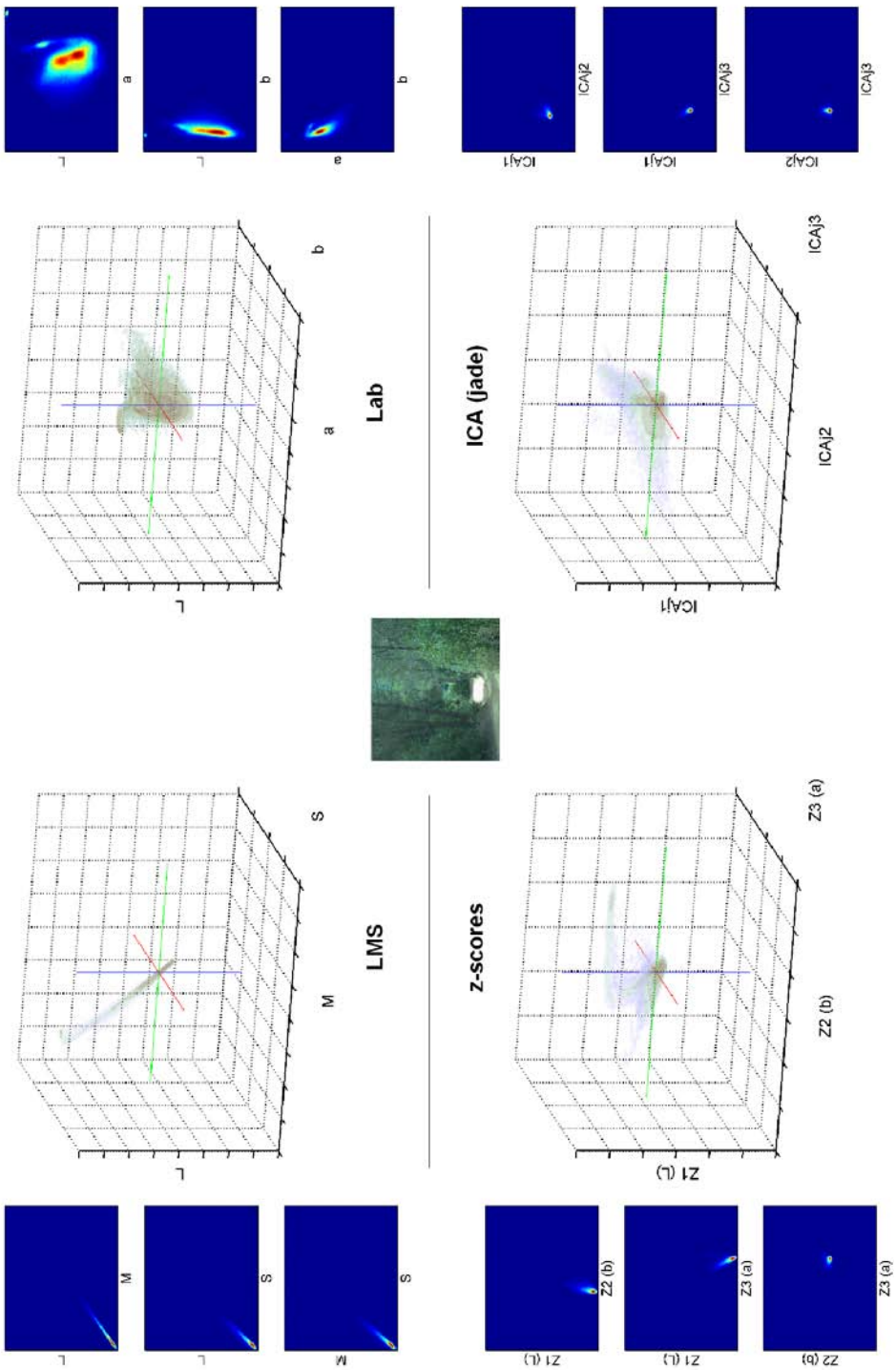

Figure 2.5: Example of color statistics in different representations (V).

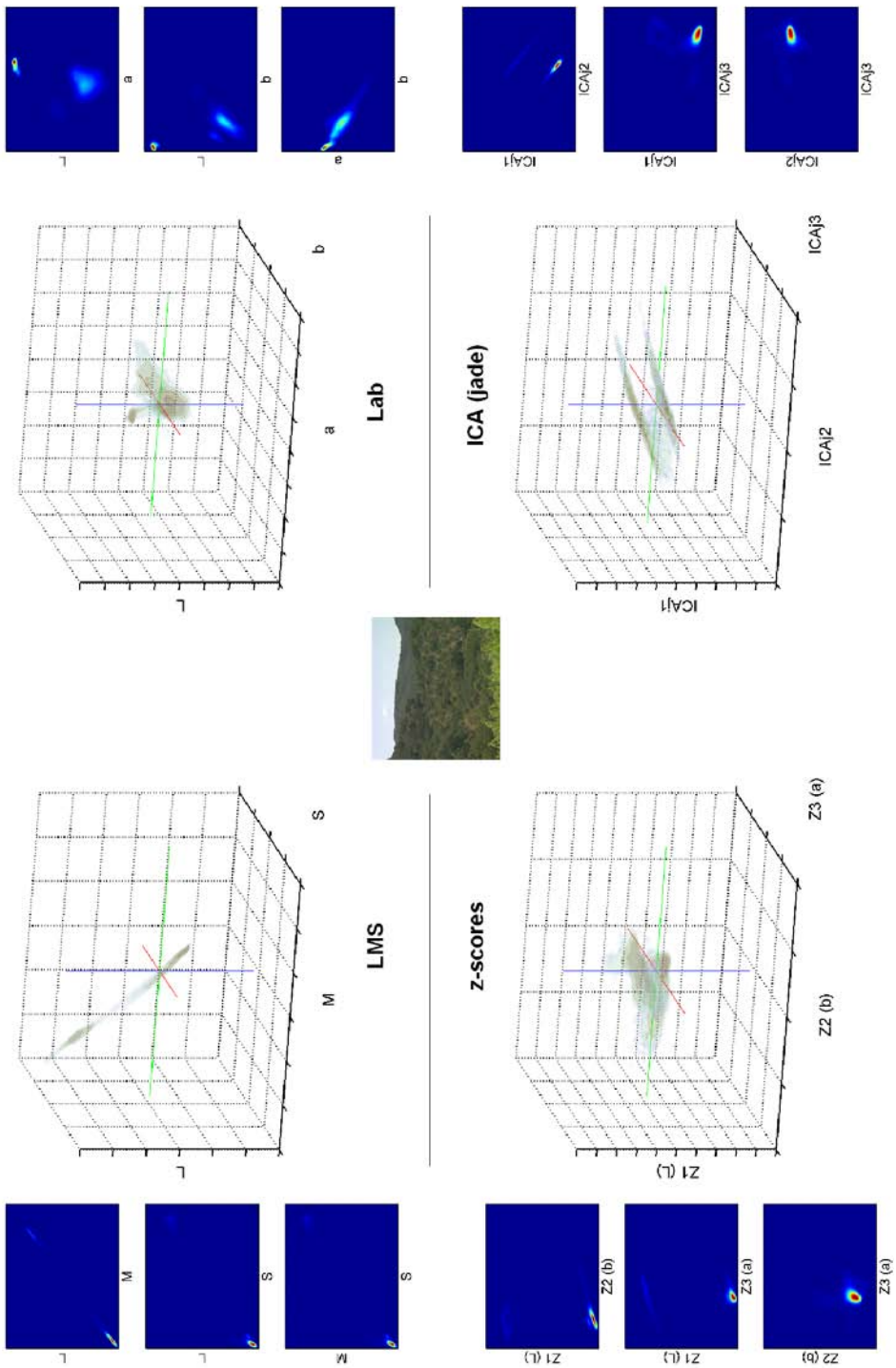

Figure 2.6: Example of color statistics in different representations (VI).

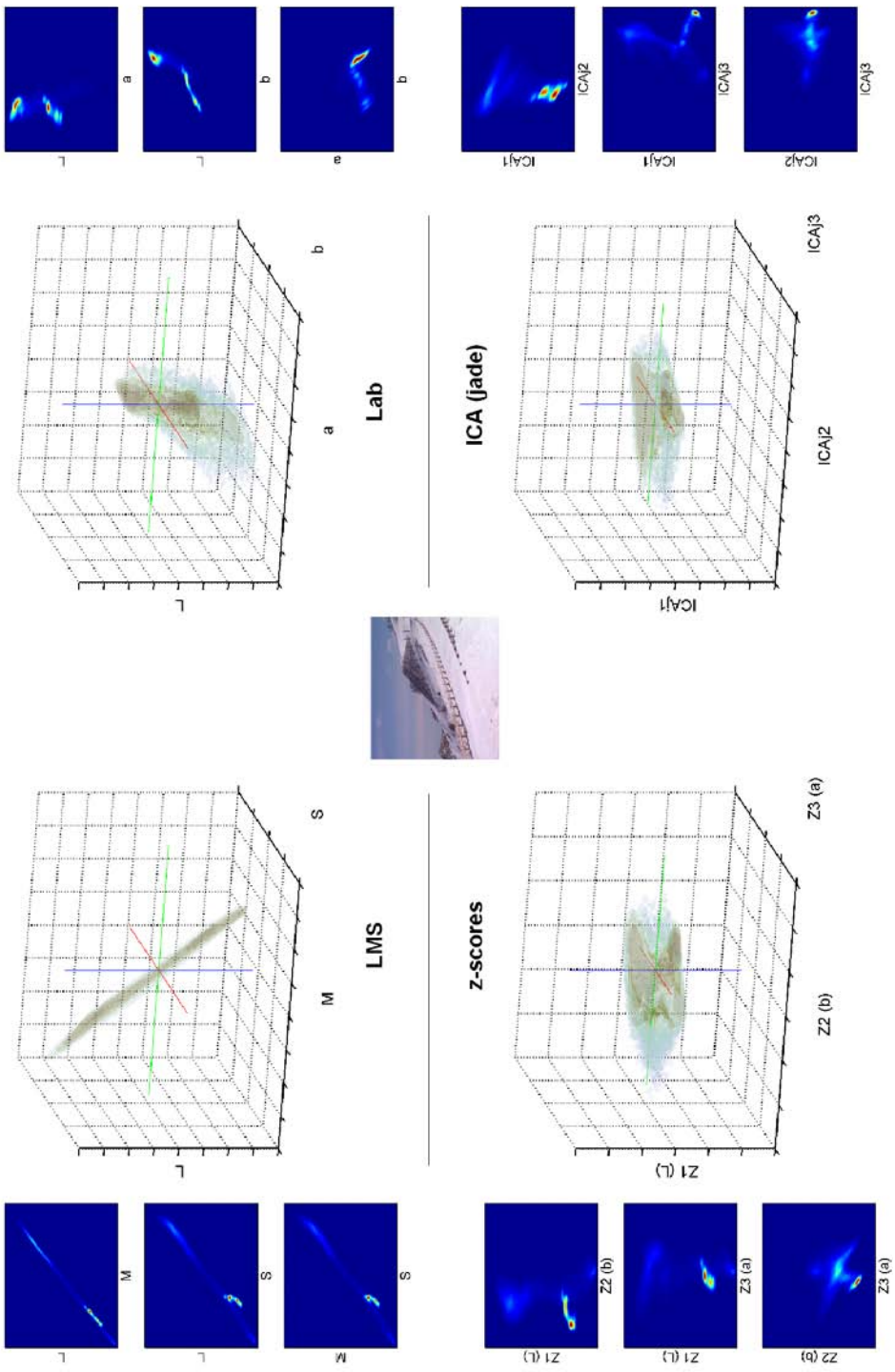

Figure 2.7: Example of color statistics in different representations (VII).

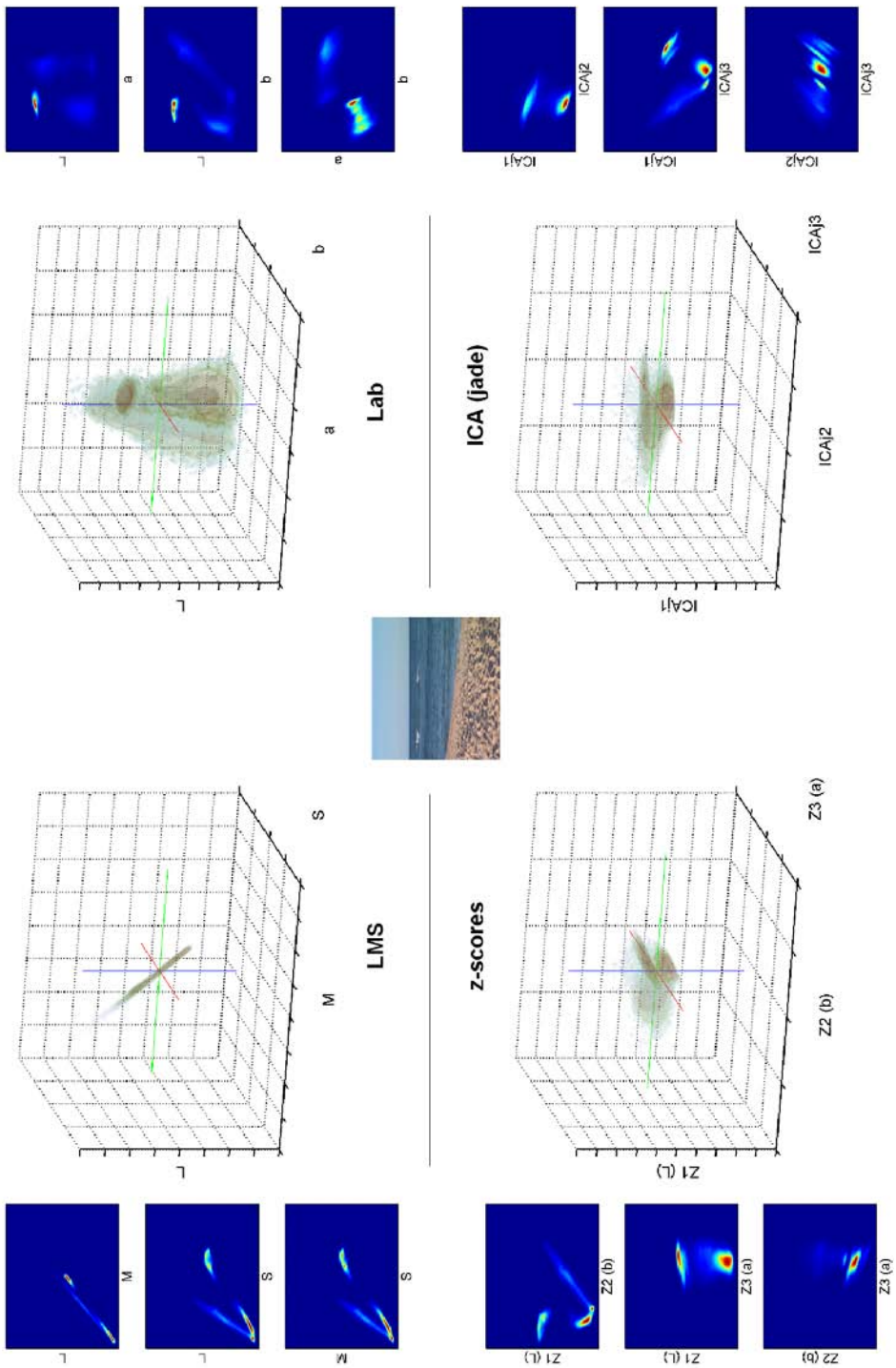

Figure 2.8: Example of color statistics in different representations (VIII).

tation like Lab removes much of the first order correlation. However, statistical distributions are still quite different from one image to another, as pointed by Webster and Mollon [WM97], thus making sense for short-term adaptation of the color representation to specific scenes. And importantly, whitened representations that suppress first order correlations are best matched to the particular statistical structure of the images, providing an improved representation for discriminative and comparative purposes, present in any visual function. To remove higher order correlations different ICA approaches are needed. An obvious benefit of whitening is a better use of the available dynamic range in the different components and a more suitable measure of distinctiveness as statistical distance.

It must be noticed that results of whitening the LMS components obtained from the characterisation of Stockman and Sharpe [SS00] also available in the used dataset, were equivalent and do not delivered any relevant difference for the sake of this discussion.

Once the statistics of the images in relevant representations have been examined, the next step deals with the analysis of the information that retains each of the components of each of the representations considered. As well, the suitability of the modulus in these representations to measure color distinctiveness is an important aspect to take into account. Of course such a measure of distinctiveness is useless in natural images, since it does not account for spatial structure. With the aim stated above, the figures 2.9 and 2.10 are provided.

In general it can be observed the known fact that a LMS representation is not suitable for detecting color distinctiveness and that the Lab color model do manage much better to catch it. Since color disctinctiveness can be considered as closely related to perceptual distance, and Lab color model has been conceived to account in average for perceptual distances, this fact is simply in agreement with known properties of these color spaces.

More interesting are the results provided by whitened representations. Since they span exactly the same space with the same norm (variance), both z-scores and independent components provide the same measure of color distinctiveness. However, z-scores show a remarkable closeness to Lab color components, so much so that in all the examples examined in the dataset, each of the z-scores components was closer to one different Lab component. To check this fact, we have taken into account that intensity is mainly composed by a summation of L+M responses, red-green opponency arises from the difference of L-M responses, and blue-yellow arises from the difference of S-0.5(L+M). Labeling each z-scores with L, a, or b by projecting its LMS composition in these component and checking which was the closest component, always delivered an arrangement of the three Lab components. This

was not true for independent components which often delivered two components that were closer to the same Lab component than to any other.

Moreover, in all the cases the first z-score component was related to L (intensity), the second component was related to b (blue-yellow opponency) and the third component was related to a (red-green opponency) . Regarding to intensity, this fact seams very reasonable, since the order of z-scores is inherited from the PCA before whitening through normalisation by variance, since principal components are ordered by decreasing value of variance, and since most of variance in a natural image is related to intensity, and only a small part of it is related to opponent color components. Again this interesting behavior was not observed for ICA components, with an order not related at all to variance.

Regarding to opponent components, the fact that a and b Lab components are always related respectively to the third and second z-score components has been quite surprising, since we have not found any reference in this sense in the literature. However, this behavior has not been observed when decorrelation is done from RGB components of uncalibrated images from a digital camera, instead of LMS components. In this case, a and b are approximated by R-G and B-0.5(R+G), while L is computed R+G+B. Again, the three z-scores correspond to different Lab components and the first z-score is related to intensity, but the second and third swap the correspondence with a and b from one image to another. Therefore, this question deserves further investigation to rule out possible biases in the LMS representation of the employed dataset.

From the observations pointed above, a natural justification for the hypothesis that z-scores approach better the coding of opponent components in the HVS, arise as a slight shift of opponent components from an average Lab representation, associated to contextual chromatic influences. Such a behavior would demand that the opponent components coded early in the HVS, which are in average decorrelated in the space of natural images, interact to match the statistics of a given image, providing a specifically decorrelated representation of the same. This interaction and adaptation, from a functional perspective, is biologically plausible as discussed in the previous sections.

2.4.2. Data-driven perceptual grouping and segregation and illusory contours from scale whitening

For each of the whitened color components -whenever existing-, an spatial decomposition in terms of oriented scales has been performed. This is in fact

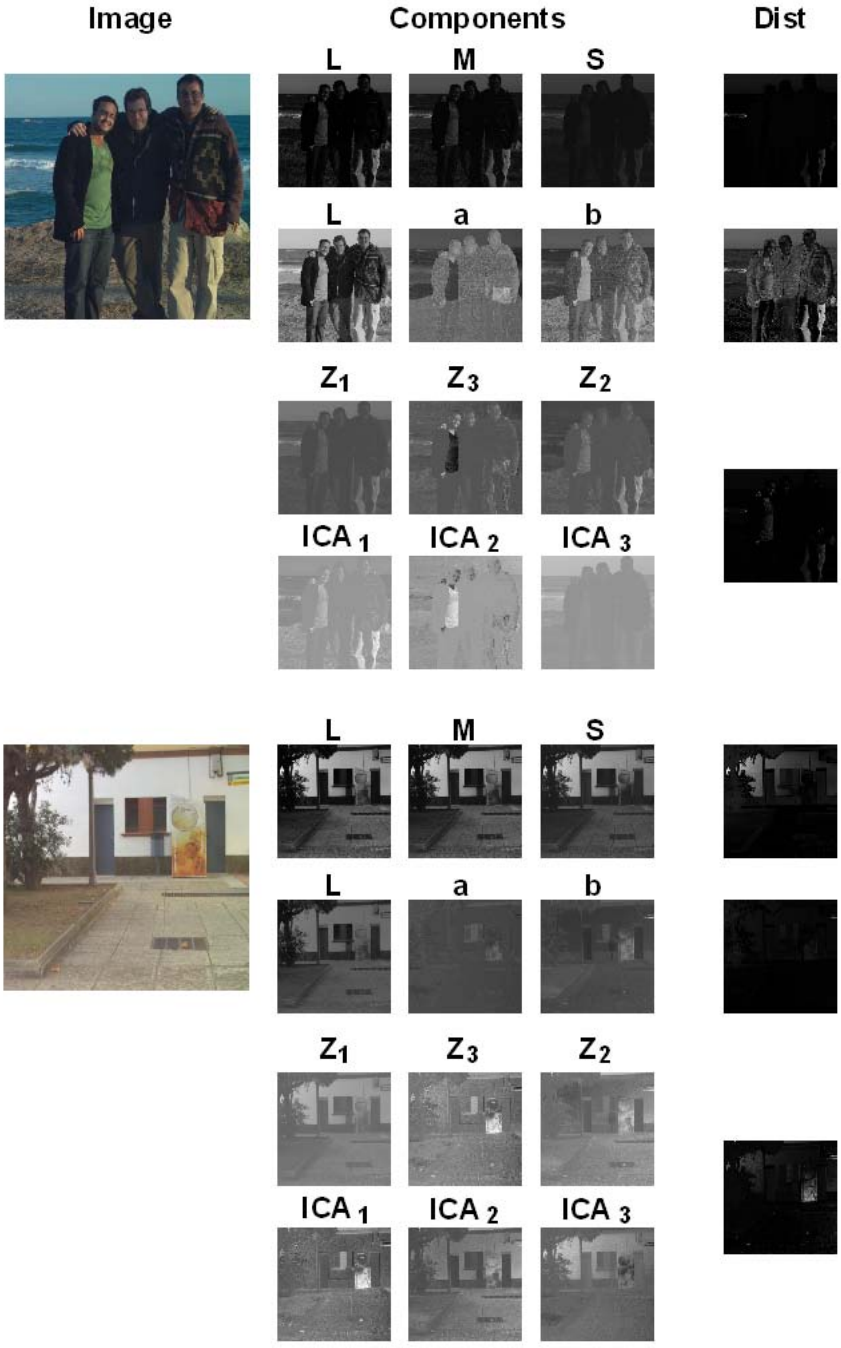

Figure 2.9: Image components in different color representations and the corresponding measures of distinctiveness from a squared euclidean distance (for images I and II)

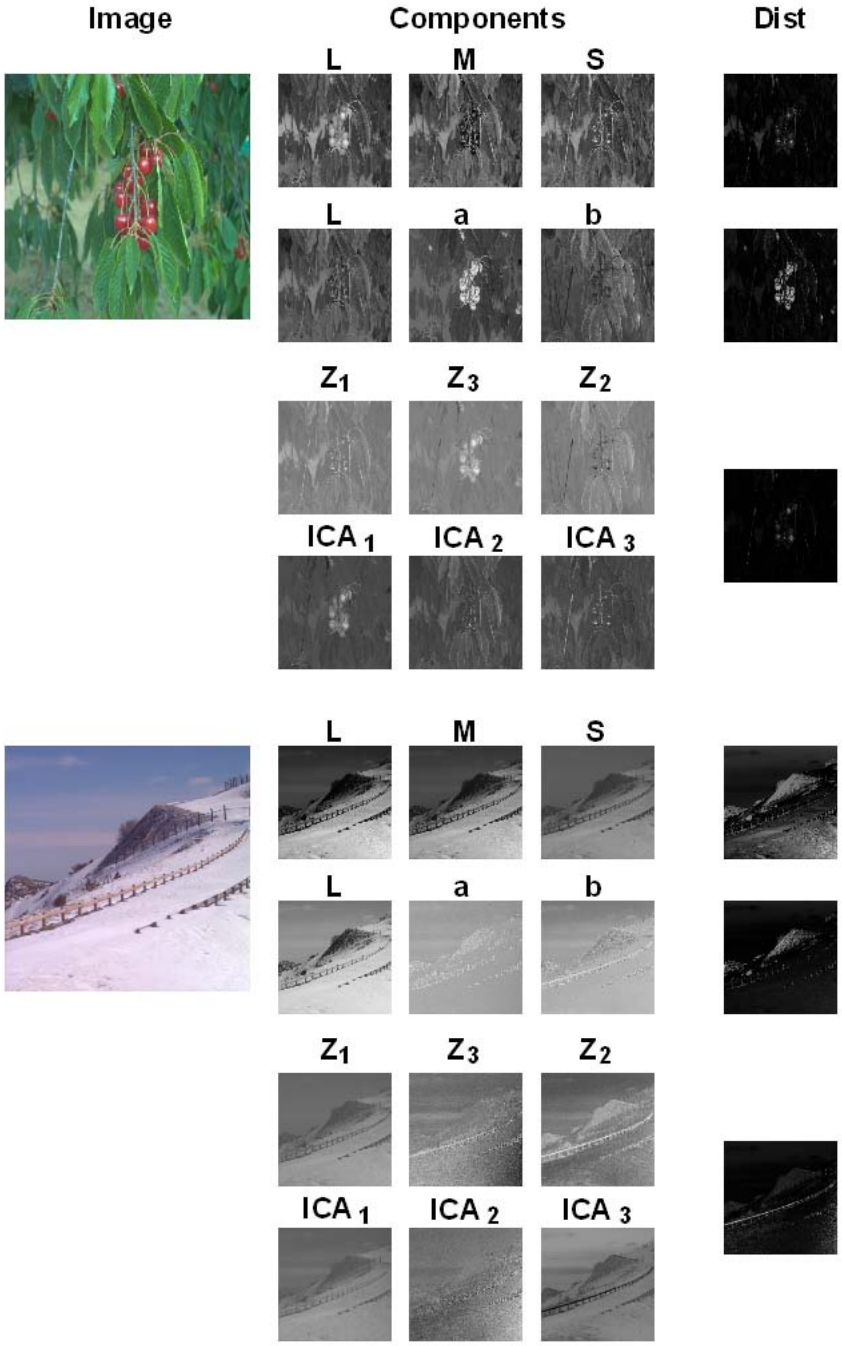

Figure 2.10: Image components in different color representations and the corresponding measures of distinctiveness from a squared euclidean distance (for images IV and VII)

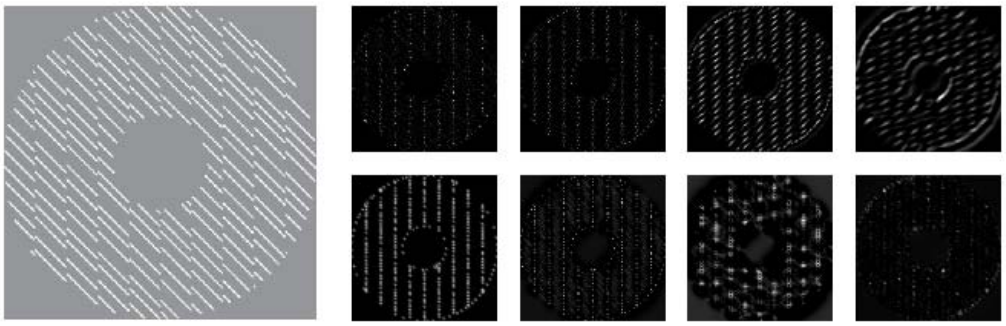

Figure 2.11: This typical psychophysical image that clearly generates vertical illusory contours has been adapted from [MLHL07]. Scale responses for a 45º orientation (top row) and the corresponding decorrelated scales that manage to catch the illusory contours (bottom rwo)

a decomposition of the image in terms of bands of spatial frequencies in the Fourier domain for each of the available whitened chromatic components. The decomposition has been acomplished using a bank of logGabor filters that are described and examined in detail in the next chapter. According to the simplest scheme of whitening, each group of oriented scales has been whitened following the same procedure than color components.

In this section, a qualitative comparison between responses to a given set of scales and the resulting whitened scales is tackled using representative and illustrative images. Figure 2.11 shows two examples of illusory contours, the first is an adaptation of an image used by [MLHL07] in fMRI experiments involving illusory contours, the second is a representative painting of so called Op-art (optical art) by Victor Vasarely. As can be seen, decorrelation of scales in the selected orientations catchs well the obvious illusory contours present in the image, while simple band-pass filtering is unable to do it.

Op-art is a style of painting and in general of visual art that relies on the use of visual illusions, concerned with the relation between understanding and seeing. It provides hence numerous examples of visual illusions that constitute an intermediate step between synthetic images most often used in psychophysical experiments and natural cluttered images. This feature makes this type of artistic creations very interesting for testing computational models of visual processing. Indeed, though not very frequent, examples of its use can be found in literature like, for instance, in the work by Troncoso et al. dealing with corner salience [TMMC05]. This work starts from Vasarely nested squares to create novel visual illusions that allow to quantify corner salience in a psychophysical experiment. Instead of squares, they propose to use nested stars and in the limit star-shaped gradients to study corner salience

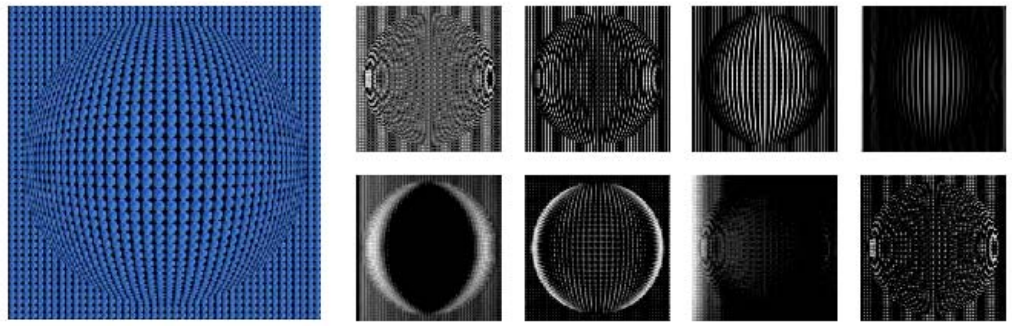

Figure 2.12: Reproduction of a circular illusory contour in a Vasarely's picture (left). Top row shows logGabor responses of four scales for the vertical orientation of the first z-score component of color (intensity), where no circular contour appears. Bottom row shows the result of whitening the previous set of scales, with a strong contour in the first and secon z-score scales

for different corner angles. This question will be taken up again in the chapter 5 when explaining the reported results in terms of the model of saliency proposed in this dissertation. Here we will focus in the fact that the corners of nested stars and the corners present in star-shaped gradients generate illusory contours that although not catched by simple bandpass oriented filters, are well captured however through the decorrelation of their responses. This is shown in figures 2.13 and 2.14 that are based on images adapted from [TMMC05]. Clearly, simple bandpass responses catch the contours of bright regions at different scales, while the whitened scales also capture very well the center lines of bright and dark regions, that is the illusory contours produced by corners.

Figure 2.15 shows results on Vega, another famous op-art work by Vasarely, for all the four orientations as well as for an isotropic set of bandpass filters. From left to right and from top to bottom, five blocks of two rows with four images are shown, coresponding respectively to 0^o, 45^o, 90^o, 135^o, and isotropic filters. Similarly to previous figures, the top row of each block shows the bandpass filters responses, while the bottom row shows the corresponding *whitened scales*. The results show well how scale whitening can produce an unsupervised and data-driven perceptual grouping and segregation for a given image, which provides suitable components allowing for quite direct figure-ground segregation. Figures 2.16 and 2.17 show similar results for two exemplars of the Zebra series and for several orientations. The zebras, the frame and the background are well captured by different whitened scales, while they were mixed in the corresponding bandpass responses.

In a further step of analysis, the figures 2.18 to 2.23 show results on nat-

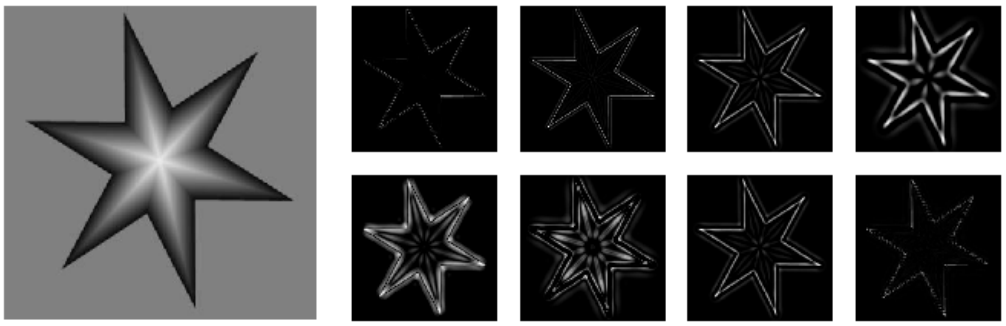

Figure 2.13: Reproduction of illusory contours on a star of grayscale gradients (left). Image adapted from the Martinez-Conde lab. Top row shows the responses to four scales of isotropic logarithmic Gaussians. Bottom row shows the results of whitening the previous scales that are able to detect all the illusory contours

ural images for different orientations of different whitened chromatic components, most corresponding to first component (intensity), but also some corresponding to the other adapted components (color opponent ones). Again, the facilitation of figure-ground segmentation in many whitened scales becomes apparent for natural images, with different scenes combining different foregrounds and backgrounds.

Finally, an interesting behavior of whitened scales in a given orientation is that they appear to allow for shifts in orientation selectivity. We can observe, for instance in the figure 2.15 how from responses tuned to horizontal features, the corresponding whitened components provide vertical lines, as well as horizontal lines from responses tuned to vertical spatial frequencies. Within the natural images we find a similar behavior, for instance a clear example is found for the diamond textures in the costumes of figure 2.19. From vertically tuned responses, whitened scales are able to catch separately different scales of the tilted structures produced by diamonds. Therefore, adaptive whitening of scales with orientation specificity can be seen as a possible contributor to the contextual adaption of orientation sensitivities.

2.4.3. Other levels of whitening

As advanced at the beginning of the section, three other levels of whitening have been considered: whitening of orientations using multiscale components, additional whitening of color components using multioriented and multiscale features, and joint whitening of orientations and scales.

In the best cases, the results did not improve the already described be-

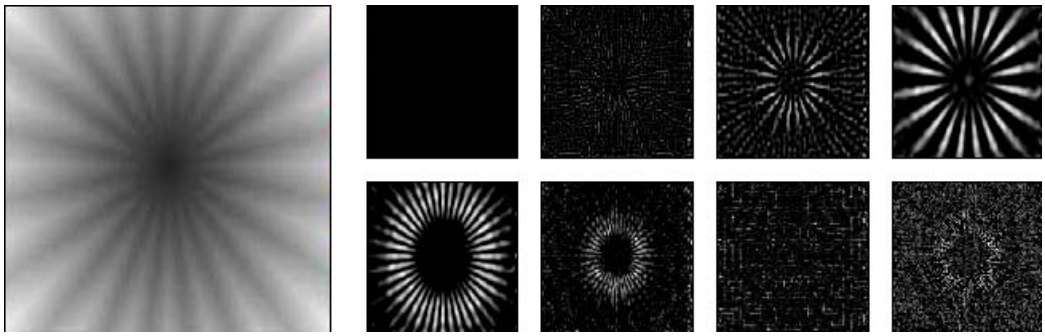

Figure 2.14: Similarly to figure 2.13 a star pattern of a gray scale gradient (left), but without the outline produces several illusory contours. Image adapted from the Martinez-Conde lab. Top row shows how the different scales of isotropic logarithmic Gaussians seem to be only sensitive to the dark fringes. However, bottom row shows how the corresponding whitened scales are able to catch all the illusory contours.

havior of the simple scheme of whitening of raw color components first and whitening of oriented scales next. No advantage was found in terms of capability to reproduce illusory contours or to reproduce perceptual grouping and segregation. An additional assessment has relied on the capability of these alternative adaptation schemes to provide an improved measure of saliency. In this regard, only whitening of color components using the responses to a Gabor-like bank of filters have shown marginally improved results in the prediction of human fixations (see chapter 4).

Therefore, the proposed scheme, under the assumptions of separate whitening of chromatic features and scale features with orientation specificity has been adopted in the following chapters as a suitable functional approximation to dominant mechanisms of contextual adaptation. Such a parsimonious approach is the ground for a definition of visual saliency and the specification of the corresponding model that is evaluated in a number of applications, to all of which the following chapters of this dissertation are devoted.

2.4.4. Mechanistic considerations

The proposal of concrete mechanistic schemes to implement the functional framework proposed above with biological plausibility is out of the scope of this dissertation. However it is worth pointing that there is a wide variety of approaches that would fit both the proposed whitening framework and the requirement of plausibility at a mechanistic level. Likewise, new mechanistic models with increasing performance and explanatory capability are being developed.

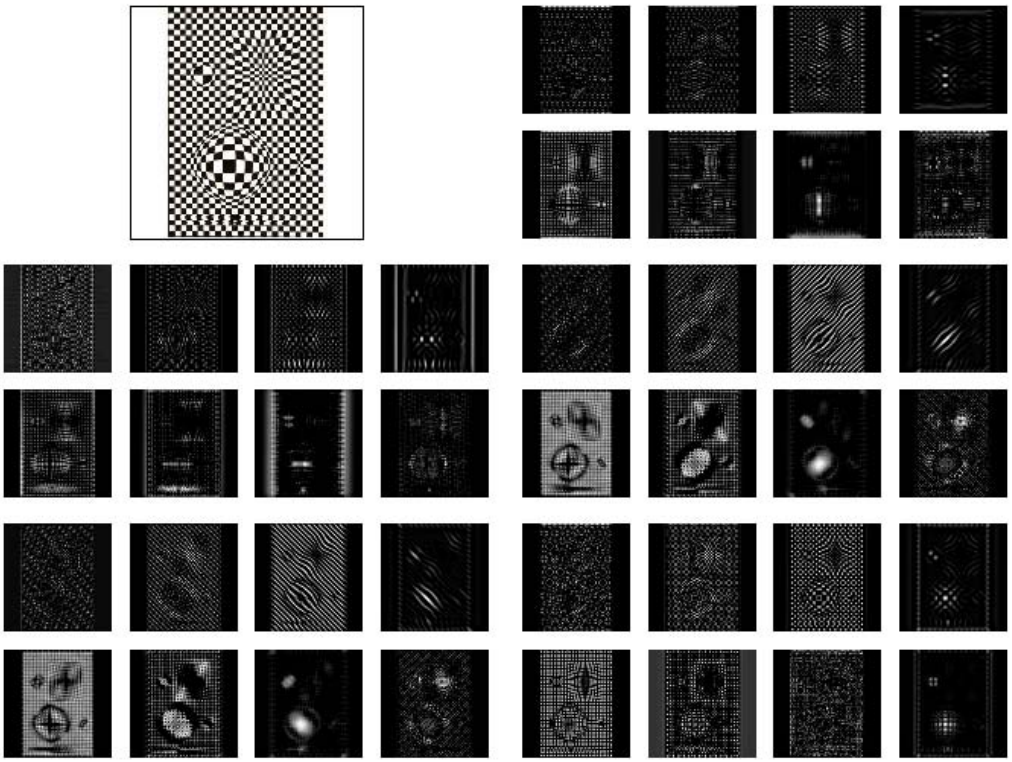

Figure 2.15: Examples of figure-ground separation through scale whitening in the Vasarely's *Vega* artwork (top left). Image. From top to bottom and left to right five blocks are shown corresponding to log Gabor responses for four orientations, respectively $0^o, 45^o, 90^o, 135^o$, and lastly to isotropic logarithmic Gaussian responses. In each block, top row contains the filters responses to four scales and bottom row the result of whitening each of these sets of scales. The geometric elements and the background are catched by different whitened components

At the lowest level, that of a neural network layer, several ways to implement computation of principal or independent components, and other whitening schemes exist [DK96, CA02, Hay09]. In general the different possible schemes involve lateral connections and/or feedback connections that produce tuning of different neurons to different decorrelated components. Many of these schemes are biologically plausible, since they are based in an Hebbian rule and hence they satisfy the requirement of locality. An early example of a mechanistic model to deal with color adaptation in cortex was proposed by Atick et al [ALR93]. It was indeed a neural network able to compute decorrelated and gain controlled components of input cone signals thanks to lateral feedback connections.

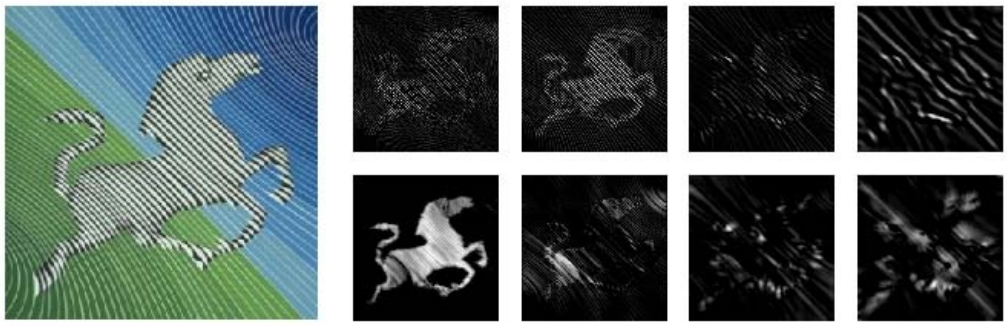

Figure 2.16: Example of figure-ground separation on an artwork from the Vasarely's *Zebra* series (left). Top row shows log Gabor responses for a 45° orientation. Bottom row shows the whitened scales, with the first component providing a proto-segmentation of the zebra

Recent powerful proposals seek to take advantage of non-linearities to produce ICA-like learning of inputs in models of neural cortical networks, yielding highly plausible computational schemes. In a remarkable example, Clopath et al. have tried to explain the connectivity patterns often found in cortex using a model of spike-timing-dependent plasticity (STDP). They found that different connectivity patterns could arise from different coding principles, and that rewiring the network can be very fast. Moreover, they showed that their model can be taken as an ICA model that is consistent with a large body of plasticity experiments [CBVG10]. Other outstanding work by Savin et al. proposes a model of intrinsic plasticity that regulates the firing rate of a neuron to guarantee sparse output. This approach yields an speeded up and robust model of ICA learning, based on a local rule. [SJT10]

At a higher level involving populations of neurons rather than few nearby ones, population codes support the use of bayessian schemes to model neural computations [ALP06]. In such schemes it is possible to formulate the computation of probabilistic PCA, ICA or Whitening [B$^+$06].

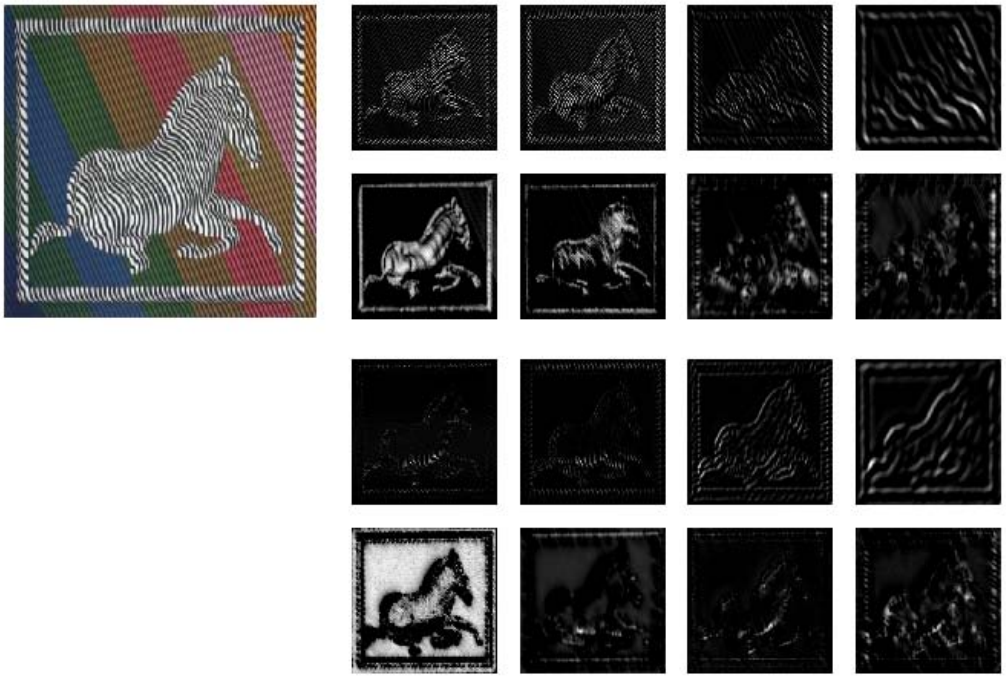

Figure 2.17: Another example based on an artwork from the Zebra series (left). Two blocks from top to bottom are shown corresponding to log Gabor filters at 45º and 135º. For each block, top row shows the responses while bottom row shows the whitened responses. Again the zebra, the frame, the zebra outline and the background are catched by different whitened components providing a suitable figure-gound separation.

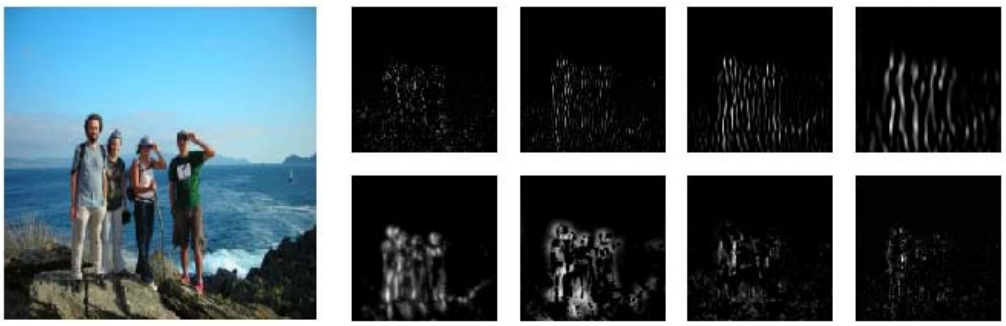

Figure 2.18: Example of figure-ground segregation on a natural image (left). Top row shows responses of log Gabor filters on the firs color component (intensity) and bottom row shows the whitened components where the group of people and its outline is catched in the first z-score component.

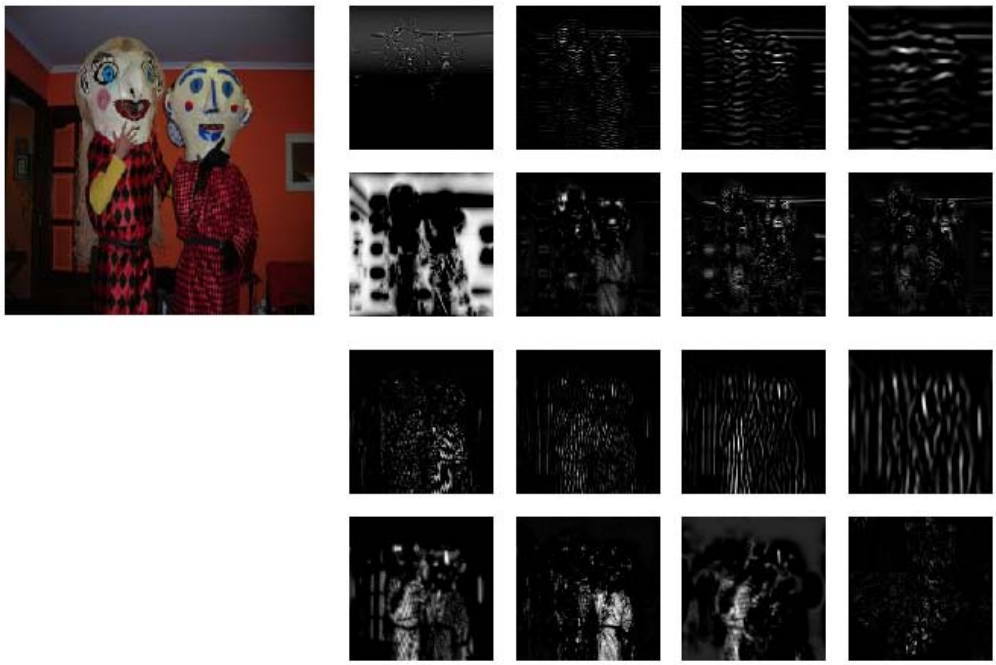

Figure 2.19: Example of spatial-chromatic whitening on a natural image (left). Two blocks of responses from top to bottom are shown corresponding respectively to the first and second z-scores of color. The first block shows the reponses to log Gabor filters oriented at 90º (top row) and the result of whitening these four scales (bottom row). The second block shows the responses to log Gabor filters oriented at 0º and the result of whitening them (bottom row). Different elements of figures and background are well cached by the whitened components.

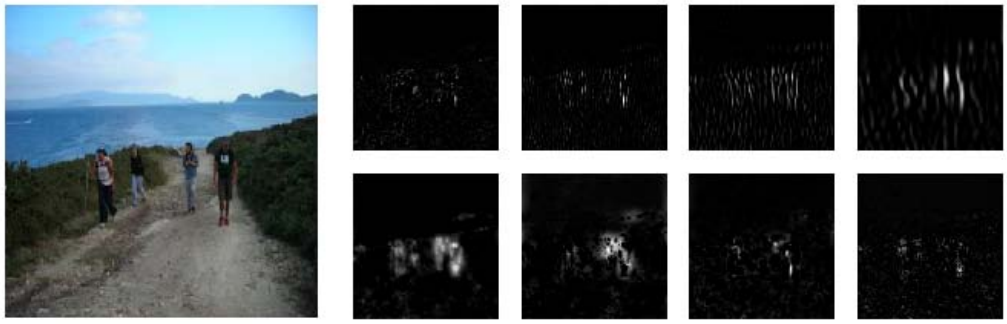

Figure 2.20: Example of figure-ground segregation on a natural image (left). Top row shows responses of vertical log Gabor filters on the firs color component (intensity) and bottom row shows the whitened components where the differen people and their outlines are catched in the first z-score component.

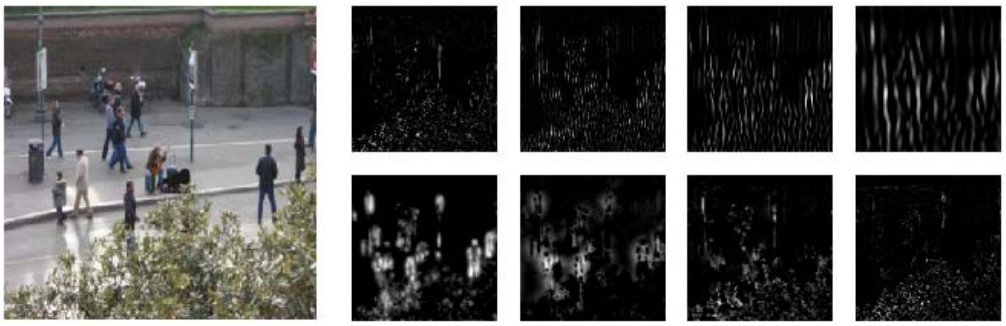

Figure 2.21: Example of figure-ground segregation on a natural image (left). Top row shows responses of vertical log Gabor filters on the firs color component (intensity) and bottom row shows the whitened components where the different people and urban object and their outlines are catched in the first and second z-score components.

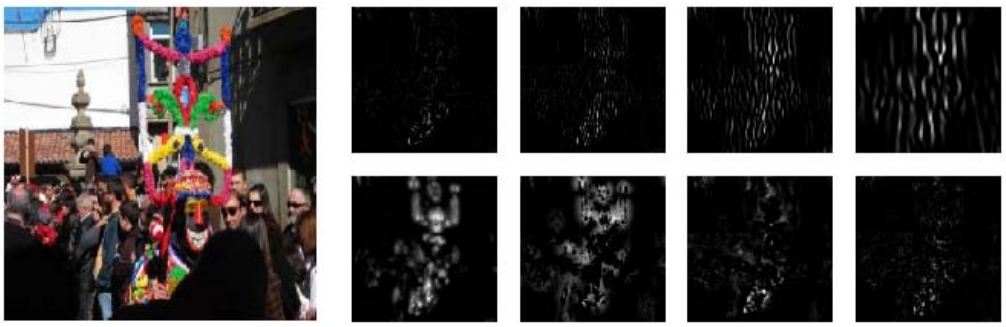

Figure 2.22: Example of spatial-chromatic whitening on a natural image (left). The top row shows the responses to log Gabor filters oriented at 0^o for the second z-score color component. The bottom row shows the result of whitening them. Different colored elements are well cached by the two first whitened components.

Figure 2.23: Example of spatial-chromatic whitening on a natural image (left). Three blocks of responses from top to bottom are shown corresponding the first tow to the intensity z-score and the last to the second z-score, correponding to a opponent component of color. Therefore, the first block shows the reponses to log Gabor filters oriented at 90º (top row) and the result of whitening these four scales (bottom row); The second block shows the responses to log Gabor filters oriented at 0º and the result of whitening them (bottom row); The third block shows the responses of the color opponent component to log Gabor filters oriented at 0º and the result of whitening them (bottom row). Different elements of the scene are well cached by the whitened components

Chapter 3

Optical Variability and Adaptive Whitening Saliency

In this chapter, a definition of saliency is proposed in terms of the optical variability present in a given image. This definition is shown to be closely related to the schemes for representational efficiency studied in the previous chapter. The modulus in these whitened schemes provides a suitable way to compute point distinctiveness as an estimation of point contribution to optical variability in the image. Therefore, this definition allows to interpret invariance to saliency of the visual system as invariance to cope with the optical variability in the surrounds. Subsequently, a simple and light implementation of saliency based on the adaptive whitening of low level features is described in detail.

3.1. Saliency as a measure of the optical variability in the visual window

3.1.1. Early coding, between optics and vision

A variety of impairments of vision has an optical nature, and optical performance of the eyes is closely related to visual performance. Indeed subjective measures of visual performance have shown remarkable correlations with optical objective measures of image quality. The need to correct visual impairments and to evaluate the effect of optical corrections of vision (both prosthetic or surgical) have led to the study and definition of different techniques and metrics to evaluate image quality. Some of them take into account not only the optical part but also incorporate a modulation component accounting for neural mechanisms. An illustrative example is the computa-

tion of the visual Strehl ratio that makes use of the optical transfer function (OTF) of the eye, weighted by a neural contrast sensitivity function, shortly the VSOTF. Its polychromatic version can be obtained from its value for the different wavelengths involved. The overall polychromatic value can be taken as the integral of its product by the visual sensitivity function [THBA04]. The VSOTF has recently shown a remarkable ability to predict visual performance in a subjective task, beyond other classical measures do [MTA04]. However the important lack of non-linearity and the rigidity of these measures makes them weak against changes in viewing conditions [AMT06], and of low value to gain insights on neural coding of images.

Since saliency attributes a single value to each point of an image, it seems of main interest to explictly ground its definition in simple optical magnitudes. Such a foundation would provide a direct formal link between meaningful physical magnitudes and a variety of psychophysical phenomena related to bottom-up visual attention. Moreover, since saliency does provide worthy insigths on early visual coding, and is probably closely related to it, such a definition would offer an additional bridge between optical description of images and the description of visual coding. In this section, saliency is defined in such a manner that naturally roots in a simple optical description of the image.

3.1.2. Optical variability

The concept of optical variability is usually found in the context of astronomical observations. Most frequently it designates different measures of combined variance of intensity and spectral composition of observed stars or other celestial bodies during time. Vision is concerned with images that have spectral -chromatic- characteristics, but that also have spatial structure as a main unavoidable feature. Thus, a definition of optical variability in space is needed. Since this dissertation is devoted to still images, the temporal dimension will not be tackled, but a very similar handling would be possible.

In Fourier optics any image can be considered as a wavefront piece and approached as a superposition of ideal monochromatic plane waves (see appendix). The local contribution to such a superposition of monochromatic plane waves can be described in terms of the spatial power distributions of chromatic components -related to electromagnetic wavelength- and of the corresponding power distributions of magnitude and orientation of the spatial frequencies present for each of them -related to the wave number vector (i.e. the direction of propagation of the plane wave).

Consequently, the contribution to optical variability by a given point from an image can be computed from the overall variability shown at that

point in the image plane by these magnitudes: spectral power and spatial frequencies power. To obtain the contribution of a particular sample from a set of samples to variability in a multidimensional space, one typical approach is to use a measure of generalized or statistical distance. It is given by the modulus -or a monotone function of it- of the vector associated to the sample in a decorrelated and whitened representation of the set of samples. It provides a measure of the distance to the center of the distribution in a system of orthogonal coordinates that have variance as the norm. Indeed, it is a measure of sample distinctiveness.

Therefore, we could think of each point as a sample with different components of luminous intensity corresponding to each combination of spectral wavelength and 2D spatial frequency. In a continuous domain the number of components would be infinite and the problem of whitening would be intractable. It is necessary to impose a discretization, considering a finite number of spectral wavelengths and spatial frequencies with a certain bandwidth. This means to assume the corresponding aproximations and change integrals by sums in the equations drawn in the appendix. Even so, the number of components can be too large for the typical whitening schemes that have a complexity cubic or higher against the number of components. Another way to reduce complexity consists in whitening separately chromatic and spatial components, as done in the previous chapter, and even only scales of each orientation.

This strategy has been observed to even improve capability of predicting fixations. Of course it is a particular definition of optical variability that assumes that enough reduction of redundancy is achieved by independently whitening chromatic and scale components. That is, independent whitening of components of spectral wavelength, and of components of the modulus of spatial frequencies for a number of orientations.

Formally, being M_c the number of discrete values of spectral wavelengths, W the whitening unmixing matrix, and λ_i^{white} a given whitened spectral wavelength the idea is to compute the transformation

$$I(\lambda_1^{white}, ..., \lambda_{M_c}^{white}) = WI(\lambda_1, ..., \lambda_{M_c}) \tag{3.1}$$

that is a coordinate transformation in the spectral domain. Besides, from equation 3.25, impossing discretization and omitting the point index variable, we have that

$$I = \sum_{i=1}^{M_c} I(\lambda_i) \tag{3.2}$$

while the squared norm in the whitened representation is the statistical dis-

tance or T^2 of Hottelling, that is

$$T^2_{chromatic} = \sum_{i=1}^{M_c} I^2(\lambda_i^{white}) = ||I(\lambda_1^{white}, ..., \lambda_{M_c}^{white})||^2 \tag{3.3}$$

which is in fact a multivariate measure of variance. Since the samples are the pixel values, each point has a T^2 value that gives its contribution to variance through the ensemble of samples. It is hence a measure of pixel contribution to variance of chromatic spectral components on the image plane.

It must be noticed however, that the relation 3.2 does not hold any more for the whitened spectral coordinates, that is

$$I \neq \sum_{i=1}^{M_c} I(\lambda_i^{white}) \tag{3.4}$$

Otherwise, the original monochromatic spectral components can be expressed by equation 3.26, which impossing discretization of spatial frequency coordinates and omitting again dependency against point index, remains

$$I(\lambda_i) = \sum_{\rho_0}^{\rho_1} \sum_{\alpha_0}^{\alpha_1} I(\lambda_i; \rho, \alpha) \tag{3.5}$$

As denoted in equation 3.1, the whitened spectral components are linear combinations of the original spectral components. Thus, a given *whitened spectral wavelength* is a linear combination of real spectral wavelengths. This means for whitened components that their composition of spatial frequencies is the corresponding combination of the compositions of the monochromatic components. As a result, an expression equivalent to equation 3.5 can be written for whitened components, that represents each of them as a combination of spatial frequency bands,

$$I(\lambda_i^{white}) = \sum_{\rho_0}^{\rho_1} \sum_{\alpha_0}^{\alpha_1} I(\lambda_i^{white}; \rho, \alpha) \tag{3.6}$$

Each of these representations of whitened components can be further whitened, using as original coordinates those of the spatial frequency bands. Instead of such an approach, the same simplification done in the previous chapter is adopted here. Whitening is proposed to be done for each set of spatial frequency bands at a given spatial frequency angle. That is being M_o the number of orientations -angles- of spatial frequencies, and M_s the number of scales (i.e. the number of values of the modulus of spatial frequency) for each whitened chromatic component λ_i^{white} and each angle of spatial frequency α_j

$$
\begin{aligned}
I^{\alpha_1}_{\lambda_1^{white}}(\rho_1^{white},...,\rho_{M_s}^{white}) &= W_1^1 I(\rho_1,...,\rho_{M_s}) \\
&\vdots \qquad\qquad \vdots \\
I^{\alpha_{M_o}}_{\lambda_1^{white}}(\rho_1^{white},...,\rho_{M_s}^{white}) &= W_1^{M_o} I(\rho_1,...,\rho_{M_s}) \\
&\vdots \qquad\qquad \vdots \\
I^{\alpha_j}_{\lambda_i^{white}}(\rho_1^{white},...,\rho_{M_s}^{white}) &= W_i^j I(\rho_1,...,\rho_{M_c}) \\
&\vdots \qquad\qquad \vdots \\
I^{\alpha_{M_o}}_{\lambda_{M_c}^{white}}(\rho_1^{white},...,\rho_{M_s}^{white}) &= W_{M_c}^{M_o} I(\rho_1,...,\rho_{M_s})
\end{aligned} \tag{3.7}
$$

As already pointed, a measure of saliency grounded on this approximation has been observed to not reduce the capability of predicting fixations or reproducing a number of psychophysical results. Indeed it has been observed to produce an slight improvement of performance in those tasks, in comparison with a measure derived from joint whitening of all spatial frequency bands.

The result is a representation of the image in whitened components respect to part of its coordinates. That is, the following overall transformation has been done:

$$
I = I(p_{xy};\lambda_1,...,\lambda_{M_c};\rho_1,...,\rho_{M_s};\alpha_1,...,\alpha_{M_o}) \tag{3.8}
$$

$$
\Downarrow
$$

$$
I = I(p_{xy};\lambda_1^{white},...,\lambda_{M_c}^{white};\rho_1^{white},...,\rho_{M_s}^{white};\alpha_1,...,\alpha_{M_o}) \tag{3.9}
$$

where point dependency has been made explicit again. From this partially whitened representation, optical variability (OV) is derived as the squared modulus

$$
OV = ||I(p_{xy};\lambda_1^{white},...,\lambda_{M_c}^{white};\rho_1^{white},...,\rho_{M_s}^{white};\alpha_1,...,\alpha_{M_o})||^2 \tag{3.10}
$$

Differently to the case of the starting color whitening, this modulus is not the T^2 of Hotelling but at most an approximation, arising from the summation of the T^2 obtained for different subsets of original coordinates. Then, it is a partial multivariate measure of variance that has indeed units of variance. To compute the real T^2, all components should be whitened jointly and the number of coordinates would be $M_c \times M_s \times M_o$. As mentioned above, the

complexity of whitening increases strongly with the number of components to the point to do such a computation too heavy or even unfeasible in practice. Besides, in the purpose of providing a measure of saliency, we have found that these approximations do not reduce its effectiveness in explaining visual behavior but they even contribute to increase it. It is worth remembering here that, as explained in chapter 2, these approximations are inspired in coarse features of the human visual system, namely the independent processing of color and spatial information as well as orientation specific decorrelation.

A simple characterization of an image closely related to its optical description in spatial frequencies, can be formulated in terms of local energy components at different scales and orientations -thus different values of modulus and angle of spatial frequencies- for different spectral components. The relation 3.2 is not true for a non-orthogonal wavelet decomposition, but can be taken as a reasonable approximation. Besides, the accuracy in that relation is not essential in our analysis, but the real importance relies on the reliability of the resulting whitened components. We have observed that the computation of whitening through PCA and ICA in our scheme, is barely affected by the overlaping between the original filters in the Fourier domain, a behavior expected for any blind signal decomposition scheme. The proposal of adaptive coding drawn in the previous chapter can then be applied directly in such a decomposition scheme. The only remarkable difference would involve the use of monochromatic spectral components rather than LMS or RGB chromatic components. Distinctiveness of a given point taken as a sample would be easily computed through the modulus in the whitened representation.

Going a step further, an additional coarse approximation would be to use the responses to broad spectral detectors rather than narrow spectral bands. For instance, RGB or LMS detectors. In this case, we can use exactly the same whitening schemes proposed in the previous chapter, and we can use the resulting modulus at each point in the image as a measure of relative variability or distinctiveness. This is the theoretical ground under the adaptive whitening saliency model, described in detail in the next section.

Otherwise, the implications of approximating chromatic whitening from broad overlapping detectors rather than from narrow quasi-monochromatic ones will be examined in more detail in the chapter 6, when dealing with hyperspectral images in the visible spectrum. There, results using narrow spectral components and responses to broad detectors will be compared and analysed. At a first glance, such approximation with LMS detectors can be understood as the computation of the variability existing in the *visual window*, that is in the optical window determined in part by the spectral sensitivities of the retinian detectors.

3.1.3. The optical visual window

The term window is usually employed to refer a given limited portion of the electromagnetic spectrum. For instance in fiber optics communications, different windows of transmission are available depending on the material of the fiber core and its absorption spectrum. It is also widely used to refer spatial limits in works in optics and computer vision. Hence, it is frequently used to denote limits in the transmission and reception of optical and visual information from a given domain.

Here the term is extrapolated to apply it to the *reception* of information from the environment by the brain, through the capture and representation of images using the visual system. Therefore, it refers the limited domain of optical magnitudes that the HVS -or any other visual system- is able to sense due to different factors. These limits and the discretizations and thresholds impossed to that magnitudes would constrain any visual transfer function.

If we think of saliency as an objective measure, resulting from the operation of an adaptive neuro-optical transfer function, then saliency must be the same for different subjects with the same visual window, when observing the same image.

Indeed, many of the approximations pointed above, related to broad sensitivities against chromatic wavelengths and spatial frequencies, but also to discretizations and to independent dimensions for whitening, can be seen as neural constraints acting on the definition of the optical visual window.

3.1.4. Invariance of saliency in bottom-up visual processing

A criticism to the efficient coding hypothesis relies in the fact that it *does not address why the coding catastrophe occurs, because it lacks specification as to the computational goal beyond representation; rather, it embraces it without further question* [SHD07].

From the previous definitions, a clear specification of the goal underlying representational efficiency and by extension the corresponding contribution to the coding catastrophe is derived: the invariance of bottom-up visual processing to cope with optical variability in the image. Saliency as a constrained measure of relative optical variability in the visual window is hence hypothesized as an invariant in biological visual systems.

Otherwise, the proposed invariance can be expected to apparently fail under two situations: artificial estimulation with statistically biased images, that will produce the corresponding artificial alteration of the visual window from long or mid term neural adaptation; and voluntary constraints on

sensed magnitudes through top-down bias, although it is not clear to which extent bottom-up and top-down representations are mixed or separated in the brain. In the absence of top-down motivations and biased estimulations, priority should be driven by saliency and thus by optical variability. As far as priority drives human behavior, like for instance eye movements, these must be invariant to relative optical variability. The exposed approach provides a simple and coherent ground to explain inter-subject consistency in the spatial distribution of fixations in terms of the efficient coding hypothesis: representational efficiency provides a suitable ground for distinctiveness computation. Since distinctiveness and improbability seem to be two sides of the same coin, a similar final interpretation can be provided in bayesian terms. Indeed, as pointed in the previous chapter, the proposed whitening of responses can be implemented in bayessian schemes.

To sum up, analysing separately bottom-up and top-down parts of visual processing, it is here proposed -regarding the bottom-up part- that it exhibits an invariance against a concrete estimation of the optical variability existing within the optical window of the visual sytem. This provides a useful and simple additional link between biological visual processing and a reduced set of physical magnitudes. Models of bottom-up visual processing aiming biological plausibility must accomplish with this requirement of overall invariance of saliency, enabling for its computation at some stage of processing.

An interesting prediction of the proposal drawn is that alterations of the visual window will affect saliency in the same way they affect optical variability. This holds for the different kinds of color blindness, for the different kinds of ocular impairments (astigmatism, myopia, etc.), and even for visual impairments of developmental nature like amblyopia. Thus, inter-subject consistency in the spatial distribution of fixations driven by saliency should suffer the same changes -if existing- for different subjects affected by equal visual impairments, or in general by equal alterations of the visual window. These last including alterations from long and mid term adaptation due to biased estimulation. This observation also raises some questions to which we have not found answers in litterature. Considering differences of visual windows arisen from different age, from different biological species, from differently biased estimulations or from different visual impairments, do they produce measurable systematic differences in the perception of saliency for the same image?. Do they produce different fixation patterns in free surveillance of images?. A comprehensive number of approaches to estimate saliency from fixations and visual behavior and to compare fixation patterns will be examined in the chapters 4 and 5, which could be used in trying to answer the posed questions.

Otherwise, alterations of top-down processing capabilities like those shown by subjects affected of visual agnosia should not have any effect in the perception of saliency. This last result has been indeed recently reported in a work comparing the spatial distribution of fixations in healthy and unhealthy subjects [MKH09].

The proposed principle can be extended to other portions of the electromagnetic spectrum and in general to any other physical-based representation (even not of electromagnetic nature) of the space to produce visual-like displays under the constraint of transforming the physical variability into perceptible visual saliency. That is, to project other physical windows on the visual window under the constraint of conservation of relative variability in the space. In the chapter 6 such extendability of the definition of saliency proposed, will be used for the proposal of a evaluation procedure of visualization techniques.

3.2. Preliminary approaches and experiments

Several approaches to compute saliency from an adaptive whitened representation of the input image have been considered in the development that led to this thesis. They all have in common the assumption of scale decorrelation as a key mechanism of adaptation that supports the computation of saliency. In a first approach, scale decorrelation has been combined with center-surround differences on a multioriented and multiscale representation of luminance. Also color features were used based on the raw definition of opponent components originally employed by Milanese [Mil93] and by Itti et al. [IKN98]. The guidelines for the design of this initial model were taken from the performance in the former visual search experiments used by Itti and Koch with their model of saliency. Therefore, a main concern was the obtaining of a highly sparse measure of saliency.

Initial experiments included reproduction of pop-out phenomena, detection of military vehicles in natural scenes, detection of traffic signs, detection of emergency triangles and a red can of coke in cluttered scenes using the open access datasets provided by the Itti's lab, except the dataset of military vehicles that has been published by Toet et al. [TBV01]. A winner take all detector with inhibition of return similar to that implemented by Itti et al. was used to succesively select detected targets. Some examples are shown in figures 3.1 and 3.2. In the figure 3.1, pop-out of color and orientation singletons is demonstrated using the same images previously employed by Itti and Koch [IK00], the graphs show the number of false fixations before the singleton is found against the number of distractors in the display. The

figure 3.2 shows results of (unguided) saliency-based visual serach of targets in natural images with several datasets previously employed in Itti et and in Itti and Koch [IKN98, IK00]. Besides, a number of results from psychophysical experiments were also reproduced. Examples of size pop-out, influence of heterogeneity of distractors and presence/absence asymmetry are shown in the figure 3.3. Further details on these psychophysical results and their importance are given in the chapter 5.

This initial approach provided a performance only slightly higher than using only center-surround differences. The analysis of the different results obtained allowed to conclude that center-surround DoG filters were too rigid, too destructive of surround activity and that were not able to provide a reliable graded mesure of saliency. Therefore its use was nearly constrained to the detection of strong pop-out phenomena. Moreover, they amplified any design bias up to make the result useless for a significant number of images. Indeed, these problems can be also observed for the measure of saliency of Itti et al. that uses rigid center-surround filters and the late version of Itti and Koch that reinforces the role of DoG filters, driving also the feature integration process.

They also revealed the risk of using visual search experiments for the assessment of saliency. This kind of evaluation, depending on the method used to select fixations can favor all or nothing strategies, inspite of providing a poorer and less stable and robust measure of saliency. Besides, depending on the type of *salient* target selected for validation, they can hide a variety of design biases. Such validation procedures are much more suitable for specific purpose models of detection or ROI selection than for a generic one, as the measure of bottom-up saliency is.

Otherwise, measures different of local energy were explored. Since local energy is the modulus of a complex response composed of a pair of filters in phase quadrature corresponding respectively to real and imaginary parts, phase sensitive multiorientation and multiscale decompositions were investigated. They were designed to catch different features like phase congruency or phase symmetry, starting from the schemes proposed by Kovesi [Kov96]. However, no improvement was found, comparing to the use of simple local energy.

Subsequently, in a preliminary version of the adaptive whitening model that will be described in the next section, it has been shown that scale decorrelation of local energy luminance features and multiscale color features in a Lab color model, without the use of rigid center-surround differences, achieves a series of results equivalent to other state-of-the-art models. This preliminary scheme is shown in figure 3.4.

Remarkably, when using a generic assesment procedure like the predic-

tion of human fixations, it clearly outperformed both versions of the classical model of Itti et al. and Itti and Koch that were based on different integration procedures of the responses to centre-surround differences. It also outperformed other state-of-the-art models in predicting fixations, and it already showed an outstanding ability to reproduce a variety of psychophysical phenomena. The resulting saliency maps were much more graded than in the initial approach, and design biases were much less amplified in the new scheme for feature integration. The overall effect was a better use of the available dynamic range for the measure of saliency. Besides, the use of a Lab representation –instead of the raw opponent components employed by Itti et al. and many other models– improved slightly the performance.

Another problem that has been also tackled is related to the integration or competition of different *preattentive* visual dimensions. Support for the biological plausibility of both a maxima and a summation strategies can be found in literature [KZ07, TZKM11]. The different schemes studied in our work allow to test both hypothesis. For instance, in the scale decorrelation model in figure 3.4 summation of the conspicuity maps can be replaced by maxima extraction at each point. Using capability of predicting human fixations as a guideline, we have found that the summation strategy performs slightly better. Consequently, this approach has been adopted.

This preliminary version showed however some problems with color handling. A detailed analysis of images where the model showed low performance in predicting fixations revealed that certain cases of evident color pop-out were not well captured by the model, which was based in a Lab representation of the image. Other models of the state-of-the-art, like the model of Bruce and Tsotsos and of course the model of Itti et al. were observed to suffer from the same problem. This behavior pointed to the need of a more flexible representation of color. It motivated a more in depth study on the possibilities of employing whitening on different low level features, and on the decomposition schemes of color. Many of the observations and results derived have been already exposed in the chapter 2 as well as in this chapter. This study also led to the simple and light model described in the next section, which makes use of whitening of chromatic and scale features. As well, an important effort in the selection and improvement of suitable assessment procedures has been done. The corresponding methods and results are provided in the chapters 4 and 5.

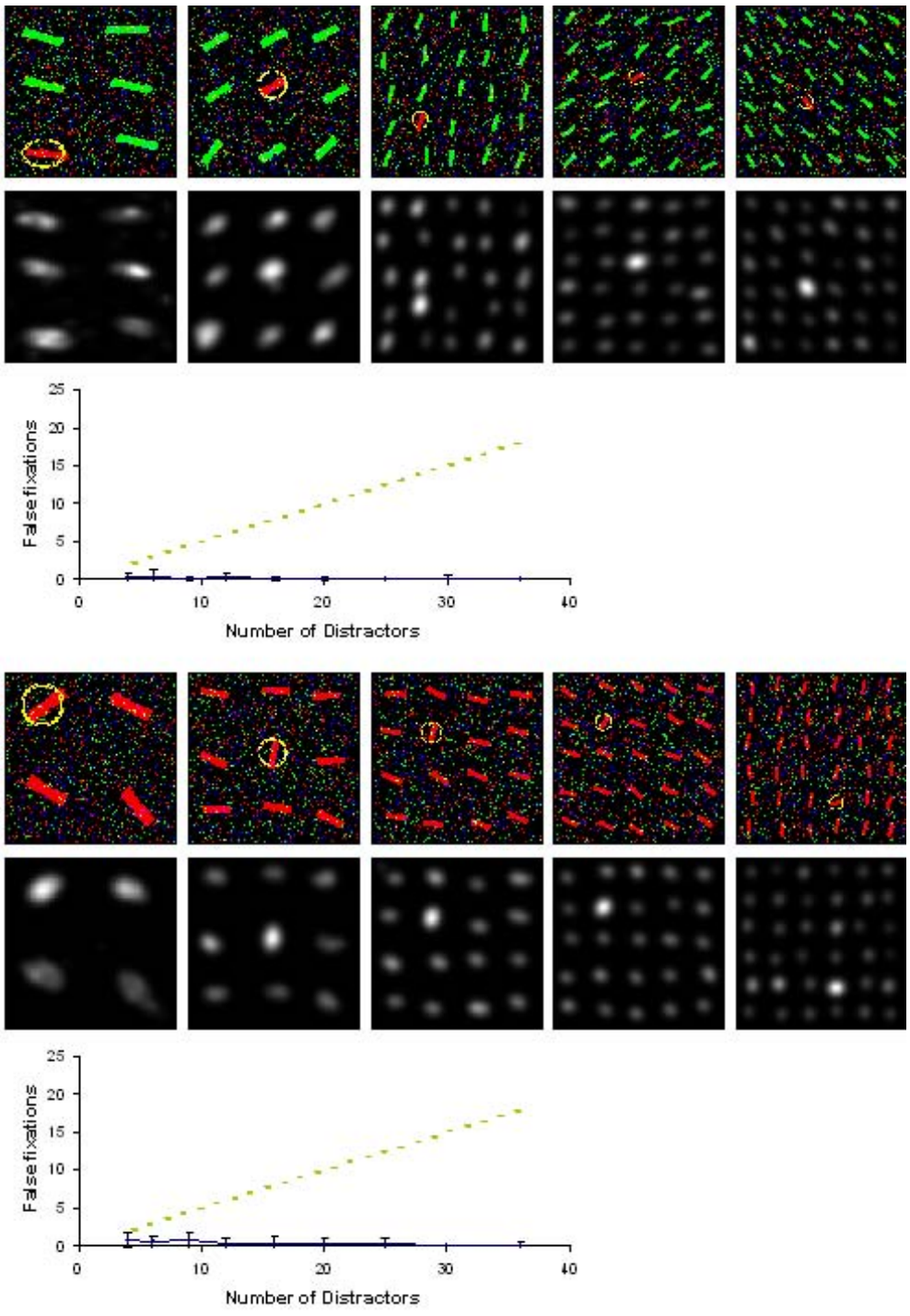

Figure 3.1: Initial experiments of reproduction of orientation and color pop-out combining deorrelation and center-surround filtering.

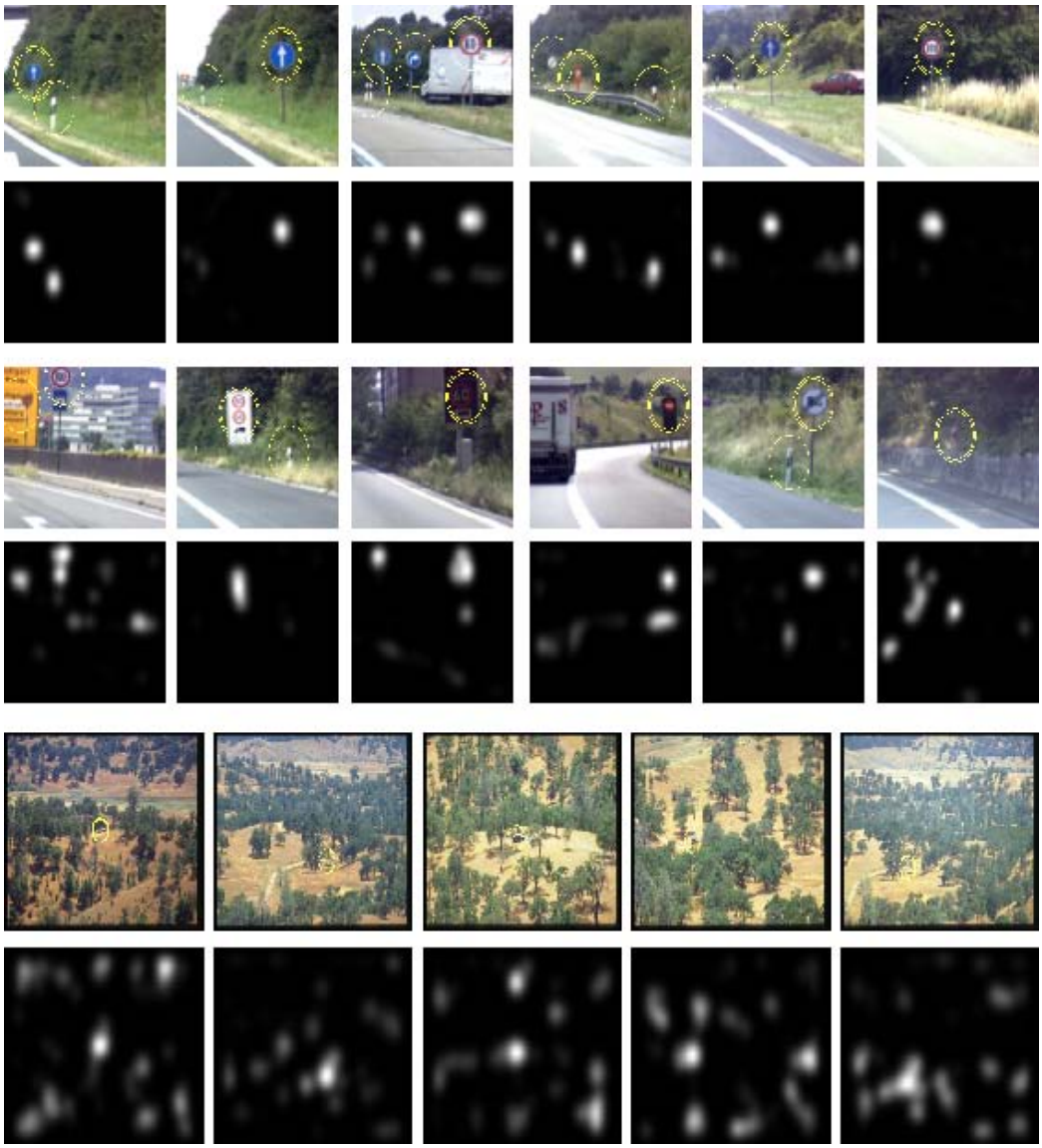

Figure 3.2: Initial experiments of visual search of concrete targets in cluttered scenes.

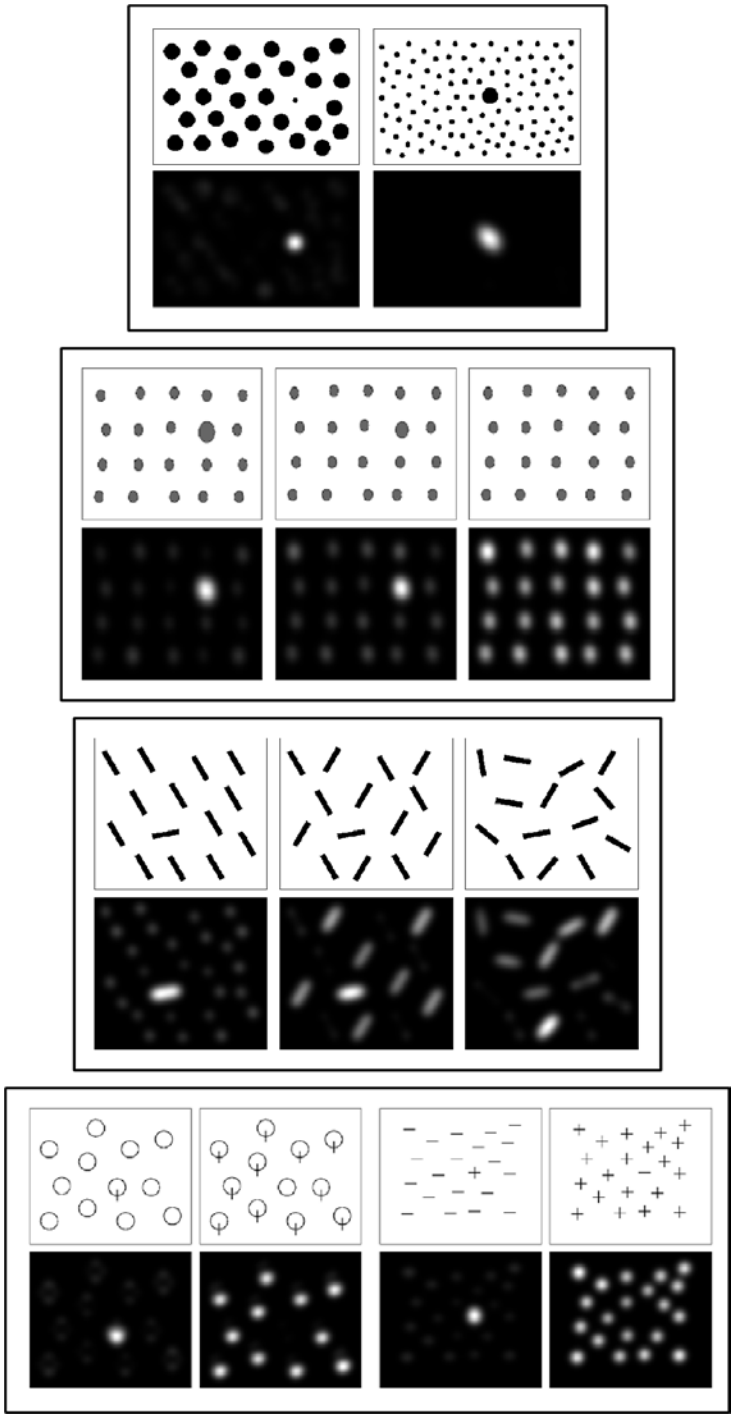

Figure 3.3: Initial experiments on the reproduction of psychophysical results.

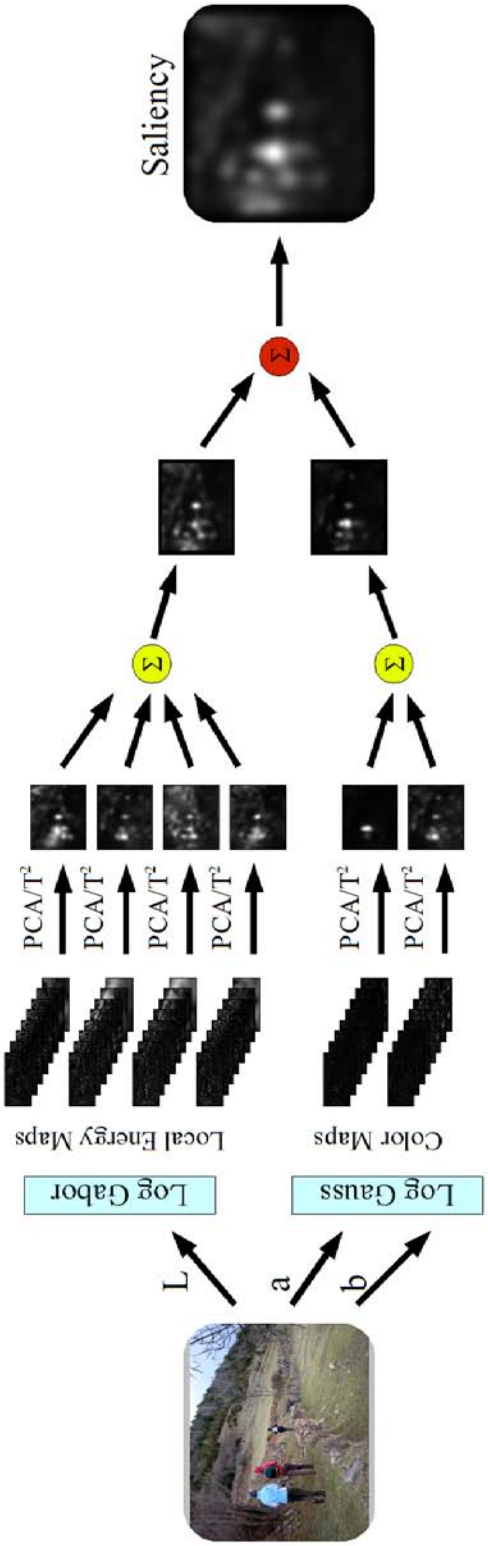

Figure 3.4: Preliminary version of the model of saliency based on the decorrelation of scales.

3.3. Description of the AWS model

In this dissertation, a simple proposal denoted by adaptive whitening saliency (AWS) is formulated to compute saliency that grounds on the whitened representation already introduced in the previous chapter. The scheme proposed here finds as well a simple foundation on the definition of optical variability drawn in the first section of this chapter.

Therefore, the starting point is the proposed adaptive whitening of multiscale features for the whole visual field and within a given orientation. The integration procedure follows the path inverse to decomposition through simple modulus computation in the whitened spaces, as well as addition of oriented conspicuities.

That is, after whitening of the chromatic features, each color component is subject to a multiscale and multiorientation decomposition, by means of a bank of band-pass filters. As previously pointed, orientation selectivity of chromatic multiscale receptive fields has been shown to take place in V1 and is thought to influence saliency [LH84, ZS06]. Then, for each color component and each orientation, the corresponding scale features are whitened. Under such process, local receptive fields are tuned to the same orientation and the same color whitened component in different positions and different scales. They interact to deliver a retinotopic representation with the same orientation and color selectivity, but with decorrelated scale information and with the variance as norm. These new receptive fields matched to whitened scales are biologically plausible, as shown in chapter 2. Indeed, they are very similar to classical receptive fields for many synthetic stimuli. To measure distinctiveness a simple squared modulus computation in the whitened feature space, for each oriented color component, is employed. The contributions of each orientation are combined through the summation of corresponding activities at each location to deliver the overall luminance and color opponent components conspicuities. The final saliency is the result of a further summation of these three chromatic conspicuities. To sum up, the AWS consists of whitening and multiscale multiorientation decomposition of colors, whitening of color oriented scales, squared modulus computation, and simple summation of location activities.

The model presented here provides a coherent approach to color and spatial short-term adaptation through adaptive whitening. It gives place to a flexible and non-parametric coding scheme, except for the design of the bank of filters, which has not shown to be crucial. Since the model is built from simple chromatic features, namely RGB, it needs neither the extraction and storage of statistical characteristics of a large set of (representative) natural images, nor the definition of functions, weights or normalization factors,

usual in color models. This fact possesses a very interesting advantage from a technical viewpoint: we can change the (R,G,B) sensors by any others with different spectral properties. This helps to directly apply the model on multispectral and hyperspectral images, as it will be shown in the chapter 6.

Next, a detailed description is presented of the implementation that will be used to perform almost all the experiments reported in chapters 4 to 6. The main particular choices are related to the whitening method applied in each case, and also to the initial decomposition of the image in color, scale and orientation components. In the figure 3.5 a flowchart summarizing the model implementation is shown.

3.3.1. Whitening procedure

Regarding color information, it has been observed that results are barely affected by the choice of the whitening procedure, by testing several approaches based on PCA and ICA [HO97, CSP93]. The results are totally independent of the whitening method employed with scale information, since they only differ in a rotation that will not alter the subsequent computation of modulus. Therefore, decorrelation is done through PCA, since it is a first order procedure that provides an ordered decomposition of the data. Its lower computational complexity is a clear advantage against higher order methods, like the diverse ICA algorithms. The benefits of an ordered representation will become apparent in the next section. Thus, the principal components are obtained, and then normalized by their variance. This last step delivers a whitened, and still ordered, representation.

Let $\mathbf{x}$ be the representation of the image in the original –color or scale– space, $\mathbf{y}$ the corresponding representation in principal components, and $\mathbf{z}$ (z-scores) the corresponding representation in the whitened coordinates. That is,

$$\mathbf{x} = (x_{ji}) \rightarrow \mathbf{y} = (y_{ji}) \rightarrow \mathbf{z} = (z_{ji}) \tag{3.11}$$

with $j = 1...M$ and $i = 1...N$, where M is the number of components, and N is the number of pixels.

The whitening procedure can be summarized in three steps. First, as well known, principal components result from diagonalization of the covariance matrix, ordering eigenvalues from higher to lower. Hence, if $\mathbf{C}$ is the covariance matrix, principal components $\mathbf{y}$ are obtained from the transformation matrix $\mathbf{U}$ and are ordered using the eigenvalues (l_j) as follows:

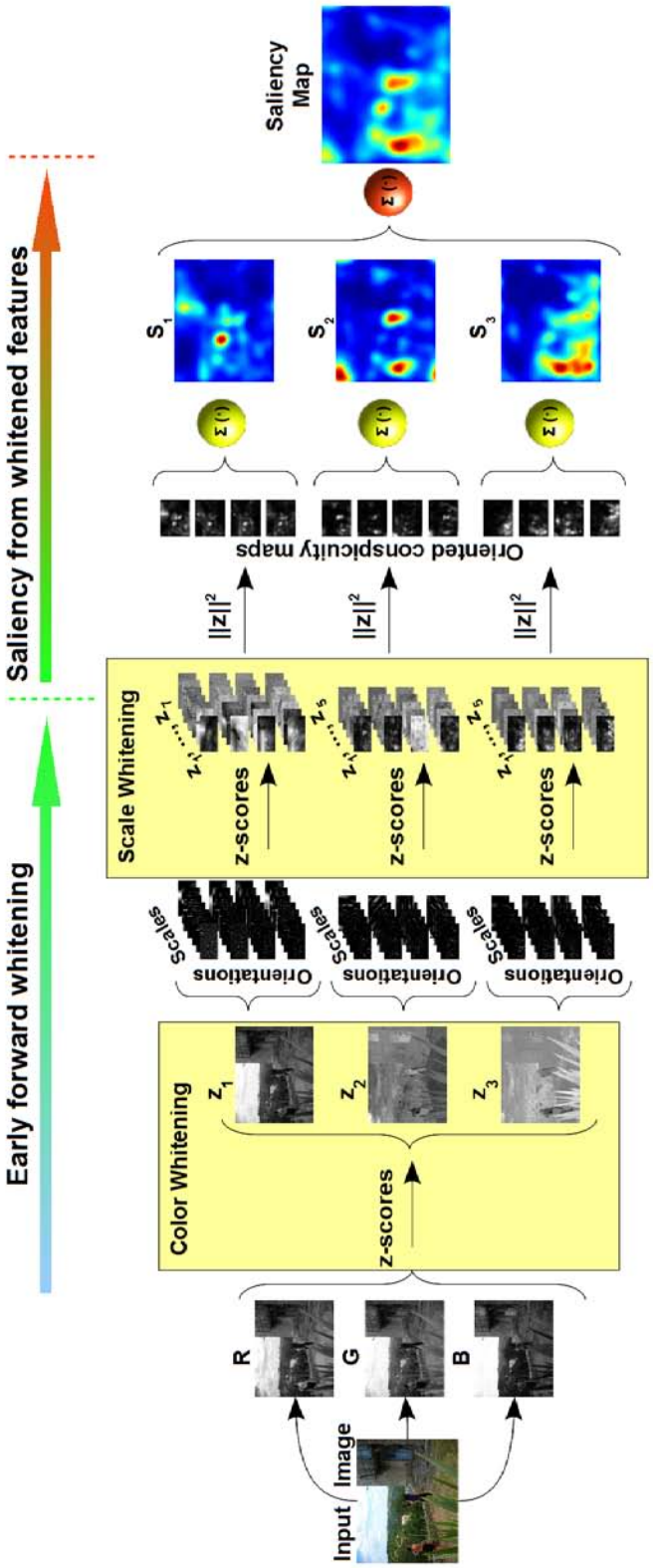

Figure 3.5: Adaptive whitening saliency model.

$$|\mathbf{C} - l_j\mathbf{I}| = 0 \; ; \; l_j \geq l_{j+1} \rightarrow L = \begin{bmatrix} l_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & l_M \end{bmatrix} \rightarrow \mathbf{U^T C U} = \mathbf{L}$$

$$\mathbf{y} = \mathbf{U^T} \left| \mathbf{x} - \bar{\mathbf{x}} \right| \tag{3.12}$$

The whitened $\mathbf{z}$ representation is then obtained through normalization by variance, given by the eigenvalues. This means that for each pixel and principal component:

$$z_{ij} = \frac{y_{ij}}{\sqrt{l_j}} \; ; \; i \in [1, N]; j \in [1, M] \tag{3.13}$$

These z-scores yield a whitened representation, and the squared modulus of a vector in these coordinates is in fact the statistical distance in the original $\mathbf{x}$ coordinates.

3.3.2. Measure of saliency

To decorrelate color information, the described whitening procedure is simply applied to the R, G, B components of the image. This whitening strategy has been also tested on other color models different from raw RGB, like Lab or HSV for instance, but the results were not so good, so they were discarded. The ordered nature of principal components is used to distinguish between intensity and color information, since luminance corresponds, in natural images, to the first principal component, that shows the maximum variance –usually higher than 90%. In turn, the second and third components correspond to typical opponent components. For other procedures, like ICA, this differentiation can be done by looking at the eigenvectors of the transformation matrix. In case of luminance, all R, G and B components contribute constructively, so that the three components of the eigenvector must have the same sign.

Once color information is whitened, each color component ($\mathbf{z_c}$) is decomposed with a multiscale and multioriented bank of filters. Log Gabor filters, which better fit the receptive fields of cortical cells are chosen [Fie87]. These filters only have analytical expression in the frequency domain, given by

$$\begin{aligned} \log Gabor_{so}(\rho, \alpha) = & \exp\left(-\frac{\left(\log\left(\rho/\rho_s\right)\right)^2}{2\left(\log\left(\sigma_{\rho s}/\rho_s\right)\right)^2} \right) \cdot \\ & \exp\left(-\frac{\left(\alpha - \alpha_o\right)^2}{2\left(\sigma_{\alpha o}\right)^2} \right) \end{aligned} \tag{3.14}$$

where (ρ, α) are polar frequency coordinates and (ρ_s, α_o) is the central frequency of the filter, s is the scale index and o is the orientation index. One of the advantages of the log Gabor filters is that they have zero DC component and zero value for negative frequencies, unlike the Gabor filters. Besides, their long tail towards high frequencies yields a more localized response. The impulse response is a complex valued function, with components being a couple of functions in phase quadrature, f and h. The modulus of the complex response of this filter is in fact a measure of the local energy ($\mathbf{e}$) of a frequency band for the color component c, with scale (s) and orientation (o) given by $(\rho_s, \alpha_o, \sigma_{\rho s}, \sigma_{\alpha o})$ [Kov96, MB88]:

$$\mathbf{e}_{cso} = \sqrt{(\mathbf{z}_c * f_{so})^2 + (\mathbf{z}_c * h_{so})^2} \tag{3.15}$$

In our bank of band-pass filters, four orientations $(0°, 45°, 90°, 135°)$ are used, seven scales for luminance, and only 5 scales for each of the opponent color components. This difference is justified by the observation that the finest and coarsest scales of color components barely showed any relevant information. Accordingly, while the minimum wavelength for luminance is 3 pixels, 6 pixels for color have been used instead. The use of orientations in color components has been observed to improve performance, compared to the use of isotropic responses, in agreement with a variety of experimental observations that show its existence in the HVS [LH84]. It has been also tried to include isotropic responses to luminance in addition to the oriented responses, but the results were practically the same. Consequently, they were considered redundant in the computation of saliency, and discarded for efficiency reasons.

The described whitening transformation is applied on the scales of each orientation. From the resulting components, the statistical distance is calculated between the feature vector associated to each point in the image to the average feature vector of the global scene, by simply computing the squared modulus:

$$\|\mathbf{z}_{ico}\|^2 = \mathbf{z}_{ico}^T \mathbf{z}_{ico} \tag{3.16}$$

This provides a retinotopic measure of the local feature contrast. In this way, a measure of conspicuity is obtained for each orientation of each of the color components. The next steps involve a Gaussian smoothing and the addition of the maps corresponding to all of the orientations. That is, for a given color component $c = 1...M_c$ and pixel i, the corresponding saliency (S_{ic}) is calculated:

$$S_{ic} = \sum_{o=1}^{M_o} \|\mathbf{z}_{ico}\|^2 \tag{3.17}$$

Color components undergo the same summation step to get a final map of saliency. Additionally, to ease interpretation of this map as probability to receive attention, it is normalized by the integral of the saliency in the image domain, i.e. the total population activity. Hence, saliency of a pixel i (S_i) is given by:

$$S_i = \frac{\sum_{c=1}^{M_c} S_{ic}}{\sum_{i=1}^{N}\sum_{c=1}^{M_c} S_{ic}} \tag{3.18}$$

Regarding the computational complexity of this implementation, PCA implies a load that linearly grows with the number of pixels (N), and in a cubic manner with the number of components (M), specifically, $O(M^3+M^2N)$. Several approaches can be used to reduce this complexity in relation to the number of components. Since the number of components (color channels and scales) remains constant and it is low, the asymptotic complexity depends on the number of pixels. This is determined by the use of the FFT in the filtering process, which is $O(Nlog(N))$. Most saliency models have a complexity which is $O(N^2)$ or higher.

3.4. AWS versus existing measures of saliency

The adaptive whitening approach proposed for computation of saliency, provides a unified framework suitable to explain the results of different previous approaches in terms of their capability to measure optical variability in a given scene.

As pointed in the previous chapters, most models of saliency ultimately resort to the same theoretical foundation: a suitable and plausible estimation of the inverse of the probability density for a given set of low level magnitudes.

Models that seek a close to uncommitted set of low level features, use a representation of the image in terms of independent components of natural image patches. Differences in this group of models are found related to the method to compute the independent components, or the size of the used patches. But the most important characteristic of each method is related to the details of the approximation proposed to estimate the inverse of the probability density. This estimation can be done comparing the distributions of features within the image in a global [BT09] or in a local manner [SM09], or comparing that distribution of features against *remembered* distributions from a set of training images [ZTM+08].

Many other models give a rigid approach to compute color distinctiveness, using a given fixed color representation and relying the measure of saliency only in the competition of spatial features. These spatial features can be those obtained from linear filtering of the image with a bank of Gabor-like filters [GMV07, TOCH06], but also the power of spatial frequencies in the Fourier domain [HZ07a]. Such a competition is again usually performed through the computation of a measure of local [GMV07] or global [TOCH06, HZ07a] comparison of distribution of features. These models have achieved a quite good performance, in spite of the poor treatment given to color. This occurs because in natural images most of saliency arises from spatial structure, and part of color saliency can be well captured in a rigid opponent component scheme. As shown in the chapter 2, a representation in the Lab color model achieves a high degree of decorrelation.

Ultimately, all of these models -grounded on a particular estimation of the inverse of the probability density of features, in a given predefined space of low level features- allow the interpretation that they rely in an estimation of the optical variability present in the image. Indeed, the measure of optical variability proposed here is a multivariate measure of variance, and thus a global measure of the distance of the local optical composition, from the distribution of optical magnitudes over the image. Models using local comparisons may be interpreted as computing optical variability in a reduced neighborhood. Besides, models using a learned distribution [ZTM$^+$08] would compute a kind of experienced optical variability that would include previous experience. Such a measure would introduce an additional rigid component also in the measure of distinctiveness, that is against the proposed strategy of contextual adaptation. Anyway, as far as these models use a predefined set of low level features, different from the optical magnitudes involved, they run the risk of introducing biases in the measure, as comparisons with eye fixations appear to suggest (see chapter 4). Otherwise, some of them do not use at all a plausible scheme for pooling of color and spatial features, being a remarkable example the spectral residual approach [HZ07a].

Models of saliency based on object segmentation can be linked to adaptive whitening, since whitening of scales provides responses that in many cases represent separately foreground and background. For instance, in the model of Achanta et al. saliency is a simple measure of distance to the mean value in the Lab color model [AEWS08]. This saliency map is segmented to retain the most salient regions as salient objects. As shown in chapter 2 such distance in a Lab color space, roughly approximates distance in a color whitened space. Hence, these models would constitute constrained implementations that take advantage of one remarkable consequence of an efficient representation, that is figure-ground segmentation. This solution would be somewhat equivalent

to compute the distinctiveness in a reduced set of whitened components -those that provide a data driven foreground segmentation-, catching hence the optical variability retained by these components.

Other approaches are based on a bioinspired but theoretically unbounded modelling of visual functions. They are mainly inspired by the hierarchical organization of the visual cortex [GM04]. Consequently, they have been referred to as *hierarchical* models in a recent review by Le Meur et al. [MC10]. Therefore they rely in the realization of a series of visual functions like center-surround competition, perceptual grouping or visual masking. These models seem harder to explain in terms of optical variability [IKN98,IK00,MCBT06]. Nevertheless, they claim to look for points *distinctive* from the surround, and thereby they ultimately allow a similar interpretation. Besides, the particular definition of optical variability proposed here is based upon the separate whitening of color and scale features. This approach is indeed inspired in the hierarchical functioning of the HVS, thought to seggregate the processing of color and form. As well it considerably reduces the computational load in comparison to a joint whitening of optical magnitudes.

The hypothesis underlying most of strategies in modelling of saliency, can be formulated and formalized then in a simple manner: Human visual system is naturally tuned to prefer sources of optical variability in the environment, and only specifical training and/or voluntary behavior is able to tune it to other kind of relevance that can also be inferred from images using knowledge or experience. There is a large support for this hypothesis, and we have provided here evidences that further reinforce this support.

AWS is however the only approach to saliency that can be explicitly linked to a simple measure of optical salience, defined as the relative contribution to the spatial variability of few optical magnitudes. At least of the interval of them comprissed within the optical visual window, that is, the part of these magnitudes -in terms of range and resolution- that survives the filter of early vision. As shown in the chapter 6 when dealing with hyperspectral images the optical visual window retains a main portion of the physically existing optical variability.

Otherwise, the AWS is compatible as well with a top-down approach to visual attention like the proposal of contextual influences by Oliva and Torralba. Both approaches are based on early feedforward representations of images. According to the experimental results obtained by them, this representation is already available in the first fixation, and it retains low spatial frequencies, to further accumulate information of higher spatial frequencies. This fast feedforward representation would explain that when a contextual knowledge influences a visual search, it modifies the spatial distribution of fixations. As a result fixations are directed to regions of high probability

to find the target searched. But what happens if no target is looked for and consequently contextual information does not affect surveillance strategy?. It is reasonable to expect that saliency arising from such a feedforward representation would guide the deployment of attention. It is striking that most of saliency is retained in AWS, in spite of a remarkable downsampling of the input image. This points to intermediate scales as determinant in a kind of layout saliency that would explain most of early fixations in a free surveillance task.

3.A. Appendix

In Fourier optics any image can be considered as a wavefront piece and approached as a superposition of ideal monochromatic plane waves [ST91]. A monochromatic plane wave can be characterized by means of its amplitude A, its spectral wavelength λ and its wave number vector $\mathbf{k}$ (i.e. its direction of propagation).

$$E(x, y, \lambda, \mathbf{k}) = A(\lambda, \mathbf{k}) \exp\left(\mathbf{k} \cdot \mathbf{r} - i(c/\lambda)t\right); \tag{3.19}$$

with $\mathbf{k}$ being a vector of free orientation and with modulus $k = 2\pi/\lambda$, and being c the speed of light.

The visual system is only sensible to light intensity, that means to the squared modulus of the different plane waves, and not to the ultrafast phase of ligth wavefronts. Besides natural images are in general illuminated by diffuse or extended sources like the sun, hence the eye can be assumed to be an incoherent system which is linear in intensity [Goo05, Gas78]. Consequently, the image intensity can be described by the expression:

$$I(x, y, \lambda, \mathbf{k}) = EE^* = A^2(\lambda, \mathbf{k}) \exp\left(2\mathbf{k} \cdot \mathbf{r}\right); \tag{3.20}$$

Hence, being u and v the rectangular components of the two dimensional spatial frequencies on an image plane parallel to the $x - y$ plane, they are related to the wave number vector through the expression

$$\mathbf{k} = 2\pi u\mathbf{i} + 2\pi v\mathbf{j} + k_z\mathbf{k}; \tag{3.21}$$

so that the spatial frequencies contributed by a given plane wave depend on the projection of its wave number vector on the $x - y$ plane. That means, they can be derived from both the angle with the image plane and its spectral wavelength, so that:

$$u = (1/\lambda)\sin\theta_x \approx (1/\lambda)\theta_x$$
$$v = (1/\lambda)\sin\theta_y \approx (1/\lambda)\theta_y \tag{3.22}$$

where θ_x and θ_y are the angles that the wave number vector makes with the planes $y-z$ and $x-z$, respectively, and the sinus becomes the angle in the paraxial approximation (for small angles).

That said, the spectral value determines the chromatic properties of the plane wave, while both the spectral value and the angle between the wave number vector and the image plane determine the spatial frequency contributed by the plane wave [ST91]. Besides on an image plane, the plane wave can be represented by a intensity value at each point. From the previous argumentation, it follows that the intensity of an image can be obtained from the integral of the light intensities in the continuum of plane waves, that is:

$$I(x,y) = \int_{\lambda_0}^{\lambda_1} I(x,y;\lambda)d\lambda = \int_{\lambda_0}^{\lambda_1}\int_{u_0}^{u_1}\int_{v_0}^{v_1} I(x,y;\lambda;u,v)d\lambda du dv \tag{3.23}$$

where

$$I(x,y;\lambda) = \int_{u_0}^{u_1}\int_{v_0}^{v_1} I(x,y;\lambda;u,v)du dv \tag{3.24}$$

Since spatial information is coded in the spatial frequencies, a given point can be referred by a single unidimensional index $(x,y) \rightarrow p_{xy}$. Using more conveniently polar instead of rectangular coordinates to represent spatial frequencies, an image can be formalized by the expressions:

$$I(p_{xy}) = \int_{\lambda_0}^{\lambda_1} I(p_{xy};\lambda)d\lambda \tag{3.25}$$

and

$$I(p_{xy};\lambda) = \int_{\rho_0}^{\rho_1}\int_{\alpha_0}^{\alpha_1} I(p_{xy};\lambda;\rho,\alpha)d\rho d\alpha \tag{3.26}$$

where ρ and α are respectively, the modulus and the angle of the spatial frequency.

The local contribution to such a superposition of monochromatic plane waves can be described in terms of the spatial power distributions of chromatic components -related to electromagnetic wavelength- and of the corresponding power distributions of magnitude and orientation of the spatial

frequencies present for each of them -related to the wave number vector. The spectral power distribution is given by the left side of equation 3.26, while the power distribution of spatial frequencies for a fixed λ can be represented by the argument of the integral in the right side of the same equation.

Chapter 4

Prediction of Human Fixations

In this chapter, the results obtained in validation experiments founded on the predictive capability of human fixations are shown. It must be remarked that no kind of parameter tuning has been made to obtain the following results. The setup of the model is exactly the same that has been described in detail in the previous chapter.

A variety of approaches have been proposed in the assessment of computational models of saliency. Some of them use an objective measurement of technical performance in a given task. An example of this can be found in the use of the improvement on recognition performance by van de Weijer et al., as a measure of usefulness of their color saliency approach [vdWGB06]. Also, the ability to detect salient objects –like traffic signs– has been used to assess saliency model performance [IKN98, AL10]. However, the most widely employed evaluation methods rely on the comparison with human fixations like for instance in [BT09, GMV08, ZTM$^+$08, SM09, HZ08]. This approach is more general than specific purpose recognition tasks, and it is clearly related to saliency. Besides, the use of natural images prevents from experimental design biases. This is why quantitative measures of prediction of human eye fixations on natural images are currently seen as the most reliable way to assess a given measure of saliency. Moreover, most models try to show their capability in pushing this benchmark further on.

4.1. Capability of predicting fixations

Human fixations are usually obtained through experiments in which subjects observe images or sequences in a screen. In these experiments, eye movements are recorded and analyzed through the use of an eye tracker, and the angles of fixations and saccades –or directly their positions on the

screen–are provided. Depending on the device and the software associated, a variety of data (like landing time or precision) is recorded, and different analysis are possible. Usually, to assess models of saliency only the fixations -positions around which the eye is steadily landing during enough time- are used. Fixations refer to the positioning and accommodation of eyes that allows the image to be brought into the fovea, and they are hence thought to be essential for the impression of clarity in the visual perception. Thereby, they appear to have a major role in visual attention.

In the assessment of saliency with eye-tracking results, the objective is not the reproduction of fixation times or order of fixations. Most frequently, the purpose is to test the explanatory capability of saliency for the spatial distribution of early fixations from a group of subjects, which observe images without any specific purpose. These conditions try to minimize the influence of top-down mechanisms on the behavior of the subjects.

4.1.1. Procedure, datasets and results

The saliency maps have been compared with human fixations through the use of the area under the curve (AUC), obtained from a receiver operator curve (ROC) analysis, as proposed by Tatler et al [TBG05]. The method has been employed to validate a wide variety of state-of-the-art saliency models, providing a reliable measure for comparison. In this procedure, one unique curve is drawn for a whole set of images. The area under this curve can be used to measure the capability of saliency to discriminate between fixated and non fixated points. To avoid center-bias, in each image, only points fixated in another image from the same dataset are used as non fixated points. As suggested by Tatler et al., standard error is computed through a bootstrap technique, shuffling the other images used to take the non fixated points, exactly like in [ZTM$^+$08] and in [SM09]. The appendix at the end of the chapter shows results with two other measures based on the Kullback-Leibler divergence.

The particular implementation of the method proposed by Tatler et al. and done by Zhang et al. has been adopted by two main reasons. Firstly, it has been recently used to assess several state-of-the-art models both by Zhang et al and by Seo and Milanfar [ZTM$^+$08, SM09]. This fact clearly facilitates comparison in a fairly fashion with existing approaches. Secondly, it is robust against tricks like border supression used in many models.

Two open-access eye-tracking datasets of natural images have been used. The first of them has been published by Bruce and Tsotsos and has 120 images and fixations from 20 subjects [BT06b]. This dataset has already been used to validate several state-of-the-art models of bottom-up saliency like for

Table 4.1: AUC values obtained with different models of saliency for both of the datasets of Bruce and Tsotsos and Kootstra et al. Standard errors, obtained like in [ZTM+08], range 0.0007-0.0008. For the groups of the Kootstra et al. dataset, standard errors range 0.0010-0.0018. (* Results reported by [ZTM+08]; ** Results reported by [SM09]).

Model	Bruce and Tsotsos dataset	Kootstra et al. dataset					
		Whole dataset	Buildings	Nature	Animals	Flowers	Street
AWS	0.7106	0.6205	0.6105	0.5815	0.6565	0.6374	0.7020
Seo and Milanfar	0.6896**	0.5933	0.6136	0.5530	0.6445	0.5602	0.6907
Hou and Zhang	0.6823**	0.5750	0.5902	0.5426	0.6086	0.5659	0.6419
AIM	0.6727*	0.5842	0.5766	0.5628	0.5953	0.5881	0.6393
SUN	0.6682*	0.5705	0.5514	0.5484	0.5401	0.6100	0.6458
Itti et al.	0.6456	0.5702	0.5814	0.5478	0.6200	0.5217	0.6509
Gao et al.	0.6395*	–	–	–	–	–	–

instance in [BT09, GMV08, ZTM+08, SM09, HZ08]. The second dataset has been published by Kootstra et al. and consists of 99 images and the corresponding fixations of 31 subjects [KNdB08, KS09]. One interesting property of this dataset is that it is organized in five different groups of images (12 images of animals, 12 of streets, 16 of buildings, 40 of nature, and 19 of flowers or natural symmetries). The main purpose of the use of different datasets in this work was to assess the robustness and reliability of the evaluation procedure.

In the table 4.1, the obtained rankings of models are shown using both image datasets of Bruce and Tostsos and Kootstra et al. The results shown for the model of Itti et al. [IKN98] are higher than in other works [ZTM+08, SM09] because instead of using their saliency toolbox, the original version has been used, as made available for Matlab (`http://www.klab.caltech.edu/~harel/share/gbvs.php`). In the figure 4.1, the saliency maps obtained with ten images from each of the datasets with the three best models are shown. As well, figures 4.2 to 4.12 show all the results for the AWS on both datasets as well as the fixations density maps provided by the authors.

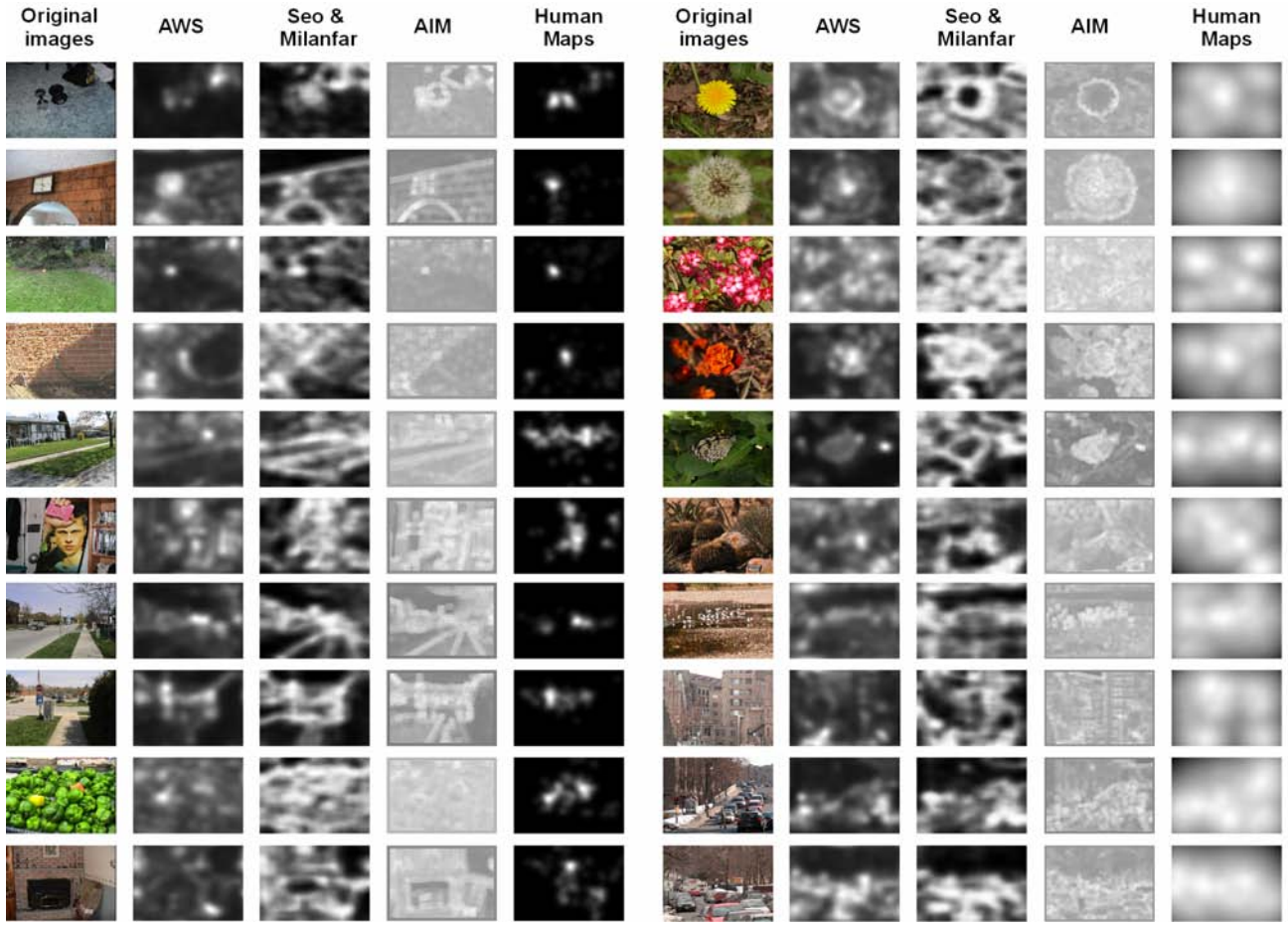

Figure 4.1: Illustrative results for comparison with state-of-the-art models and humans for 10 images from each of the datasets of Bruce and Tsotsos (left) and Kootstra et al. (right) (see the text for details).

4.1.2. Discussion of results

As we can see in the tables, this evaluation method suffers from an obvious problem. The tight uncertainty values of AUC obtained for a given dataset or group of images are clearly incompatible with the ones obtained with the others. Hence, it is well-grounded to question the validity of these uncertainties, below the thousandth part, and their use as the minimum relevant difference between models.

We could be tempted to attribute the differences between datasets to differences in experiments. But this would not explain the even higher differences found between the categories provided by Kootstra et al.. The variation of results between types of scenes is really high, to the point of making uncertainties and even differences between models seem irrelevant.

Therefore, it could happen that each of the models catches different kinds of saliency better and, hence, some models might work better with certain images than others. That is, we could think that the results are scene-biased or feature-biased for a given dataset. But there is also something that seems to question this explanation: despite the high variation in the AUC values, the resulting ranking is quite stable. It is the same for both datasets, although it is not the same for all of the groups of images. AWS gets the highest AUC value for both of the datasets and four of the groups. Only in the buildings group is slightly outperformed by the model of Seo and Milanfar.

Other factors that could explain differences between datasets are those related to a different relative influence of saliency on the behavior of humans. For instance, there might exist differences on the strength of the influence of saliency in driving fixations, owing to a different spatial concentration of saliency itself. There could also be top-down mechanisms affecting differently to different types of images, like in [ESP08] and in [BBK09]. Any of these two factors could explain the disparity in results observed for the several groups of images.

4.2. Comparison with humans

From the last observations, it seems reasonable to compare the capability of the model of predicting human fixations with that shown by humans themselves, this is, to use a measure of relative prediction capability, rather than a simple measure of prediction capability, as usually done. This would suppress the effect of particularly strong subjective –be random or top-down– behavior in certain images, and as a consequence, would provide a more robust measure of performance, less affected by inter-scene variance. In case there

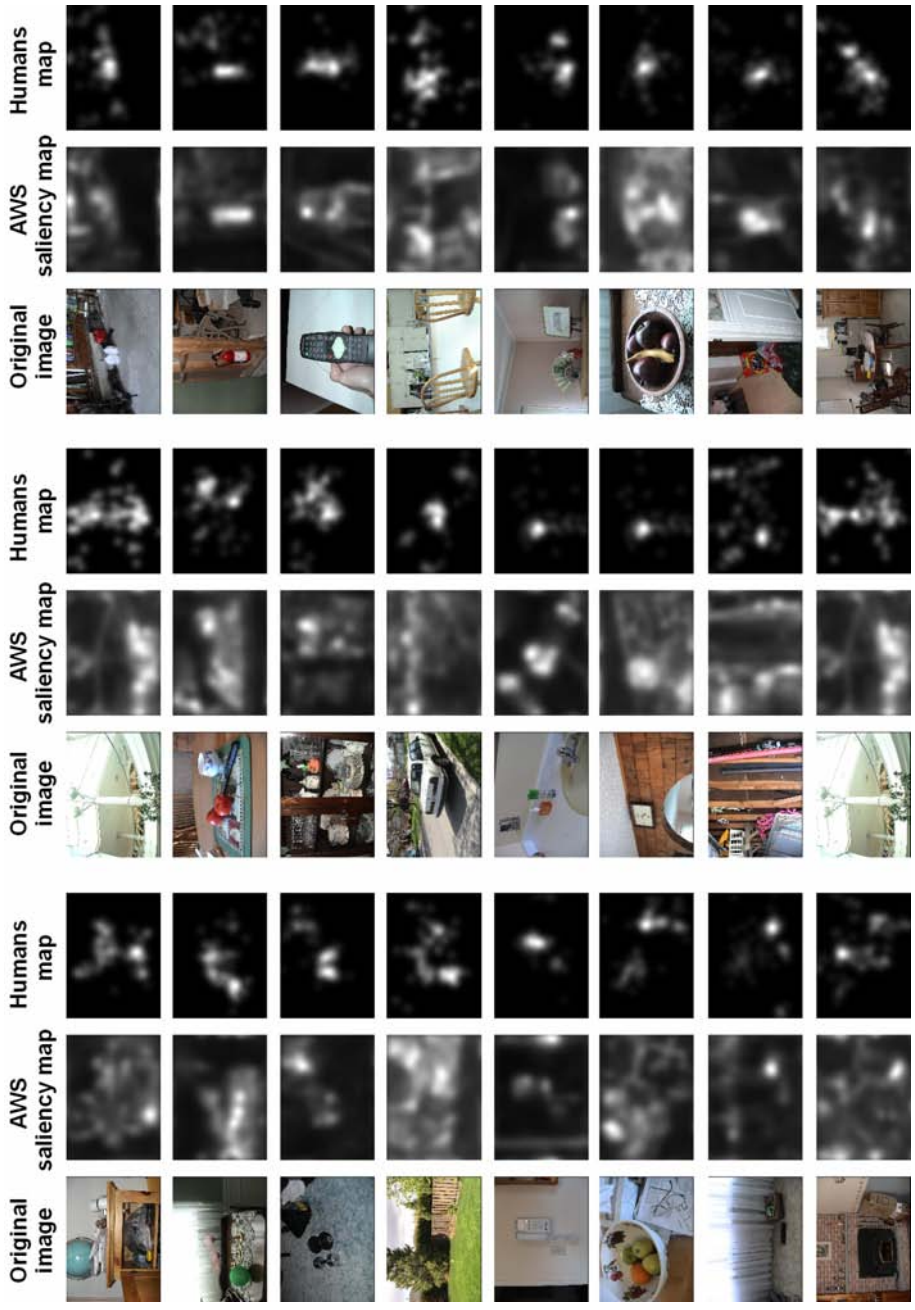

Figure 4.2: Complete results on the dataset of Bruce and Tsotsos. The human maps are those provided by the authors, obtained through Gaussian kernels applied on fixations and averaging through subjects (I).

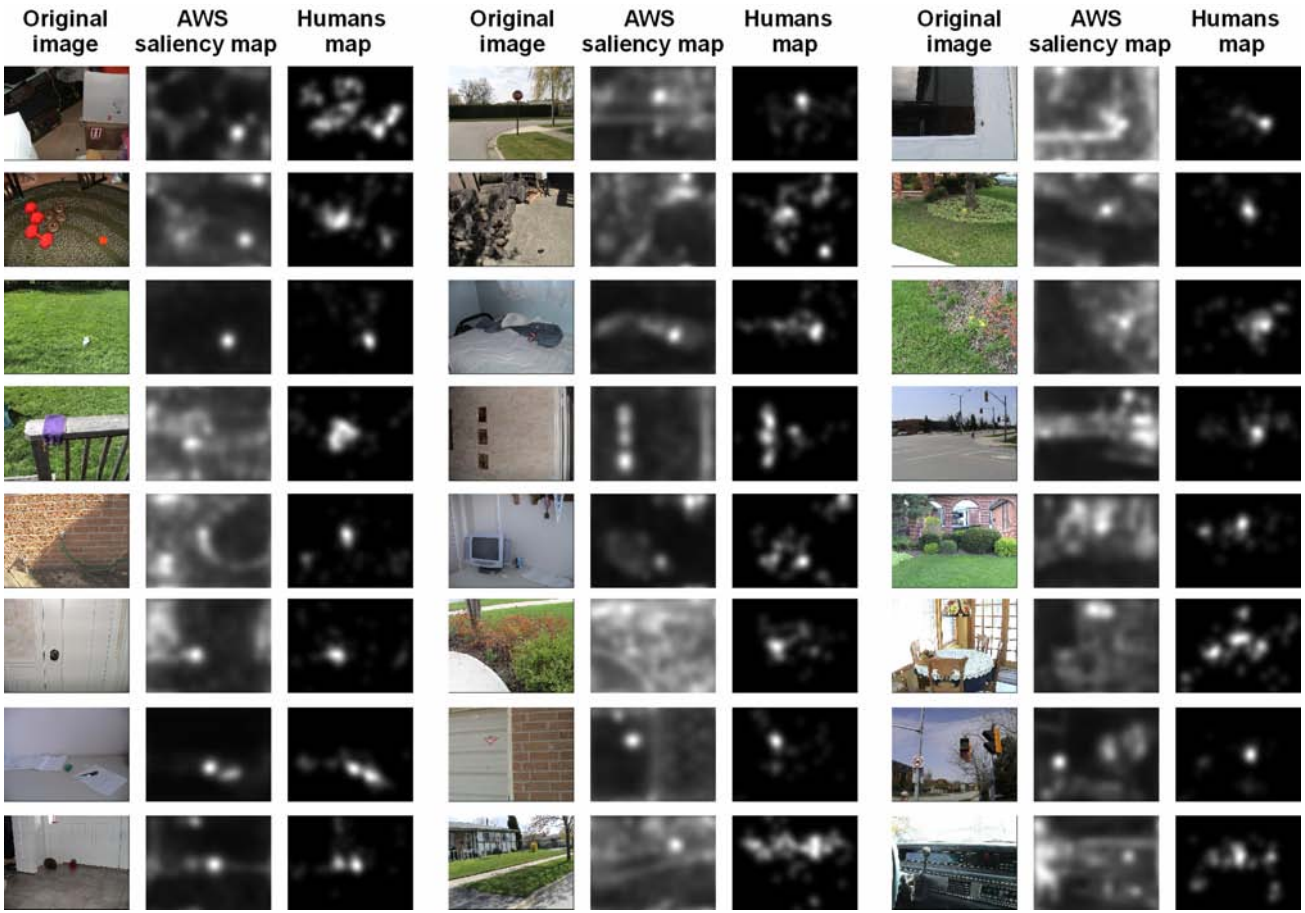

Figure 4.3: Complete results on the dataset of Bruce and Tsotsos. The human maps are those provided by the authors, obtained through Gaussian kernels applied on fixations and averaging through subjects (II).

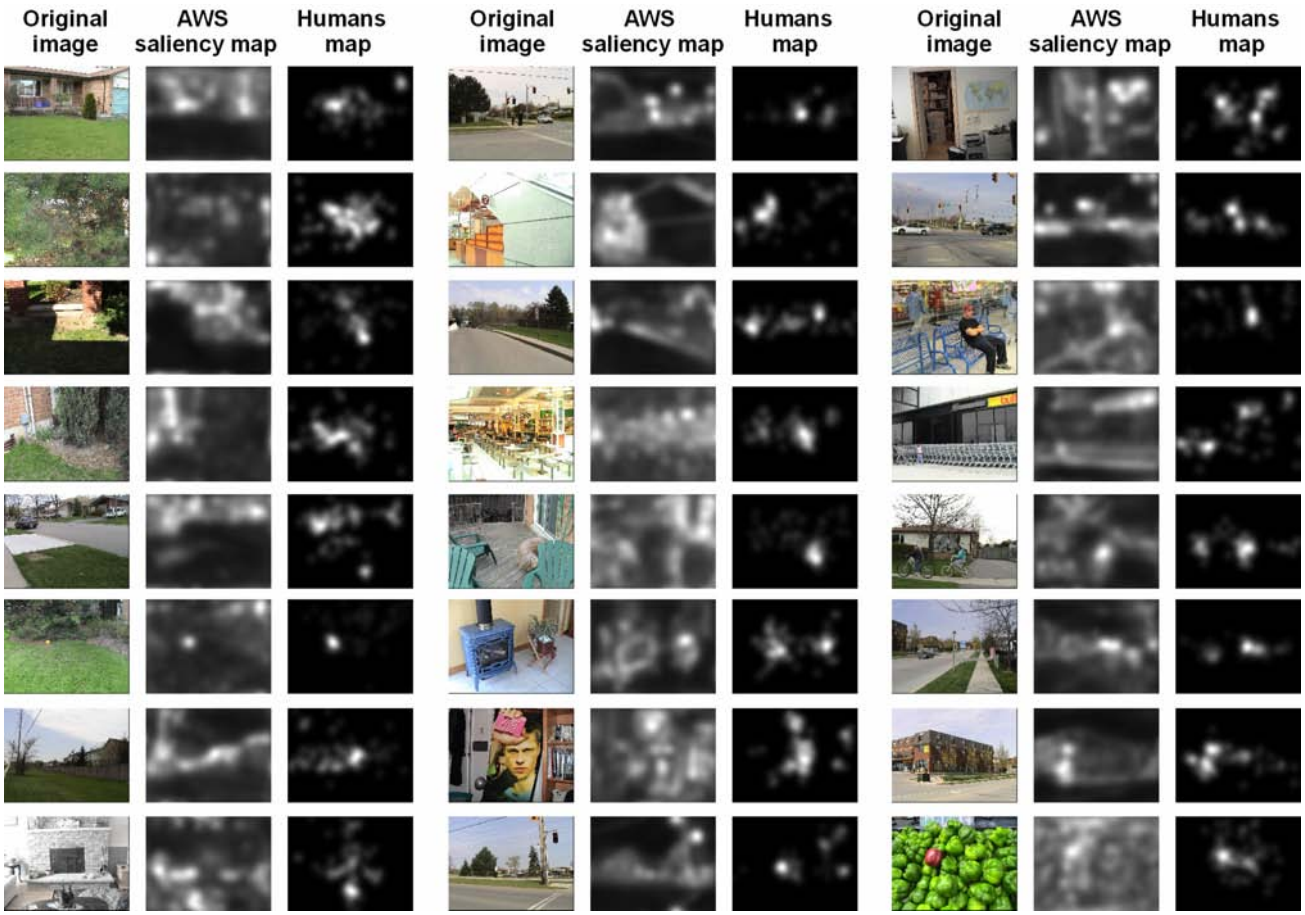

Figure 4.4: Complete results on the dataset of Bruce and Tsotsos. The human maps are those provided by the authors, obtained through Gaussian kernels applied on fixations and averaging through subjects (III).

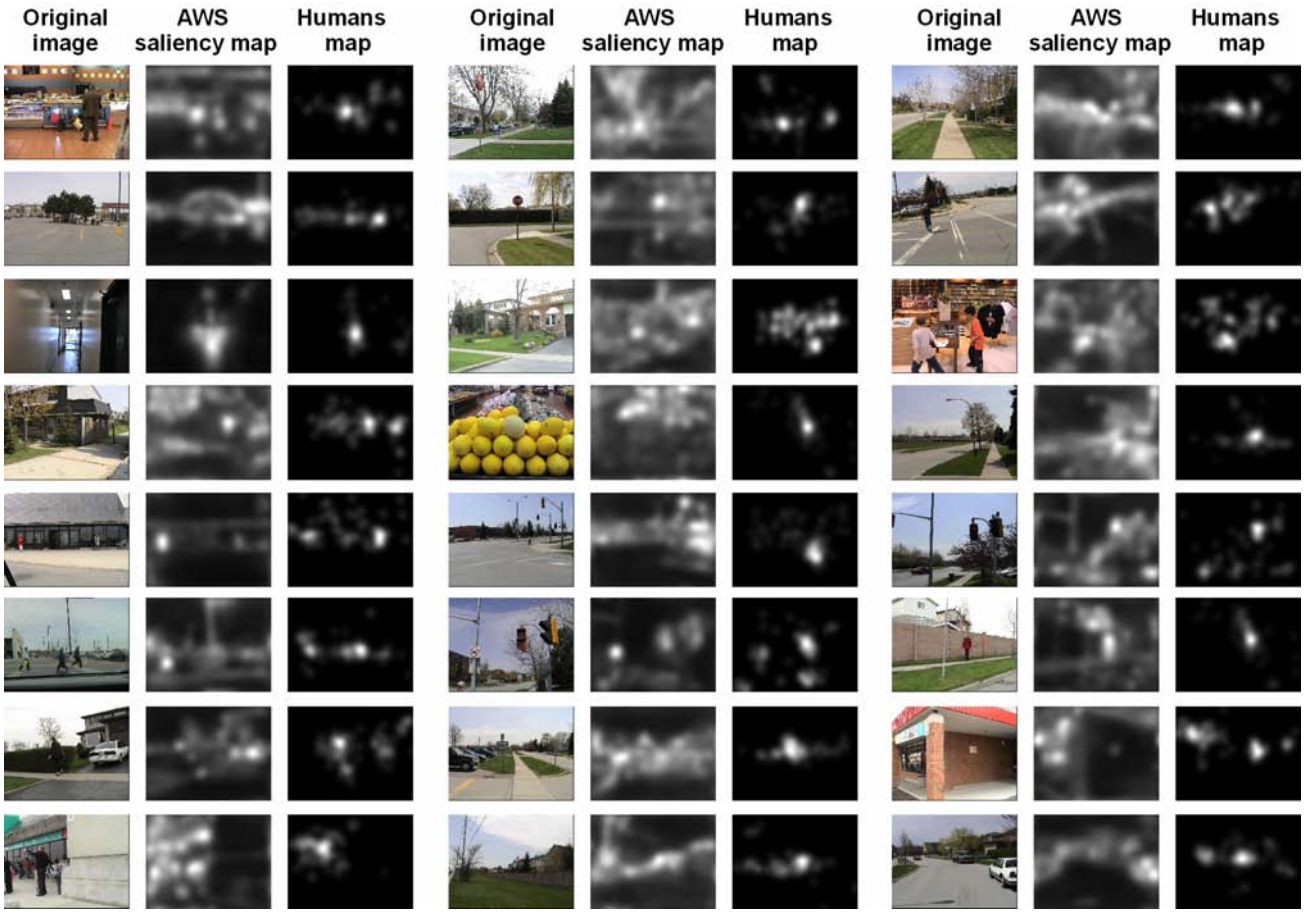

Figure 4.5: Complete results on the dataset of Bruce and Tsotsos. The human maps are those provided by the authors, obtained through Gaussian kernels applied on fixations and averaging through subjects (IV).

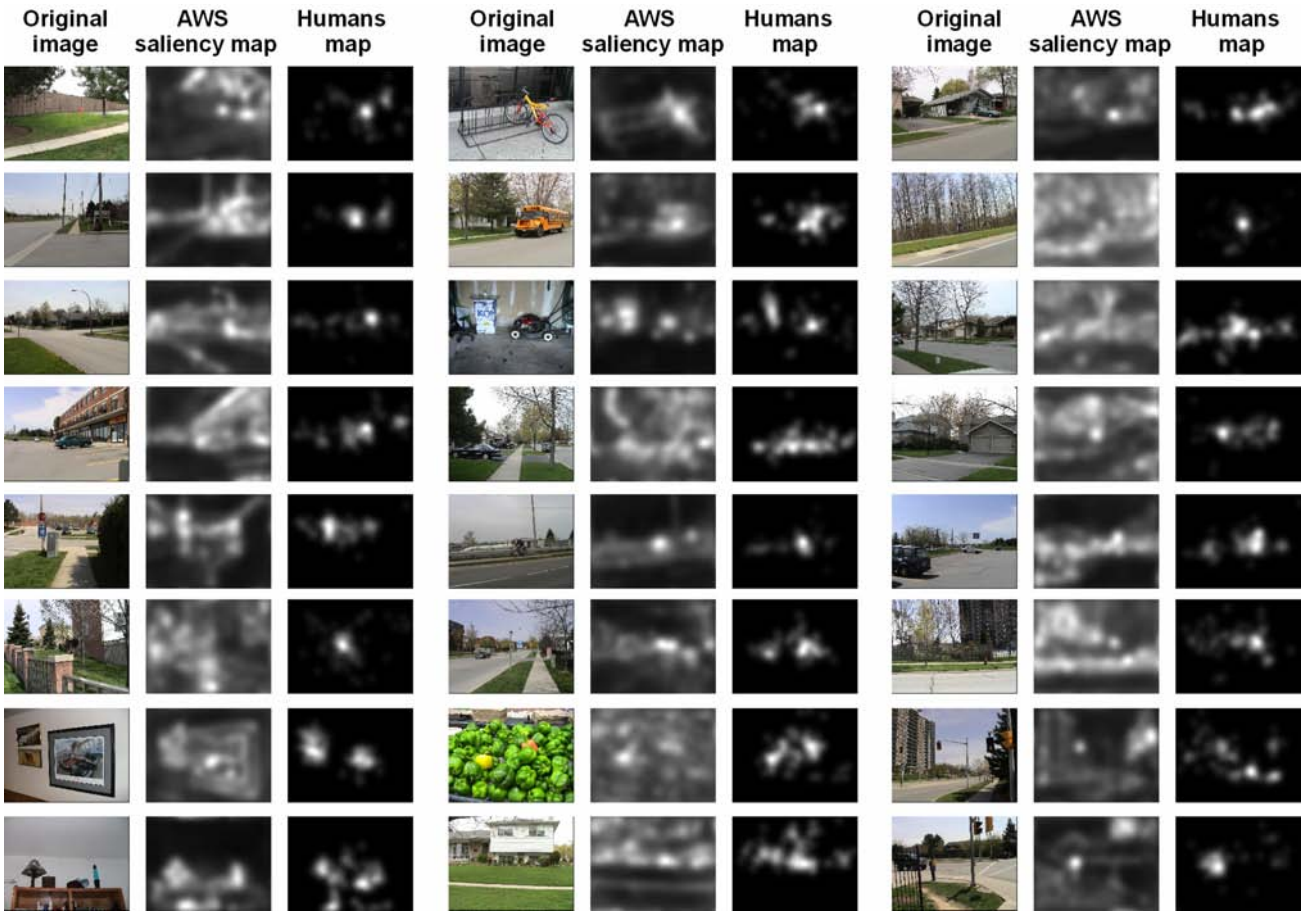

Figure 4.6: Complete results on the dataset of Bruce and Tsotsos. The human maps are those provided by the authors, obtained through Gaussian kernels applied on fixations and averaging across subjects (V).

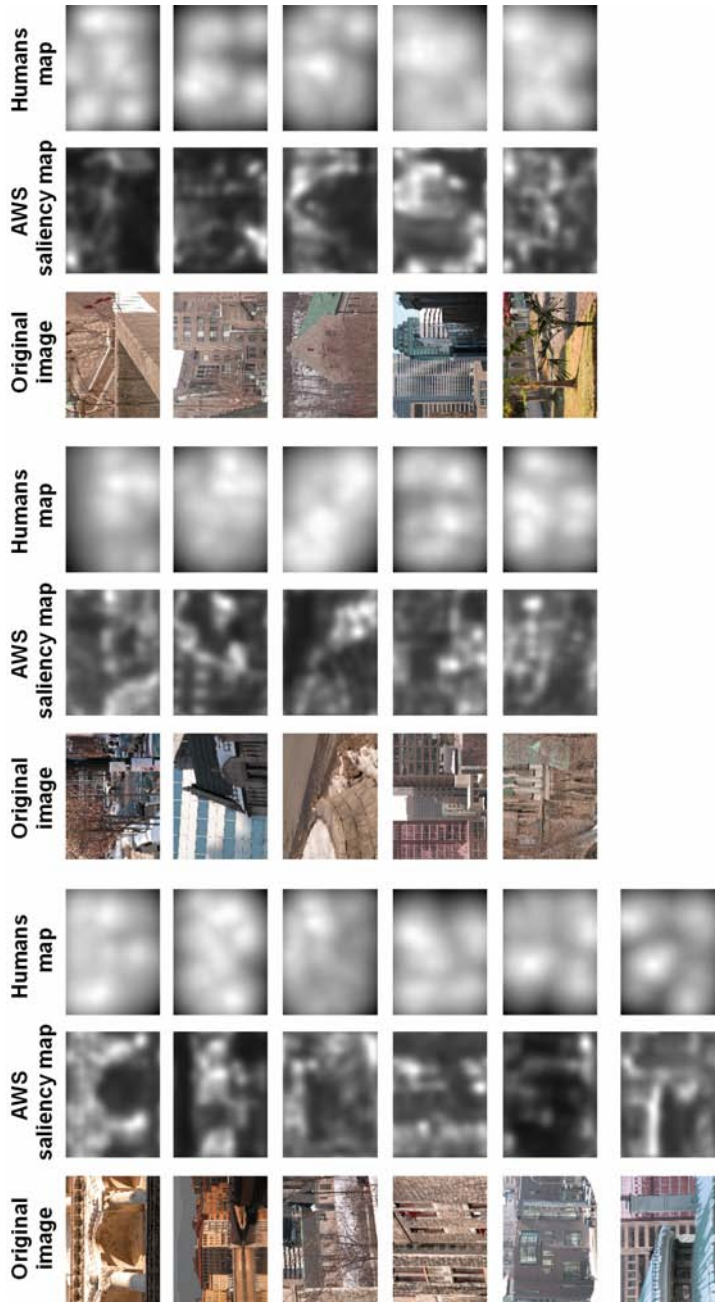

Figure 4.7: Complete results on the buildings group from the dataset of Kootstra et al.. The human maps are those provided by the authors, obtained through distance to fixation transform for each observer and averaging across subjects.

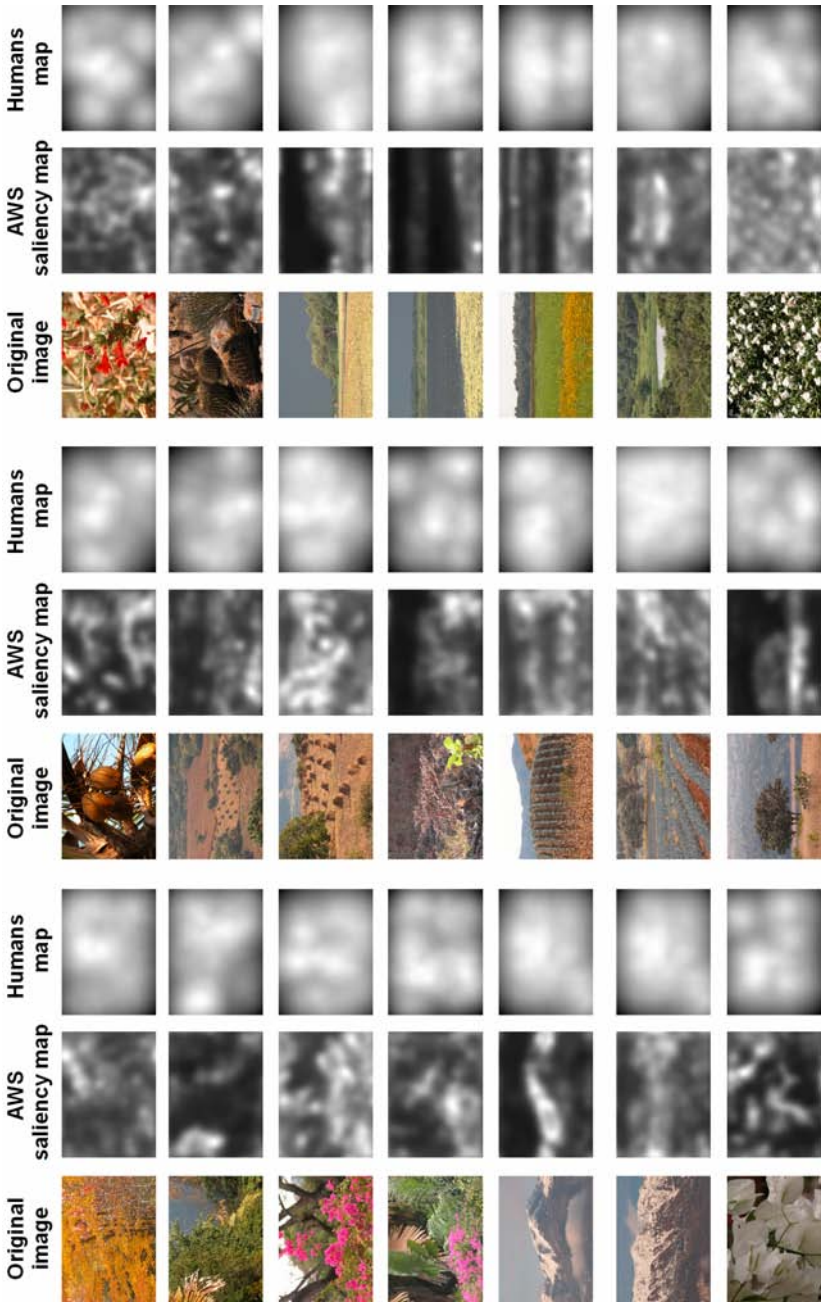

Figure 4.8: Partial results on the nature group from the dataset of Kootstra et al. (I). The human maps are those provided by the authors, obtained through distance to fixation transform for each observer and averaging across subjects.

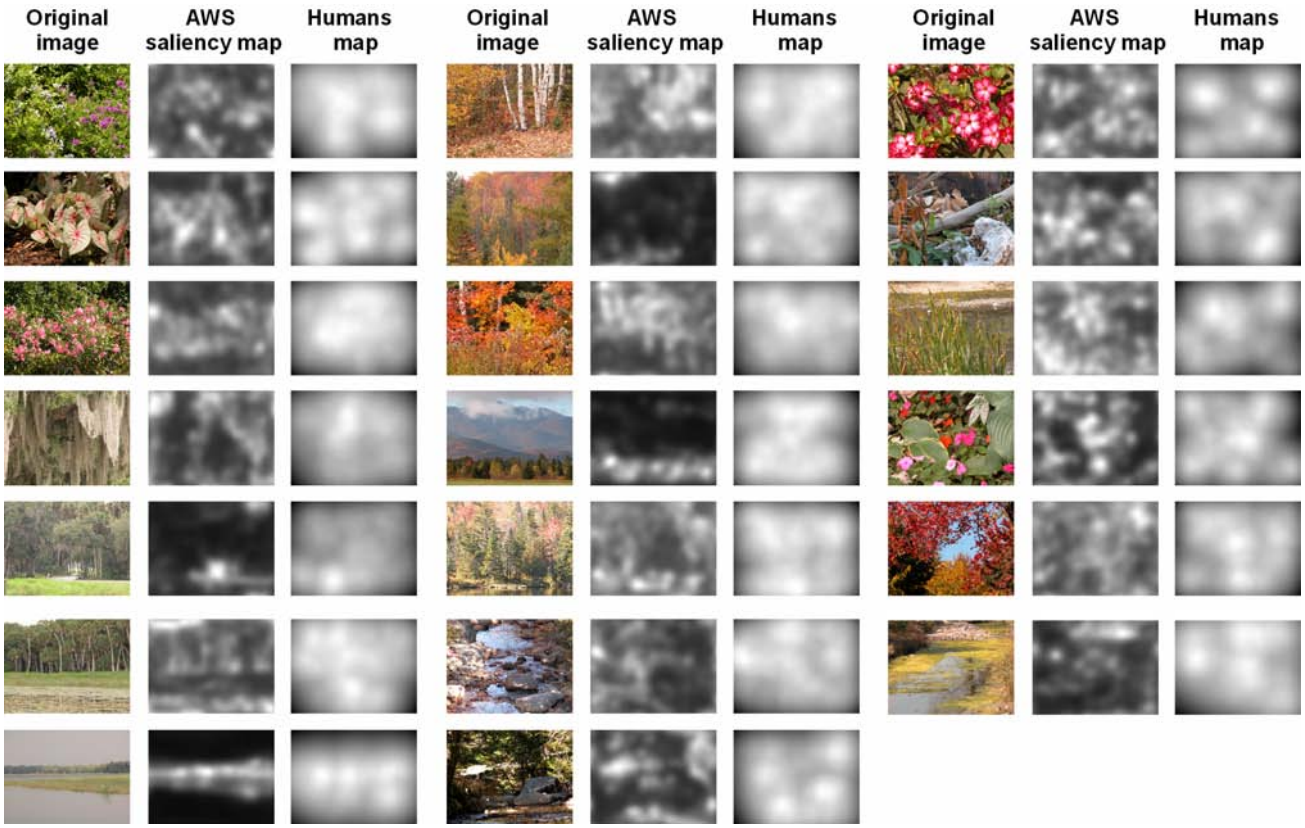

Figure 4.9: Partial results on the nature group from the dataset of Kootstra et al. (II). The human maps are those provided by the authors, obtained through distance to fixation transform for each observer and averaging across subjects.

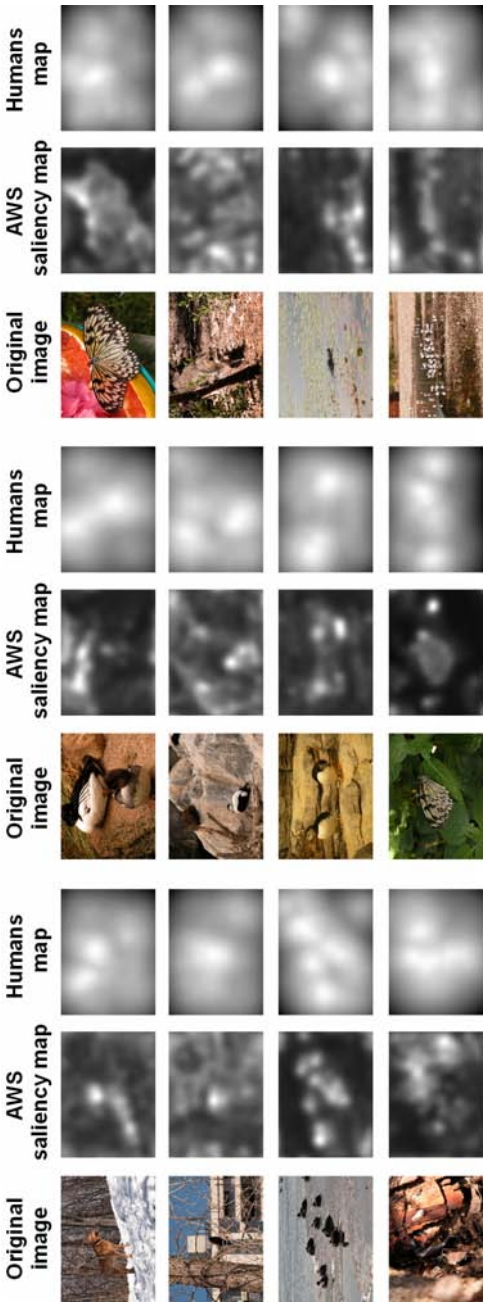

Figure 4.10: Complete results on the animals group from the dataset of Kootstra et al.. The human maps are those provided by the authors, obtained through distance to fixation transform for each observer and averaging across subjects.

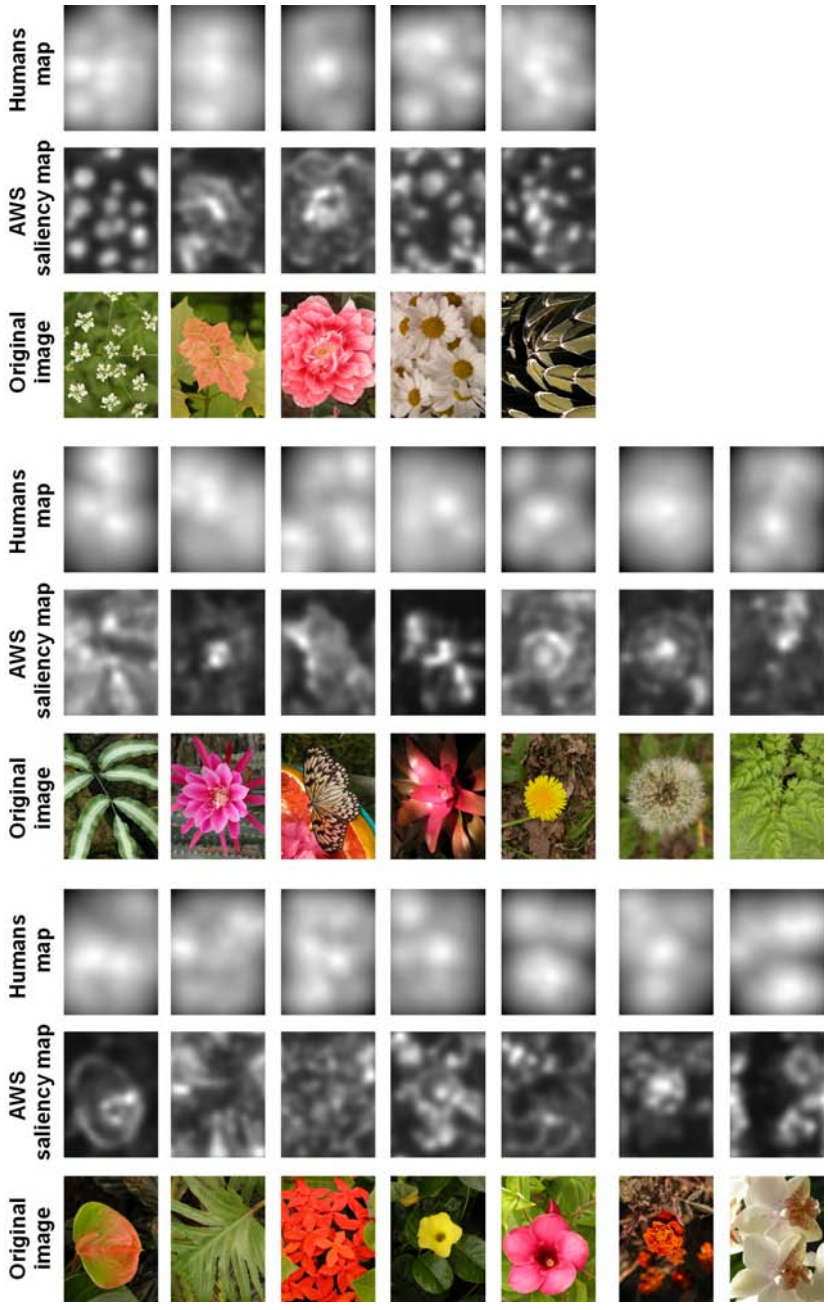

Figure 4.11: Complete results on the flowers group from the dataset of Kootstra et al.. The human maps are those provided by the authors, obtained through distance to fixation transform for each observer and averaging across subjects.

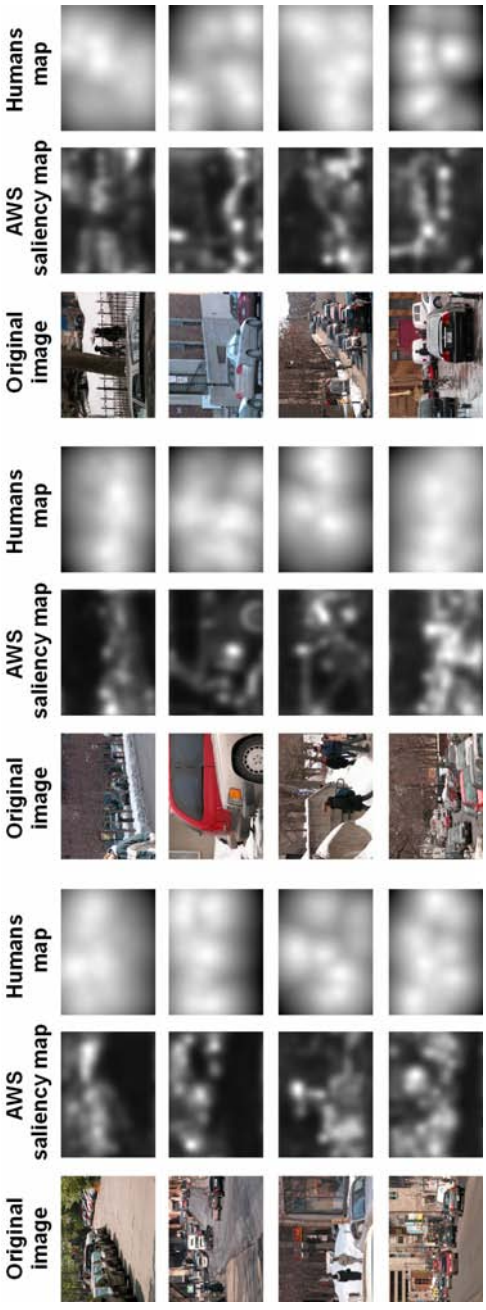

Figure 4.12: Complete results on the street group from the dataset of Kootstra et al.. The human maps are those provided by the authors, obtained through distance to fixation transform for each observer and averaging across subjects.

was an important amount of top-down behavior, common to all subjects, this would result in a decrease of the relative capability of prediction. Thus, such an evaluation can also give interesting additional information about human behavior in a visual surveillance task.

4.2.1. Human predictive capability

To implement this measure, priority maps derived from fixations will be used, following the method described by [KS09]. This method lies in the subtraction of the distance between each point and its nearest fixation from the maximum possible distance in the image. As a result, fixated points have the maximum value and non fixated points have a value that decreases linearly with distance to the nearest fixation. The resulting maps can be used as probability distributions of subjects fixations (priority maps), and can be considered as subjective measures of saliency.

At least with few fixations per subject, as it is the case, this method yields improved predictive results than the approach to compute priority maps based on filtering of fixations with truncated Gaussians [Oue04,BT06b]. This other approach has the problem of assigning zero priority to most points, despite the fact of having different distances to the nearest fixation. Furthermore, the linear distance-based method is parameter free. Of course, it can be argued that it is not justified to assume that priority drops linearly with distance to fixations. Nevertheless, it seems actually reasonable to assume that priority drops monotonically with distance to the nearest fixation. If the method to compare and evaluate maps is invariant to monotonic transformations, as ROC analysis is, then there is no issue with using linear, quadratic or any other monotonic maps.

Hence, through a ROC analysis, the same one employed to evaluate models of saliency, the capability of these maps to predict the fixations of the set of subjects can be assessed. This implies the capability of predicting fixations of all subjects from fixations of a single subject.

The described differences between maps with both methods can be qualitatively compared in the figure 4.1 for a selection of examples, or seaching in the figures 4.2 to 4.12 (in both cases the *Human maps* column), where the average priority maps provided by the corresponding authors with each dataset are shown. In Gaussian-based maps most of the points have zero priority, while in distance-based maps a gray-scale continuum with several local maxima covers the image. It must be noticed that we have not used these averaged maps, but distance-based maps computed for each of the subjects. Such a procedure emphasizes still more the differences.

The previous evaluation for each subject has been done, only for those

Table 4.2: Average predictive capability of humans using distance-to-fixation priority maps.

Bruce and Tsotsos dataset		Kootstra et al. dataset							
		Whole dataset		Buildings		Animals		Street	
Mean	2σ	Mean	2σ	Mean	2σ	Mean	2σ	Mean	2σ
0.6946	0.0248	0.6254	0.0224	0.6154	0.0330	0.6672	0.0356	0.6923	0.0402
				Nature		Flowers			
Max	Min	Max	Min	Mean	2σ	Mean	2σ		
0.7156	0.6805	0.6462	0.6056	0.5874	0.0194	0.6419	0.0245		

with fixations for all of the images. One individual has been excluded of the dataset of Bruce and Tsotsos, whose deviation from the average of humans was larger than the standard deviation, and who also had just one fixation in many images. This yields 9 evaluated subjects for the dataset of Bruce and Tsotsos, and 25 subjects for the dataset of Kootstra.

From there, the predictive capability of each of these subjects has been obtained for the fixations of the whole set. Computing the average, there is also the predictive capability of the average subject. Besides, the double of the standard deviation provides an estimation of the range of predictive capabilities for the 95% percent of humans, assuming a normal distribution for AUC human values. This was true for the datasets and groups studied, with a kurtosis value very close to 3. Moreover, this interval can also be used as a measure of the minimum relevant distance between two models.

4.2.2. Human-model performance comparison

Examining the obtained results in table 2, it has been found that the AWS model is compatible with the average human, for both datasets and for each of the five groups of the Koostra et al. dataset. The model of Seo and Milanfar is also compatible with the average human in the dataset of Bruce and Tsotsos, and in two of the five groups of Kootstra et al., but not in the whole data set of Kootstra et al., clearly outperformed by all subjects. The model by Bruce and Tsotsos is only marginally compatible with the average human with their own data set. Furthermore, AWS is the only one that outperforms several subjects in all the cases, representing nearly half of the observers of Kootstra et al. dataset, and more than half of them in the Bruce

Table 4.3: Results of comparing predictive capabilities of saliency models, subtracting the average predictive capability of humans. Positive sign means better, and negative sign means worse, than the average human. (All results derived from tables 4.1 and 4.2).

Model	Bruce and Tsotsos dataset	Kootstra et al. dataset					
		Whole dataset	Buildings	Nature	Animals	Flowers	Street
95% of humans	± 0.025	± 0.022	± 0.033	± 0.019	± 0.036	± 0.025	± 0.040
AWS	0.016	-0.005	-0.005	-0.006	-0.011	-0.004	0.010
Seo and Milan-far	-0.005	-0.032	-0.002	-0.034	-0.023	-0.082	-0.002
AIM	-0.022	-0.041	-0.039	-0.025	-0.072	-0.054	-0.053
SUN	-0.026	-0.055	-0.064	-0.039	-0.127	-0.032	-0.047
Itti et al.	-0.049	-0.055	-0.034	-0.040	-0.047	-0.120	-0.041

and Tostsos data set. In this last, our model performs even slightly over the average human, as well as in the street group of Kootstra et al.

To provide a numerical measure that synthesizes these results, the simplest one has been chosen: the difference between the results of each model and the average human. Positive values imply higher predictive capability of the model and negative values imply higher predictive capability of the average human. As proposed above, the interval of the 95% of humans will be used as relevant difference, given that AUC errors are comparatively negligible. The values obtained are shown in table 3, for the different models.

4.2.3. Discussion of results

Results achieved by AWS in both datasets and the groups of Kootstra et al. dataset are highly consistent, not only in ranking position but also in the absolute value. Then, it seems to be robust against scene change. This means that it does not show any kind of scene bias and that it is not specially fitted for particular kinds of saliency, present in different scenes. Also, it clearly exhibits the highest performance among the analyzed models that constitute a representative sample of the state of the art. The model by Bruce and Tsotsos shows marginal compatibility of results among datasets. The model of Itti et al. presents also consistency among datasets –with the

lowest performance in both cases. The rest of the models do not show robustness when comparing with human performance and AWS is the only that maintains consistency when the groups of images of the dataset of Kootstra et al. are considered. This points out to scene or feature biases in the different models, and to a difficulty to catch certain salient features that are frequent in natural images. One clear example of such failures that can be observed in fig. 2 is the existing symmetry in many natural images, that only AWS is able to catch, but also the low sensitivity of other models to high frequencies and small scales that sometimes are very salient.

As it has been shown, the proposed measure to assess saliency fulfills the requirement of invariance against the kind of scene. As a result, it can be used to detect possible biases and lack of robustness in models. It also provides a realistic estimation of the minimum relevant difference between them. Therefore, it must be noticed that the 95% of humans is always among 0.02-0.04 around the average value. This leads us to two conclusions. Firstly, for any set of different natural images, as large as those employed here, differences of 0.02 or higher can be considered significant. Secondly, a ranking could be reverted with a different dataset due to a possible combined feature bias of the set of images used to perform the assessment and of the sensitivity of models being evaluated.

Other aspect that it is found to be particularly relevant is the fact that AWS completely matches the predictive capability of humans, and it always behaves like another human. In our understanding this means that the model is able to explain the shared factor that drives human fixations, during the free-viewing of natural images. Therefore, there do not seem to be shared top-down effects driving fixations up to increase the predictive capability of humans to a level that saliency is not able to explain. From our viewpoint, this fact reinforces the importance of bottom-up saliency in the explanation of the shared (inter-subject) behavior of the HVS. It also questions the real implications of results like those provided by [ESP08] or by [BBK09], involving top-down influences. In relation to this topic,we believe that more efforts are needed to clarify when shared behavior of humans follows the physical saliency, and when it is driven by a shared top-down mechanism (interest, motivation, abstraction, etc.).

4.3. Robustness of performance against spatial resolution

An interesting aspect related to the performance of a measure of saliency is the impact of spatial resolution on its capability of predicting fixations. The figures 4.13 and 4.14 explore this question for both of the studied datasets and a wide number of state-of-the-art and reference models. All models have been labeled with abbreviations for the sake of clarity, except the model of Itti et al, which has been labeled with the name of the first author. Most of these abbreviations have been already introduced, but two new labels are used, namely Srf for the model based on selfresemblance by Seo and Milanfar [SM09] (already assessed in this chapter), and ICL for the incremental coding length model by Hou et al. [HZ08].

Clearly, the AWS model presents not only the highest maximum performance as already claimed in the previous sections, but also an unrivaled robustness against the spatial resolution of the input image. This fact reveals that differently to other state-of-the-art models the AWS model is not biased to deal with certain scales that are most often involved in the determination of saliency. As expected from its adaptive nature to the specific context, the AWS model is able to deal with a wider range of scales (i.e. a wider spectrum of spatial frequencies) and hence not only with the scales that are most frequently the salient ones.

In the dataset of Bruce and Tsotsos, there are several models –for instance the Srf and SRS [SM09, HZ07a]–that increase monotonically their performance as the spatial resolution decreases up to a given value from which they quickly decay. For low spatial resolutions the model of Seo and Milanfar manages to outperform the AWS (using the same spatial resolution). The reason is that it is optimized for a fixed (small) value of the size of the input image of 64x64 pixels, that is, for a low spatial resolution. We have checked that the AWS can also be tuned to outperform these results at such low spatial resolution, in fact we believe that the model of Seo and milanfar is also somewhat tuned to work better in this dataset. Moreover it must be noticed that the maximum AUC value achieved by their model for such a low spatial resolution value is still clearly under the maximum achieved by AWS, as shown in the previous sections. Besides, a tuned version of the AWS for these low resolution values does not achieve either that maximum value. This points to an amount of relevant saliency present in lager scales that is lost with such a drastic downsampling. The same behavior is observed in the dataset of Kootstra et al, however here none of the models outperforms the AWS, even at lowest resolution values.

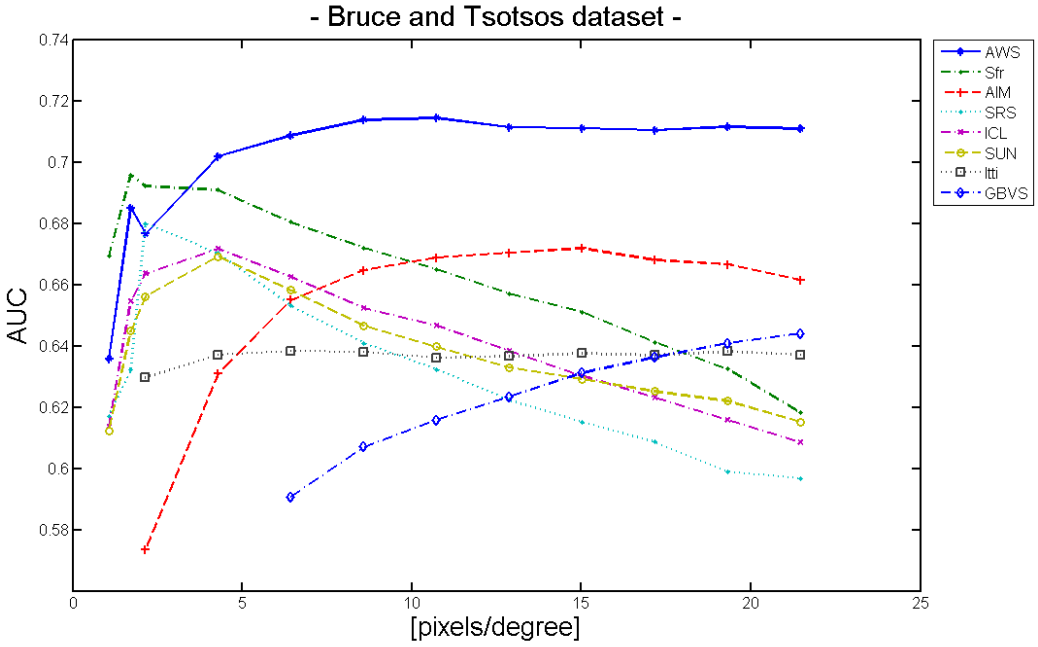

Figure 4.13: Comparison of the capability of different models to predict human fixations –measured through AUC values from ROC analysis, as explained in the first section–against the spatial resolution retained in the input image. Spatial resolution is expressed as pixels by degree of visual field for subjects. Results are shown for the dataset of Bruce and Tsotsos.

Overall we can say that several models have been designed to only work fine at low resolution values. As well, the fact that they have been assessed with the dataset of Bruce and Tsotsos used as a benchmark seems to have biased their performance. Thereby they seem to work comparatively better in that dataset but they are not able to keep the same performance in a difference dataset. This observation is in agreement with the analysis done in the previous sections when comparing with human performance.

Another question that is worth noting is that except models optimized (in the dataset of Bruce and Tsotsos) for a given size of the input image [SM09, HZ07a] that force that size, the function call of the other models have a default value of downsampling factor. We have used this unique downsampling factor in the previous sections. For the AWS this poses a problem: since the maximum spatial resolution available in the original images is different between datasets, and since the maximum performance is achieved for nearly the same value of spatial resolution, different downsampling factors should have been used to work at maximum performance. This points to

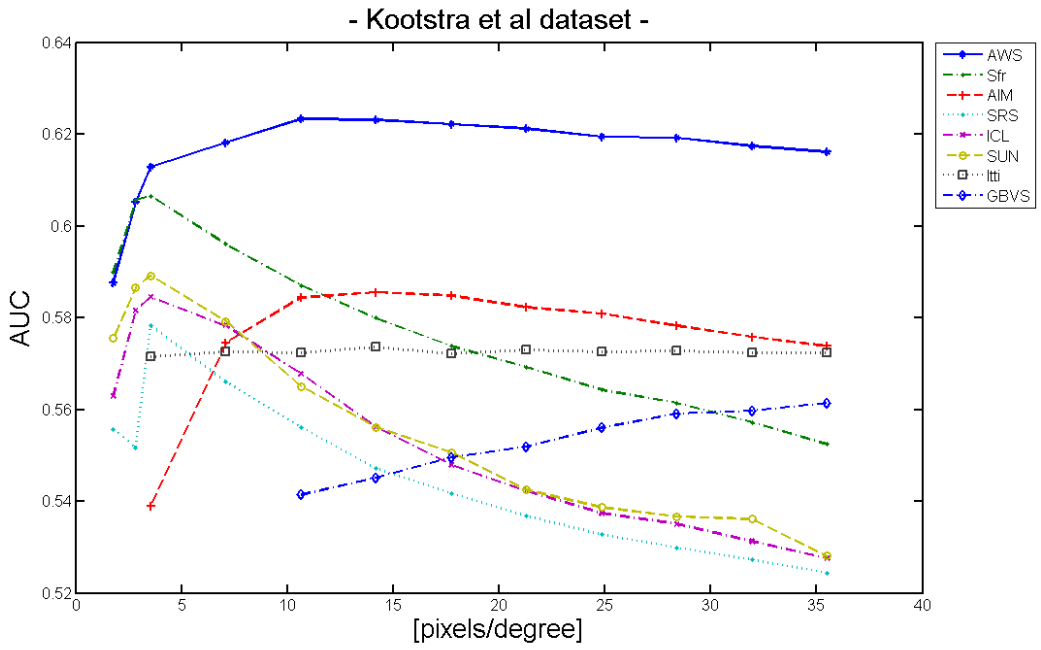

Figure 4.14: Comparison of the capability of different models to predict human fixations –measured through AUC values from ROC analysis, as explained in the first section–against the spatial resolution retained in the input image. Spatial resolution is expressed as pixels by degree of visual field observed by subjects. Results are shown for the dataset of Kootstra et al.

spatial resolution, instead of the *downsampling factor*, like the relevant parameter to keep constant across datasets. This difference is not excesive for the AWS model. For the sake of clarity and a fair comparison we have used the constancy of the downsampling factor. However, also the performance of the SUN model in the Kootstra et al. dataset has been considerably underestimated due to this issue. As well most of the models would have probably favored from a fixed value of spatial resolution rather than any other parameter. An additional problem with models that force the dimensions of the input image to a fixed *square* value of 64x64 is that they do not respect the proportions of the horizontal and vertical dimensions and then they produce a geometric deformation of the input image, of a variable amount that depends of the shape of the input image.

Anyway, the point here more than comparing performance is to compare the robustness of the models against variations in the spatial resolution of the input image. In this respect, the model by Itti et al. and in less extent the AIM model by Bruce and Tsotsos show a considerable stability in their

behavior against spatial resolution of the input image. All of the other models show an wild lack of robustness to manage with spatial resolution, revealing strong biases and a very rigid design. This is most probably related to design choices like the definition of fixed sizes for the receptive fields and their surround, or the definition of fixed ranges of spatial frequencies to compute saliency.

Otherwise, as shown in the graphics of the figures 4.13 and 4.14 the maximum of the predictive capability of the AWS model is not achieved for the maximum spatial resolution, but for a spatial resolution clearly lower, of around 10 pixels/degree for both datasets. This fact suggests the existence of a threshold of the visual acuity that is able to affect intersubject consistency in the spatial distribution of fixations. That is, for subjects with a lack of visual acuity that do not put them under such value, fixation patterns will retain the same consistency present between healthy subjects with normal visual acuity. Therefore, the hypothesis of adaptive whitening in the HVS seems to predict a strong robustness in the consistency between subjects in spite of important variations of visual acuity, that is of important variations in the spatial characteristics of the visual window. However further analysis is needed in this respect to determine such a threshold. In particular, it would be worth using a selection of biased images in which saliency is expected to be driven by small scales to find how small a scale can be to affect the spatial pattern of human fixations in free surveillance, in order to determine the value of such threshold of visual acuity.

4.A. Appendix

Other measures that have been previously used for the comparison of saliency with the spatial distribution of human fixations have been cosidered. They do not reveal remarkable differences with the analysis done on the selected measure based on ROC analysis. Consequently, they have not been included in the previous discussion for the sake of clarity.

However, for informative reasons we consider convenient to show in this appendix some results for two of these comparative measures that are based on the Kullback-Leibler (KL) divergence and that have been previously used in a number of works with the same purpose. The Kullback-Leibler divergence provides a measure of difference between probability distributions. It is not a distance, since it is not symmetric. It gives a measure of how much additional information is needed for a given distribution to equate the other. Thereby, the higher the KL divergence between two distributions is, the higher is the difference between those distributions.

Table 4.4: KL values obtained with different models of saliency for both of the datasets of Bruce and Tsotsos and Kootstra et al. Standard errors range 0.001-0.002 for both of the datasets. Higher is better.

Model	Bruce and Tsotsos dataset	Kootstra et al. dataset					
		Whole dataset	Buildings	Nature	Animals	Flowers	Street
AWS	0.4625	0.1425	0.1576	0.0830	0.2709	0.2049	0.4433
Seo and Milanfar	0.4121	0.1019	0.1592	0.0712	0.2521	0.0818	0.4053
Hou and Zhang	0.3846	0.0672	0.1107	0.0498	0.1454	0.0707	0.2478
AIM	0.3198	0.0797	0.1007	0.0654	0.1507	0.1225	0.2848
SUN	0.2934	0.0562	0.0664	0.0480	0.0712	0.1402	0.2501
Itti et al.	0.2524	0.0549	0.0994	0.0457	0.1568	0.0356	0.2885

Its formal expression is given by:

$$D_{KL}(D1, D2) = \sum_{i=1}^{N} D1_i \cdot \log \frac{D1_i}{D2_i} \tag{4.1}$$

where $D1$ is typically taken as the true probability distribution, $D2$ is taken as the distribution aiming to approximate $D1$, and N is the number of discrete values of the independent variable used for both distributions and it must be the same. If the base of the logarithm is 2, the KL divergence between the distributions is given in bits. It can be interpreted hence as the additional amount of information needed to predict $D1$ from the knowledge of $D2$.

Capability of predicting fixations

One of the typical uses of the Kullback-Leibler divergence consists in comparing the distribution of saliency values in fixated points against the distribution of saliency values in non-fixated points. It is expected that the higher is this value, the more discriminative is the measure of saliency when it is used to decide if a point in an image is to be fixated or not.

This measure still makes a direct use of fixation positions in the definition of the comparing distributions. Therefore, it can be viewed as a direct measure of the capability of saliency to predict the spatial distribution of fixations, that is, to discriminate between fixated and non-fixated points.

The formal expression is given by equation 4.1, being $D1$ the histogram of fixated points in an image and $D2$ the histogram of the same number of random non-fixated points from the same image. Like in the previous sections, center bias is avoided through the same shuffling procedure in which non-fixated points are selected from points fixated in a different image of the dataset. Besides this procedure delivers an estimation of the standard error associated to the measure.

The results are provided in the table 4.5. They support the same discussion done in the first section of this chapter for results from ROC analysis. The only minor remark that can be added is that the distances between models are amplified with this measure, making the advantage of AWS more apparent.

Comparison between spatial distributions of saliency and human priority

Other use that has been employed by Le Meur et al. relies on the interpretation of a priority map as a measure of the probability of each point to attract gaze, and the interpretation of a saliency map as a prediction of that probability [MCBT06]. From this viewpoint, it makes full sense to compare both distributions through the KL divergence.

It is worth noting that, instead of comparing gray level probabilities of fixated and non fixated points like in the previous section, the comparison is done now between distributions of probabilities in the space. This introduces an important difference related to the use of a derived –rather than direct–measure of human behavior, since priority is obtained through a particular processing from obtained human fixations. For each of the datasets we have used the average priority maps provided by the authors that have been obtained through different procedures, as previously explained in this chapter.

With this aim, we obtain probability maps simply dividing a given map by the sum of its gray level values. We denote by $h_i = h(x, y)$ and $m_i = m(x, y)$ the priority map deived from fixations and the saliency map from each model respectively –taken as probability distributions. Modifying correspondingly the expression 4.1, the KL divergence is now given by:

$$D_{KL}(h, m) = \sum_{i=1}^{N} h_i \cdot \log \frac{h_i}{m_i} \tag{4.2}$$

where N is the number of pixels, and the pixel index is the independent variable i. The pixel index is a unidimensional variable since the pixel order

or its relative position are not relevant whenever being the same for the compared distributions.

The results obtained on the Bruce and Tsotsos dataset are provided, for several state-of-the-art models, in the table 4.5. The AWS model achieves again the best value, since it shows the least difference value, pointing to a spatial distribution closer to the priority map. The model of Seo and Milanfar achieves again the second best value. The remaining models change their relative positions, comparing to the results obtained with a ROC analysis or a Kullback-Leibler comparison of fixated and non-fixated points.

Table 4.5: KL values obtained with different models of saliency for both of the datasets of Bruce and Tsotsos and Kootstra et al. Standard errors range 0.001-0.002 for both of the datasets. Lower is better.

Model	Bruce and Tsotsos dataset
AWS	0.8161 ± 0.0211
Seo and Milanfar	0.8659 ± 0.0257
Hou and Zhang	1.1185 ± 0.0383
AIM	1.0868 ± 0.0254
SUN	1.0082 ± 0.0251
Itti et al.	0.9965 ± 0.0249

However, several problems of this procedure put in question its validity. Firstly, it is an indirect procedure since it uses comparison with priority maps and not with fixations. this use is different from the use proposed in the third section of this chapter, since there the predictive capability of priority maps was used as a benchmark *zero-point* value but the comparison of saliency maps was done directly with fixations. Moreover, here there is no shuffling procedure but a KL divergence is obtained for each of the images, and standard error is computed dividing the standard deviation by the square root of the number of images (i.e. the number of averaged values). The distribution of values is not normal, making difficult to interpret the meaning of such stardard error. In agreement with the observation previously done, values vary wildly from one image to another and if the standard deviation is used as uncertainty instead of the standard error of the mean, it increases an order of magnitude up to the point to do nearly all of the models statistically compatible.

Therefore, at least in the described form this measure seems to be useless. However, its focus on the comparison of distributions in the space makes it interesting enough to take it into account, and it possibly would improve if

using a shuffling procedure to compute it in a whole set of images.

Chapter 5

Reproduction of Psychophysical Results

The AWS model is able to reproduce a series of psychophysical results, usually associated to saliency. As explained in the chapter 2, a variety of psychophysical studies have been devoted to characterize the behavior of visual attention, as well as more specifically the perception of visual saliency.

The beating of the FIT proposed by Treissman (see chapter 1) can be found underlying most of these studies. This is the case of the non-linear effects observed for certain features, arising from its parallel processing in early vision. Likewise, the search for asymmetries to discard features as simple ones is also a tool to advance in the research project proposed by the FIT. Of course, a straightforward strategy is the characterization of search efficiency, seeking for the separation of simple features able to produce pop-out from those compound features that produce a serial and inefficient search. Otherwise, saliency has also been related to phenomena of perceptual amplifying that are in the basis of some visual illusions. All of these studies are here reviewed in the light of the adaptive whitening hypothesis in early visual coding, without the need of further hypothesis of parallel processing of *primitives* nor particular phenomena of perceptual amplifying.

The chapter starts showing the results of two experiments based on perceptual comparisons. They seem to be closely related to early vision and many of them also to the perceived bottom-up saliency. Likewise, they do not involve occulomotor performance, and thus they seem in principle less exposed to perturbations from motoric (non visual) neural functioning. In the first experiment, an accurate explanation for the illusion of increasing luminance with increasing corner angle is conveyed, while the second is related with the non-linearity of saliency against orientation contrast. Next, the AWS model will be shown to suitably reproduce several phenomena related

to visual search. Namely, Weber's law and pressence/absence asymmetry, color asymmetry, and efficient and inefficient search results are shown to be correctly reproduced by the model.

5.1. Experiments based on perceptual comparisons

In this section results of two experiments that are based on the comparison of the target stimulus -of constant luminance- with a luminance value of other reference stimulus are considered. These experiments are conceived to quantify the saliency of the target stimulus by means of the luminance value of the reference that provides the equilibrium in the comparison. The change in this value is assumed to arise from the difference in the visual saliency of the target that depends on its relation to the context.

5.1.1. Linearity against corner angle

Here an experiment inspired in a series of Vasarely's op-art works devoted to nested squares is reproduced and explained in the light of the AWS. With the aim to characterize and explain the illusion of higher luminance in the diagonals of nested squares, as well as other related visual illusions, Troncoso et al. have studied the saliency of a corner in a gray scale gradient. [TMMC05]

They measured saliency as a function of corner angle. To do it, they used seven images of different corner angles, with the middle point of the gradient within the corner, always with the same luminance. Six of those images can be seen in figure 5.1. They asked observers to compare the intensity at that central point of the stimulus, with a standard stimulus. Such standard stimulus was made of a vertical stripe with 55 segments with a different luminance value. The order of segments was varied so that any segment had the same probability to appear at the same height than the central point of the corner. Given that the physical luminance of the central point was the same for all of the corners, differences in the luminance chosen in the standard stripe were attributed to an illusory enhancement, due to a different magnitude of saliency. The results obtained revealed that saliency decreases linearly with corner angle.

The authors tried an explanation of such behavior with basis on center-surround differences. They measured the responses of a DoG filter for all of the corners in the central point evaluated by observers. They succeded to explain the trend to decrease of saliency, but not the linearity observed. They stated that the results pointed to a kind of center-surround competition,

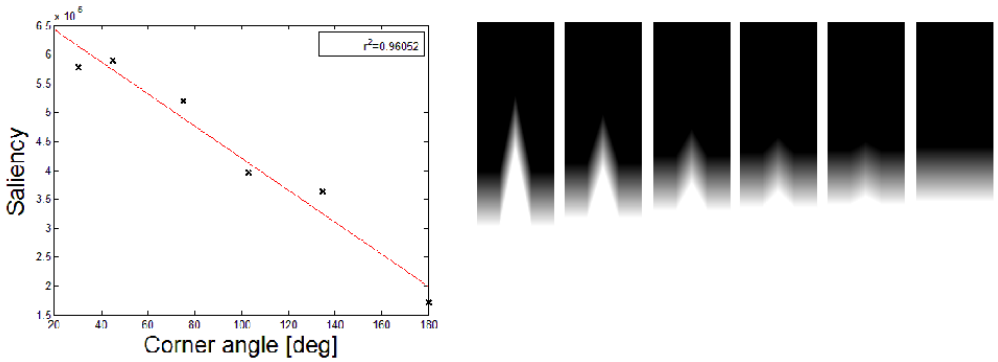

Figure 5.1: Saliency against corner angle and the six images used, obtained from [TMMC05].

and hypothesized two possibilities to explain the linear behavior obtained. Namely, a non-linear component in filtering, or the intervention of mechanisms different from center-surround differences.

In figure 5.1 the results obtained with the AWS model are shown. The saliency measured by the model decreases with corner angle, for 6 corner angles $(30°, 45°, 75°, 105°, 135°, 180°)$. This result is in fair agreement with the reported linear behavior of humans. Saliency for an additional corner of 15° used by Troncoso et al. was clearly underestimated by the model, and has not been used for linear fitting. Other models tested by us have not been able to reproduce such behavior, and to our known AWS is the only one to claim it.

5.1.2. Non-linearity against orientation contrast

The next experiment selected is already a classic in attention and saliency literature. It is related to the observed non-linearity of saliency against feature contrast that will be tackled again in the last section of the chapter. Here the feature on focus is orientation.

Nothdurft has shown that the saliency of a target stimulus as a function of orientation contrast, perceived by humans, varies in a non-linear manner [Not93]. It varies from an starting threshold value, increasing rapidly at the beginning, up to a nearly constant saturation value. The experiment here reproduced consisted in the observation of images with a target stimulus of variable orientation, among an homogeneous background of equally oriented stimuli. Four example images similar to those used in the experiment are shown in figure 5.2. To measure the saliency of the oriented target, the images included one additional target of variable luminance, embedded in the array

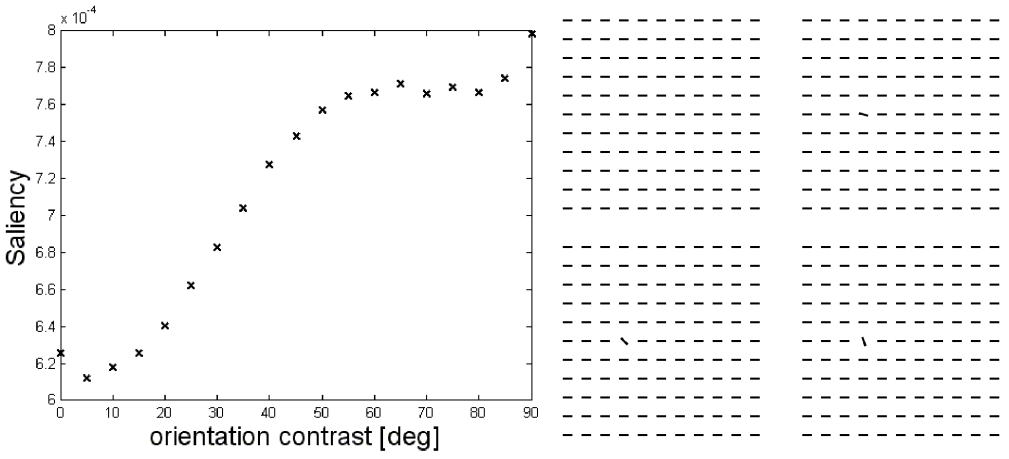

Figure 5.2: Obtained saliency against orientation contrast of the target and four examples of the images used (Images adapted from [Not93]).

of background stimuli. This luminance target had the same orientation and characteristics, except luminance, that the background stimuli had. Positions of both orientation contrast and luminance contrast targets were random but in opposite sites of the display. Observers were asked to rate the conspicuity of each of the targets. The measured rating were fitted to hiperbolic tangents, and the luminance contrast of equilibrium was determined for each value of orientation contrast. This value of equilibrium was taken as the measure of saliency of the orientation contrast target.

In the figure 5.2, a plot of the corresponding measure of saliency against the orientation contrast provided by the AWS model is shown. Clearly, the AWS perfectly matches the non-linear behavior described by Nothdurft for humans. Thereby, the saliency measured by the model increases steeply from a threshold value of orientation contrast of around 20°, up to a saturation value over 50° of orientation contrast. This result has also been reproduced by [GMV08]. However, other state-of-the-art models fail to do it, at least with the setups made public by the authors [BT09, IKN98].

5.2. Reproduction of visual search results

It has been already pointed that the FIT proposed by Treissman has been a fostering theory that provided basic concepts guiding the psychophysical research devoted to visual attention and early visual processing. Two main topics tackled by Treissman as well as many other authors are related to visual search asymmetries and to search efficiency.

The underlying goal was most often related to the identification of simple features -or *integral* features, as denoted by Treissman in her seminal paper- that would have a parallel processing and thus would guide the deployment of attention [TG80]. A great deal of psychophysical experiments has been devoted to this task. Otherwise, many of these results are being reproduced and explained by models of saliency or in general of early coding, that do not use feature *channels, parallel integral features* or a *subset* of low level *primitives*, like edges, conectedness, and many other features supposed to have an idependent and separable processing in the visual system. This is particularly important in the context of the AWS model, defined in terms of a computational scheme from few simple optical dimensions that characterize an image.

Many psychophysical studies have tackled the study of the asymmetric behavior shown by subjects in visual search tasks. In these kind of experiments, a couple of stimuli differing in a simple feature are used alternatively as target and as distractor. Search latencies are measured for both cofigurations and asymmetry is reporter wether a different value is found depending on which is the target and which is the distractor. Actually, this term encompasses phenomena of very different nature. It has been pointed in many cases that the name itself is not suitable. The reason is that the underlying assumption of symmetric design of the experiment is wrong. Two different examples of search asymmetries, the first related to the presence and absence of stimuli and the second related to the change of color of stimuli in different backgrounds are reproduced and analyzed in the light of the AWS model, and compared to the results provided by other models.

Wolfe and Horowitz have reviewed in depth the results reported by visual search studies to provide a classification of the differents studied features in several groups, in function of the evidence suggesting that they guide or not the deployment of visual attention. With basis on that review, the ability of the AWS model to reproduce the efficient and inefficient behavior of humans for a variety of features is examined at the end of the chapter.

5.2.1. Weber's law and presence/absence asymmetry

A classical example of search asymmetry is the presence/absence asymmetry, observed for a pair of stimuli differing only in the presence or absence of a given simple feature, and analyzed in detail by Treisman and Souther [TS85]. In this assymetry, while a target presenting that feature surrounded by distractors lacking it causes a clear pop-up phenomenon, the reverse target-distractor distribution does not. The interpretation of such behavior in terms of the FIT is that presence of the feature activates its fea-

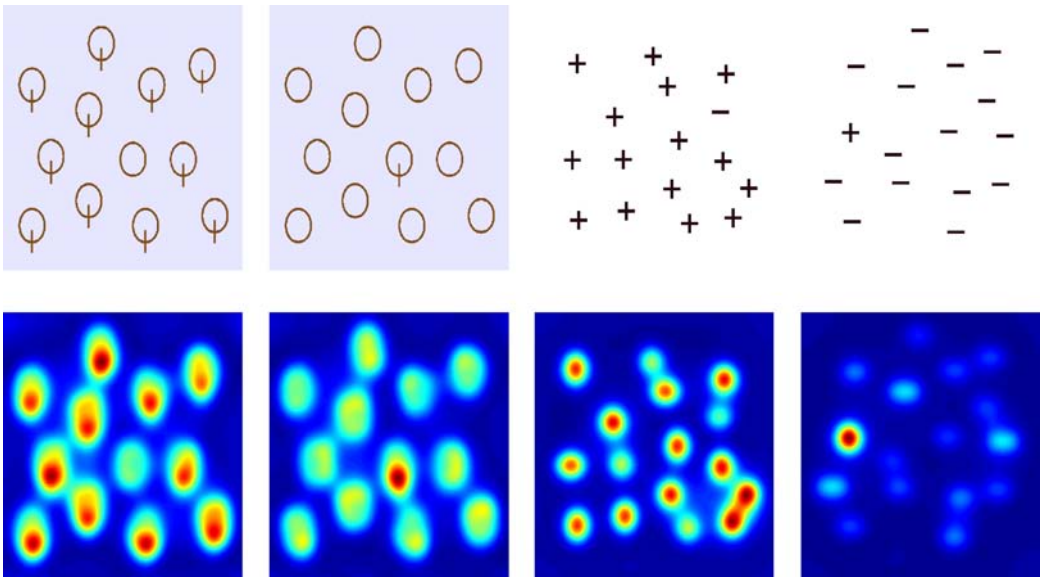

Figure 5.3: Two typical examples of the so called presence-absence asymmetry.

ture map in the corresponding location guiding attention without the need for a serial examination of the image. However the absence of the feature does not produce any particular activation, so that subjects need to examine each position of the image to check if the feature is present or absent.

In the figure 5.3 it can be seen how the AWS manages to reproduce this behavior. The images as well as the saliency maps obtained for two typical examples are shown, namely: the plus and minus symbols, and a circle with and without a bar. Clearly the presence condition targets receive the highest value of saliency, while the absence condition targets are exceeded in saliency by the remaining presence condition distractor. This is not an outstanding behavior of the AWS, since most of the state-of-the-art models of saliency have shown its capability to reproduce it. It is hence more a requirement to accept the state-of-the-art plausibility of new models of saliency.

In a subsequent study, Treisman and Gormican, analyzed in more detail this effect to achieve a quantitative characterization [TG88]. To do it, they modified a very simple dimension: stimulus length. Therefore they measured latencies for different lengths of target and distractor. They found that for a given ratio between the length of the target and the difference of length between target and distractor search time was always the same. As well, the variation of search time against this ratio was linear. Again in terms of the FIT, this linear behavior is easily explained by an increase of location activation of the corresponding feature map proportional to the relative strength of the feature in the location.

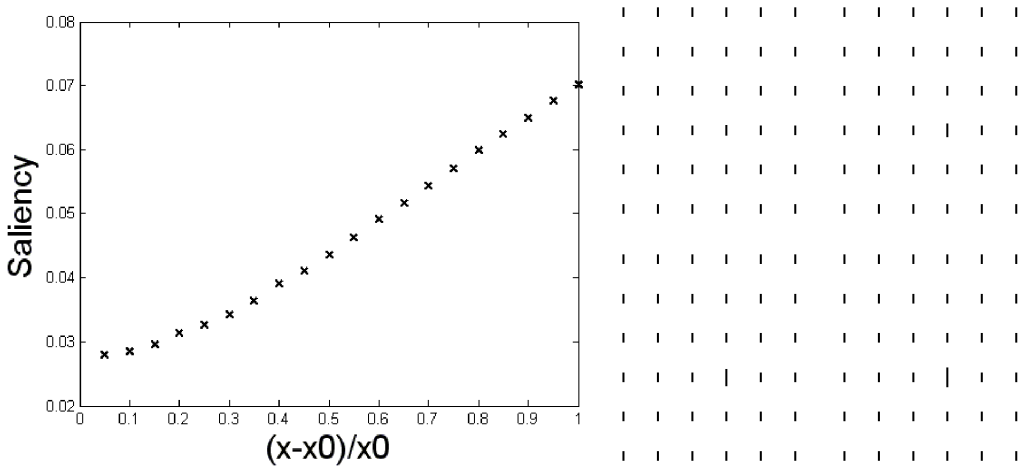

Figure 5.4: Left: Saliency against relative variation of length reproduces the Weber's law observed in humans. Right: Four examples of the images used in the experiment.

To check the behavior of the AWS model, the saliency maps have been obtained for 20 images with different relative values of length between target and distractors. The resulting plot of saliency against relative increase in length is shown in figure 5.4. The behavior of the AWS maps is clearly linear against the relative enlargement in one dimension. Thereby, the behavior shown by humans is reproduced without difficulties by the model. To our knowledge only the center-surround discriminant approach proposed by Gao and Vasconcelos has also been able to reproduce this result before [GMV08].

5.2.2. Color search asymmetry and influence of background

Search asymmetries have also been observed in experiments on the effects of color in visual search. Traditionally, psychophysical studies on color made use of the combination of colored stimuli in a grey -*achromatic*- background. Therefore, results were analyzed in the light of the color properties of the stimuli, without a particular concern aboout the influence of background.

Rosenholtz et al. have pointed a problem of such approaches. The color grey has its own place in a color model, so that color stimuli will present different chromatic distances to background. These distances are also expected to be affected by the area covered by stimuli and background. They focused their study in the so called color search asymmetries, providing a detailed characterization of the influence of background [RNB04].

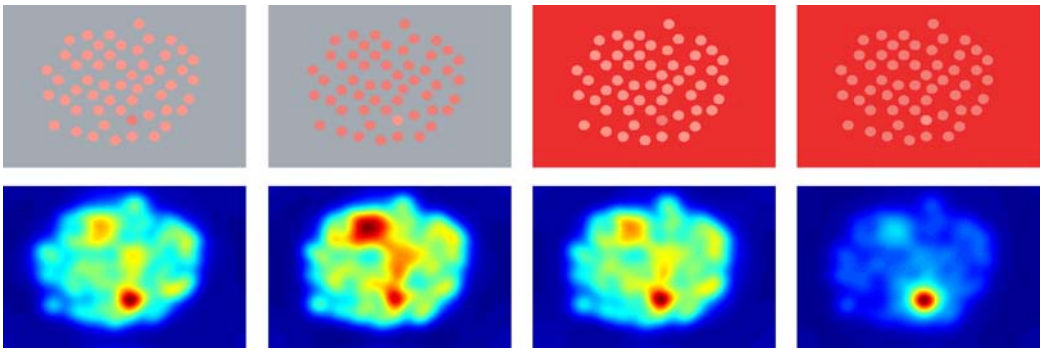

Figure 5.5: Color search asymmetry and its reversal by a change in the background color. Images adapted from [RNB04]

Again, given a pair of stimuli, now with the same luminance and differing only in one color coordinate (in a MacLeod and Boynton color space), it is found that the stimuli exhibit different search times depending on which is the target and which is the distractor. Nevertheless, as Rosenholtz et al. showed, the background influences this effect, to the point of reversing it. For instance, with a gray background, a redder stimulus is more salient than a less red one. However, if background is red, then the redder stimulus becomes less salient. Likewise, with stimuli of fixed size Rosenholtz et al. showed that search latencies were correlated with the Mahalanobis distance of stimuli in the color space. As previously mentioned in chapter 2 when dealing with color coding, this result is in agreement with a whitened representation of color as proposed here. Otherwise, Rosenholtz et al did not tackle how spatial saliency can be combined with such definition of color saliency.

In the figure 5.5 one example of the images used by Rosenholtz et al. [RNB04] is shown. As well, the saliency maps provided by the AWS model are shown. Both the described asymmetry and its reversal by a change of background are well reproduced by the resulting saliency maps. The redder stimulus achieves a higher relative saliency on a grey background, while the less red stimulus achieves a higher relative saliency on a red background. To our knowledge, the only generic (i.e. tested on natural images) model of saliency that have achieved shown to reproduce these results is the AIM model by Bruce and Tsotsos [BT09]. However they employed displays quite different from those used by Rosenholtz et al., while here reproductions of the original displays are used.

5.2.3. Efficient and inefficient visual search phenomena

An unavoidable assessment for a model of saliency relies on its ability to reproduce a series of pop-out phenomena and, more broadly, phenomena related to efficient and inefficient search. Most of recent state-of-the-art models demonstrated their ability to do it. A detailed account of the main phenomea related to visual search efficiency shown by humans and the probability of different features to guide attention has been done by Wolfe and Horowitz [WH04].

Regarding the AWS model, it suitably reproduces a variety of pop-out phenomena related to orientation, color, size or texture, widely reported in literature. The figure 5.6 demonstrate this statement. It shows different images -reproduced from popular references- with singletons of color, orientation, size or texture, as well as the saliency maps produced by the AWS model. The pop-out of these singletons is clearly captured by the AWS maps, that ascribes a higher value of saliency to each of them, comparing to other stimuli that do not pop-out.

Besides, the figure 5.6 also shows the behavior of the model in typical situations where humans perform an inefficient search. The corresponding saliency maps allow a saliency-based explanation for typical cases like a unique closed circle surrounded by randomly oriented open curves or a cross surrounded by similarly oriented intersections, which do not pop-out and undergo an inefficient search. In all of these cases the corresponding stimuli have a value of saliency equaled or exceeded by the surrounding stimuli.

Other phenomena widely studied in visual search experiments are related to target-distractor similarity and to distractor heterogeneity. These works appear to be closely related to studies on feature contrast, already considered in the previous section in relation to orientation contrast and its luminance contrast equivalent. Here the reference is not a value of luminance, but a search time. Like in the previous visual search studies, lower search times are associated with higher saliency. The observed behavior is a non-linearity against both target-distractor similarity and distractor heterogeneity, in coherence with non-linear behavior against feature contrast -for certain features- in experiments based on perceptual comparisons.

In the figure 5.7 one typical example of color similarity adapted from [WH04] and two typical examples of distractor heterogeneity are shown. Saliency maps catch well the non-linear influence of target-distractor color similarity, and from a given difference between target and distractors, saliency does not increase any more. As well, distractor heterogeneity, another important factor that affects the saliency of a color or orientation singleton in human observers, gives place to a similar behavior by the AWS model. To

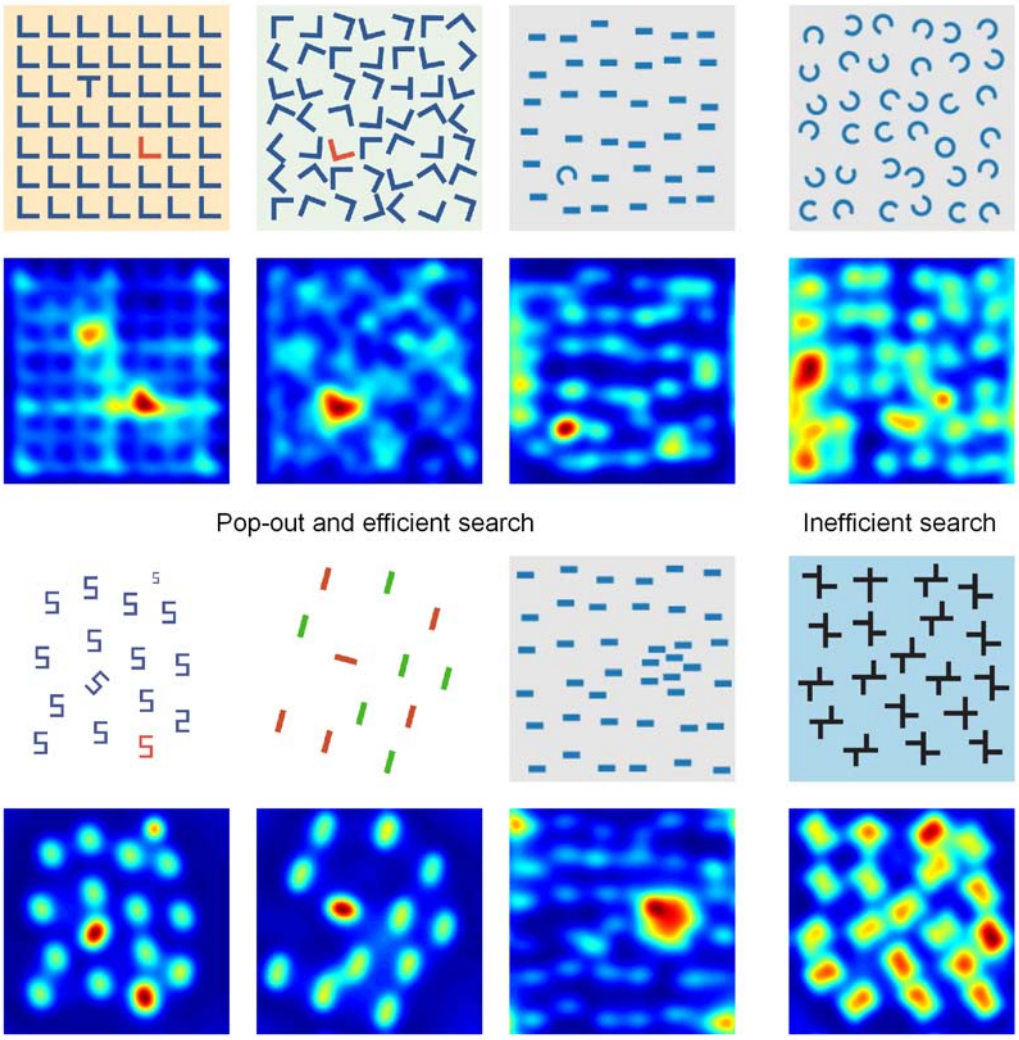

Figure 5.6: Typical examples of pop-out, efficient and inefficient search observed in humans, and reproduced by the AWS. Images adapted from [WH04], [BT09], and [HZ07b].

our knowledge, only the AIM model by Bruce and Tsotsos has shown its ability in reproducing these results.

5.3. Discussion

The early models of saliency assumed in part the use of *primitives* undergoing parallel processing, being edges an outstanding example [Mil93]. This choice was very conditioned by the interpretation of psychophysical results

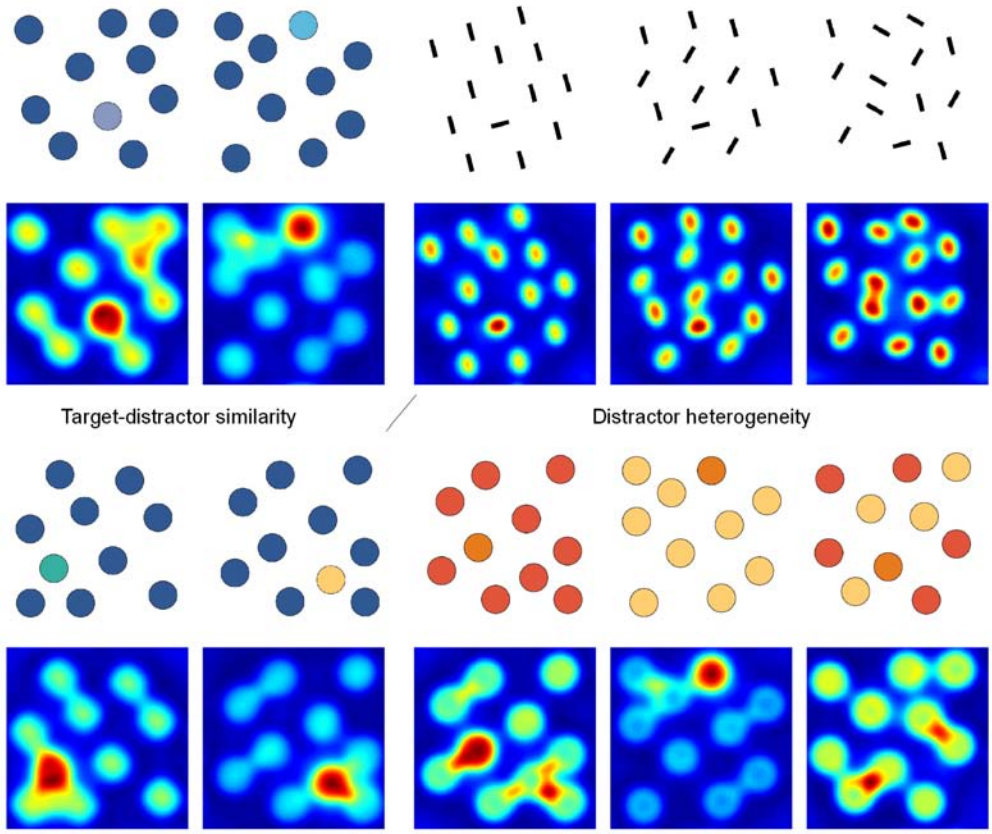

Figure 5.7: AWS matches human behavior against target-distractor similarity and distractor heterogeneity. Images adapted from [WH04] and [BT09].

by the FIT proposed by Treisman et al. In an illustrative passage, at the beginning of a reference work in the field, Treisman and Gormican state that: *Most theorists agree that the early description derives from spatial groupings of a small set of simple primitives that are registered in parallel across the visual field. These primitives, or functional features, need not correspond to simple physical dimensions like wavelength or intensity.* [TG88]

In this dissertation the inverse path has been tried, to explain behavior commonly associated to early vision in a simple scheme of recoding of few simple optical dimensions like wavelength, spatial frequency, and intensity. This trend to simplification from the use of such *primitives* found in early approaches to simple computational mechanisms is implicit in the most recent models of saliency and early coding, mostly based on a probabilistic foundation.

Here, the computational link to the original optical dimensions is explicitly formulated in a coherent and comprehensive manner, providing a functional framework for early coding as well as a highly efficient measure of saliency directly derived from this framework.

None of the other state-of-the-art models that we have tested using the code made available by the authors have been able to reproduce the ensemble of psychophysical results here selected. To our knowledge there is none, except AWS, claiming to do it. Otherwise, most of the experiments selected have been used in the validation of one or more state-of-the-art models. The most popular being probably phenomena of pop-out. We also have used an experiment not employed for the validation of any other model of saliency before, related to the linearity of saliency against corner angle.

It is worth noting the importance of reproducing this selection of psychophisical results. After a comparison with eye fixations as done in the previous chapter, it may be tempting to put all the confidence there. However as pointed in the previous chapter, the methods used for quantitative comparison -like the main methods used in literature- have a important advantage and limitation: they are invariant under monotonic transformations of the saliency map. This is not the case for several of the phenomena studied here, particularly linearity with corner angle, non-linearity with orientation contrast and Weber's law. Moreover, the linearity against corner angle, and the behavior against orientation contrast have been observed in experiments that do not involve eye movements, unlike eye-tracking and visual search experiments. This fact reinforces the generality of the effect on visual perception of a measure of saliency able to explain them.

To summarize, it has been shown that the AWS is able to reproduce a wide and representative set of psychophysical phenomena, to our knowledge not reproduced together by any other model before. Moreover, the setup of the model was exactly the same that was used in the prediction of human eye fixations. These facts reinforce the support for the validity of the adaptive whitening approach to the early visual coding and the subsequent computation of bottom-up saliency.

Chapter 6

Model Extensions and Applications

Selection of regions of interest is a problem common to practically all applications involving image analysis. It is obvious that AWS can be extended to other visual functions, incorporating new components like depth or motion. These can be incorporated in the same way as has been done for previous models of saliency, introducing the corresponding feature map in parallel to the decorrelated color components like in [MCB07] or in [FNSH], or using spatio-temporal filtering like in [SM09].

The work shown in this chapter merely exploits the possibilities derived from the formal generality of the AWS. Therefore, it is focused in direct applications that do not require any kind of adjustments or modifications of the model. Firstly, the suitability of resulting maps to extract proto-objects is shown in a qualitative manner. Secondly, the use of the resulting saliency maps to alleviate an interest-point based solution to the problem of scene recognition in robot navigation is tackled. Thirdly, the direct applicability of the model to multispectral and hyperspectral images is demonstrated, with interesting results in the visible spectrum. Finally, the definition and model of saliency described in this dissertation are proposed as the basis for a quantitative quality criterion in sensor fusion of spatial data for visualization.

6.1. Saliency-based segmentation: context and proto-object extraction

The AWS model allows the extraction of proto-objects, in a similar manner to that used with previous models of saliency [WRKP05, WK06, HZ08, SM09]. This ability is very interesting, since it can be useful to reduce the

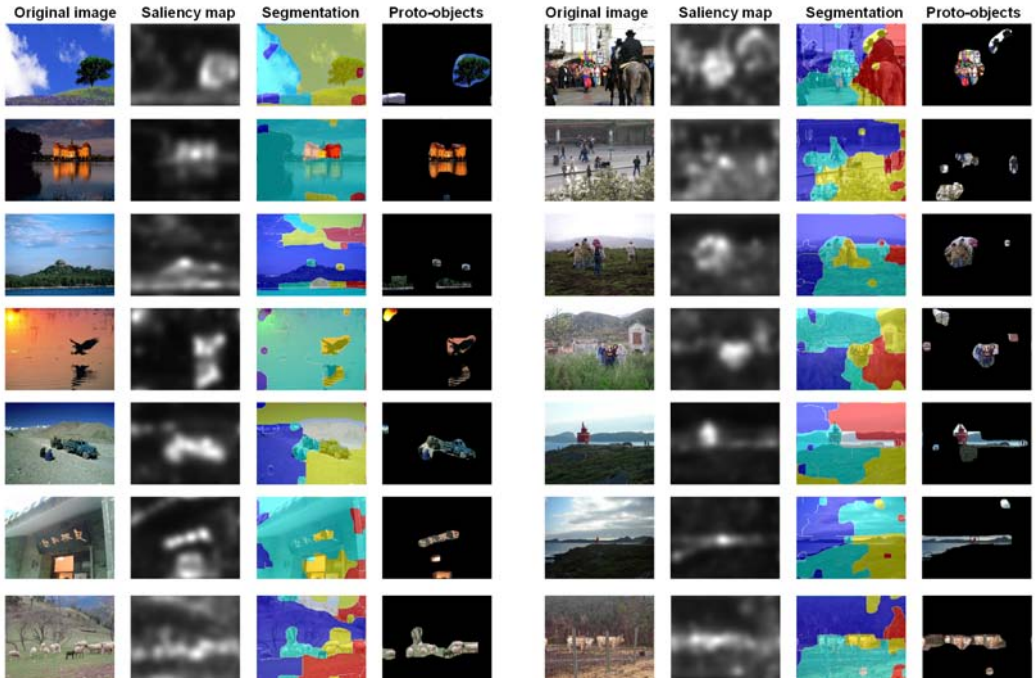

Figure 6.1: Examples of saliency-based segmentation. First six images (left) have been obtained from [HZ08]. The rest are available at http://merlin.dec.usc.es/agarcia/AWSmodel.html.

search space in many visual operations, such as object detection and recognition, unsupervised image classification, or natural landmark generation and detection.

The watershed algorithm has been used to segment images from saliency information, a state-of-the-art general purpose technique. It has the advantage of being parameter free, which eases comparison with other preprocessing approaches. To show the quality of these proto-objects, some results are provided in the figure 6.1 on 14 images with different degrees of clutter and lighting –luminance and color– conditions, as well as different relevant scales and spatial structure. For each image, the regions containing the six highest local maxima have been selected, which delivers six proto-objects.

As can be seen, in general the model extracts some proto-objects that correspond to meaningful objects, or to identifiable parts of them. Also, some salient textures are caught as proto-objects. Besides, examining the segmentation results (third column), the saliency map not only provides information related to salient objects, but also a good basis to make a contextual analysis of the images. This has been shown to facilitate object search tasks and is thought to play an important role in early vision [TOCH06]. Further valuable

information can be found in partial saliencies and oriented conspicuities for a more refined approach. These results give additional support to the validity of the model, as well as to its usefulness for the application in a variety of solutions based on automatic detection and analysis of unknown objects.

It is important to remark that, unlike other models that have tuned their setup or forced the object scale [WRKP05, WK06, HZ08, SM09], here, the exact implementation described in chapter 3 has been used, without any kind of additional bias. Furthermore, it has been used a simple and parameter-free segmentation procedure that delivers results without any kind of special tuning or adaptation. The goal here is to show the usefulness of the bottom-up saliency map in general purpose segmentation, rather than developping any specific approach to a given problem.

6.2. Scene recognition for robot navigation

In this section results are shown that illustrate the usefulness of the model of saliency proposed to improve a scene recognition application by reducing the amount of prototypes needed to carry out the classification task. The application is based on robot-like navigation video sequences taken in an indoor university facility formed by several rooms and halls. The aim of the application is to recognize the different scenarios in order to provide the mobile robot system with general location data. Saliency maps are normalised to the range [0 1]. Scene recognition is firstly performed using invariant local features to characterize the scenarios, and the Nearest Neighbor rule for classification. With regards to the invariant local features, we compare two approaches that currently focus literature attention in this area [MS05, BSP07]. This two approaches are SIFT [Low04] and SURF [BETG08]. Both provide with distinctive image features that are invariant to scale and rotation, and partially invariant to change in illumination and 3D viewpoint.

The scene recognition task is related with the recognition of general scenarios rather than local objects. This approach is useful in many applications such as mobile robot navigation, image retrieval, extraction of contextual information for object recognition, and even to provide access to tourist information using camera phones. In our case, we are interested in recognize a set of different scenarios which are part of university facilities formed by four class rooms and three halls. The final aim is to provide general location data useful for the navigation of a mobile robot system. Scene recognition is commonly performed using local features in images that try to collect enough and distinguishable information to recognize the different scenarios. For this purpose we used SIFT and SURF alternatives to extract invariant

local features.

To compute SIFT and SURF features we used the original code by Lowe and Bay et al. respectively, and the 1-NN rule for classification in all cases, which is a simple classification approach but robust and fast to compute [Low04, BETG08]. For the 1-NN rule, we needed to previously build a database of prototypes to collect the recognition knowledge of the classifier. These prototypes were a set of labeled keypoints obtained from the training frames. The class of the keypoints computed for a specific training frame was that previously assigned to the frame in an off-line supervised labeling process. The entire database was then incorporated into the 1-NN classifier, which uses the Euclidean distance to select the closest prototype to the test keypoint being classified. The class of the test keypoint was then assigned to the class of the closest prototype in the database, and finally, the class of the entire test frame was that of the majority of its keypoints.

With regards to SIFT features, we used the algorithm of Lowe [Low04]. For each key location it assigns an orientation, determined by the peak of a histogram of previously computed neighborhood orientations. Once the orientation, scale, and location of the keypoints have been computed, invariance to these values is achieved by computing the keypoint local feature descriptors relative to them. Local feature descriptors are 128-dimensional vectors obtained from the pre-computed image orientations and gradients around the keypoints.

For SURF features we used the original approach by Bay et al. [BETG08]. They make an efficient use of integral images to speed-up the process. There are two versions: the standard version which uses a descriptor vector of 64 components (SURF-64), and the extended version which uses 128 components (SURF- 128). SURF are partly inspired by SIFT, being the standard version several times faster than SIFT thanks also to a reduction of the number of features that characterize the keypoints [BETG08]. While SIFT uses 128 features, standard SURF only uses 64.

The experimental work consisted in a set of experiments carried out using four video sequences taken in a robot-navigation manner. These video sequences were grabbed in an university area covering several rooms and halls. Sequences were taken at 5 fps collecting a total number of 2,174 frames (7:15 minutes) for the first sequence, 1,986 frames for the second (6:37 minutes), 1,816 frames for the third (6:03 minutes) and 1,753 frames for the fourth (5:50 minutes). First and third sequences were taken in a specific order of halls and rooms: hall-1, room-1, hall-1, room-2, hall-1, room-3, hall-1, hall-2, hall-3, room-4, hall-3, hall-2, hall-1. The second and fourth sequences were grabbed following the opposite order to collect all possible viewpoints of the robot navigation through the facilities. In all the experiments, we used the

Figure 6.2: Salient regions in a frame.

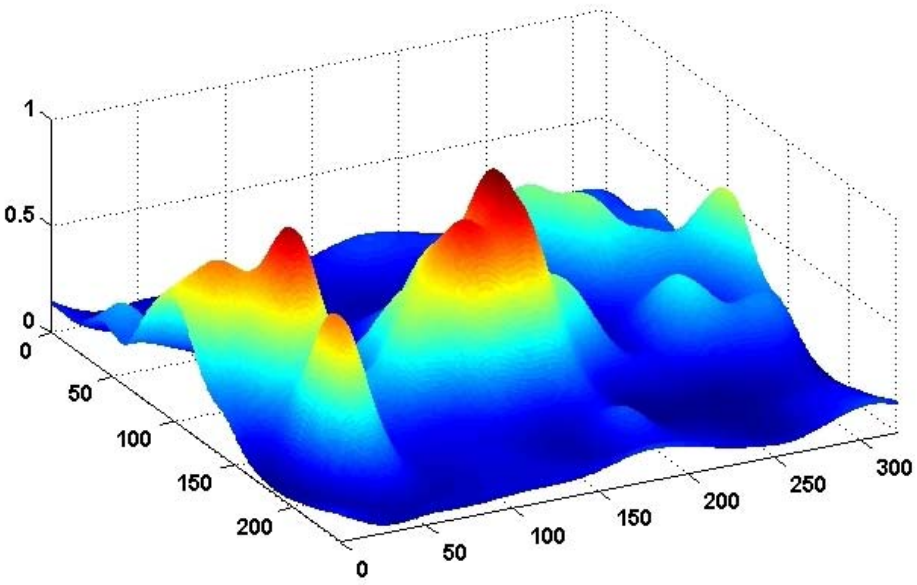

Figure 6.3: 3D contour plot of the saliency map.

first and second sequences for training and the third and fourth for testing.

In the first experiment we computed the SIFT keypoints for all the frames of the training video sequences. Then, we labeled these keypoints with the corresponding frame class: room-1, room-2, room-3, room-4, hall- 1, hall-2 or hall-3. The whole set of labeled keypoints formed itself the database of prototypes to be used by the 1-NN classifier. For each frame of the testing sequences their corresponding SIFT keypoints were computed and classified. The final class for the frame was set to the majority class among its keypoints. This experiment achieved very good performance, 95.3% of correct classification of frames. However, an important drawback was the computational cost of classification, due to the very large size of the database of prototypes which was formed by 1,170,215 samples. In next experiment, we followed the previous steps but using SURF features instead of SIFT. In this case, the recognition results were very bad achieving only 28.2% of recognition performance with SURF-128, and 25.1% using SURF-64, being the size of the database of prototypes of 415; 845.

Although there are well known techniques for NN classifiers to optimize the database of prototypes (e.g. feature selection, feature extraction, condensing, editing) and also for the acceleration of the classification computation (e.g. kd-trees), at this point we are interested in the utility of using the saliency maps derived from the visual attention approach. The idea is to achieve significant reductions of the original database by selecting in each training frame only those keypoints that are included within its saliency map. Also, in the testing frames only those keypoints lying within their corresponding saliency maps will be considered for classification. Once the database is reduced this way, optimizing techniques could be used to achieve further improvements.

In next experiments, we carried out the idea showed in previous paragraph. Nevertheless, we wanted to explore more in-depth the possibilities of using the saliency maps. As it was commented, the saliency measures is set in a range between 0 and 1, thus, we can choose different levels of saliency by simply using thresholds. We will be the least restrictive if we choose a saliency >0.000, and more restrictive if we choose higher levels (e.g. 0.125, 0.250, etc). We planed to use seventh different saliency levels: 0.125, 0.250, 0.375, 0.500, 0.625, 0.750 and 0.875. For each saliency level we carried out the scene recognition experiment achieving the percentage of recognition performance, and the size of the database of prototypes. Results using SIFT and SURF features are shown in Table 6.1 and figures 6.4 and 6.5.

In Table 6.1, S refers to the saliency threshold, Recog. is the % of correct recognition and DB Size is the size of the database of prototypes, given also in %, with regards to the original size, when no saliency maps are used. Only SURF-128 results are shown because the standard version of SURF

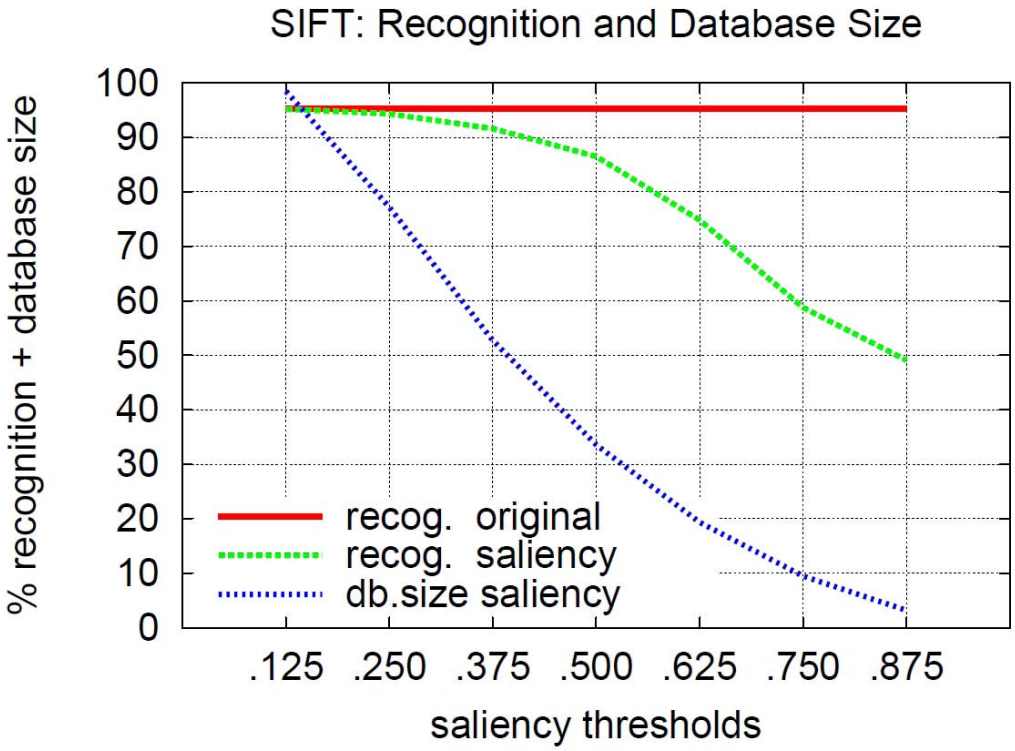

Figure 6.4: Recognition performance and database size given in % for SIFT features.

Table 6.1: SIFT and SURF-128 results on recognition rate and database size in percentage.

	SIFT		SURF	
	Recog.	DB Size	Recog.	DB Size
Original	95.3	100.0	35.1	100.0
S>0.125	95.2	98.5	25.3	99.1
S>0.250	94.3	77.2	51.8	83.5
S>0.375	91.6	52.7	46.9	59.2
S>0.500	86.5	33.6	88.4	37.6
S>0.625	74.8	19.4	64.3	21.4
S>0.750	58.8	9.5	56.6	10.3
S>0.875	49.1	3.2	40.6	3.4

(SURF-64) achieved worse results.

These results show that although SURF features collect significantly less interest points than SIFT features (approximately the half) their performance is not adequate for the scene recognition application. However, SURF features have proven to be adequate, and faster than SIFT features, in other

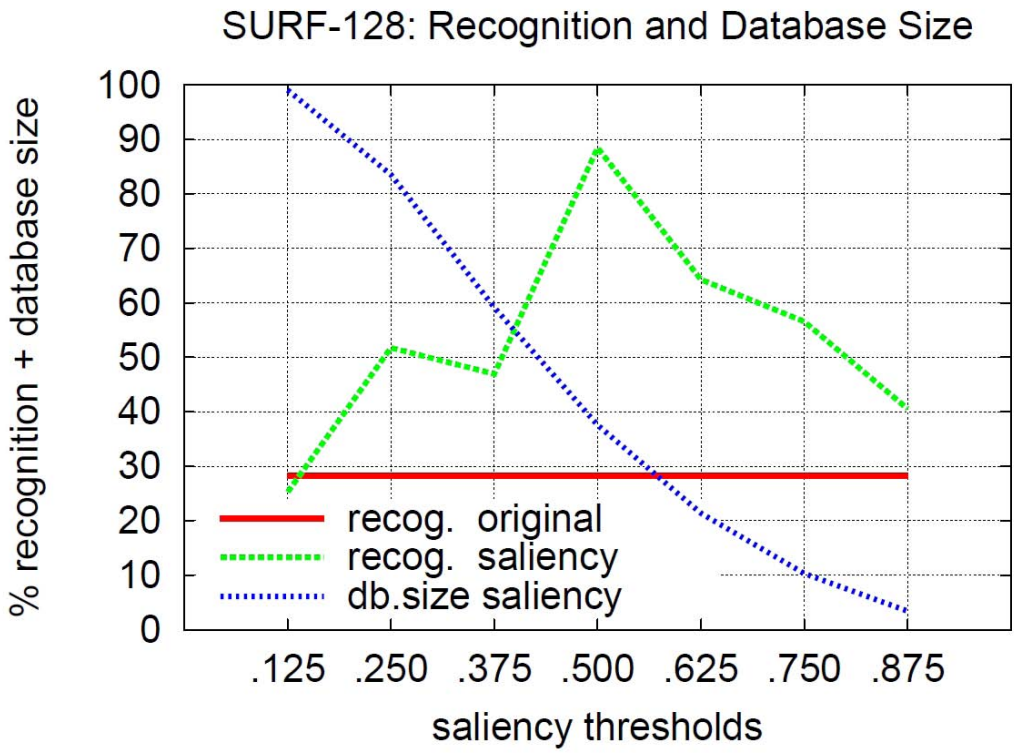

Figure 6.5: Recognition performance and database size given in % for SURF-128 features.

applications [BSP07]. Another interesting result is that the recognition performance of SURF features improves as we use more restrictive saliency maps until a 88.4% peak is reached at saliency level 0.500, then it drops in a similar way than SIFT features (Figure 6.4). This means that SURF features loose distinctiveness as more interest points are used (less restrictive saliency maps), which does not occur in SIFT features, making us to conclude that SIFT features present more distinctiveness than SURF features in very large databases of interest points, as it occurs in the present case.

The best results are achieved using SIFT features, which combined with saliency maps can reduce the amount of prototypes in the database while the recognition performance is held, e.g. saliency level 0.375 in Table 6.1 and Figure 6.4. In this case, the performance drops to 91.6% (only 3.7 points from 95.3%) while the database size is significantly reduced from 1,170,215 to 616,879 prototypes.

6.3. Multi- and hyperspectral saliency

A novel application is proposed here for bottom-up saliency, which does not require to modify our model and retains its efficiency. Slight modifications that could still improve the model will also be discussed. The main underlying idea is that, just as R,G,B color components are whitened, different spectral components can be as well adapted. Again, it has been used exactly the same version of the model that was presented in the chapter 3.

Indeed, similar procedures are common in the analysis of multi- and hyperspectral images, the main application being the reduction of the number of spectral bands. To this end, PCA, ICA, and other procedures have been extensively used [LSL01, WC06]. This reduction of components has an obvious goal: reduction of redundancies and saving of resources. Reduction to three spectral bands is also useful to produce visualizations with pseudo-colors, like for example in [CRHW09]. With that aim, these procedures of dimensionality reduction can be used to extract a number of decorrelated components, and alternatively to establish other reduction criteria, like the percentage of variance to retain when using PCA.

A different approach found is the use of the Karhunen-Loeve transform to obtain a decorrelated representation, without the purpose of reducing the number of spectral bands [HS03]. In a recent work, the combination of this decorrelation procedure with a 2D discrete wavelet transform is evaluated in the frame of the JPEG 2000 standard, for lossy compression of hyperspectral images [PTMO07]. This approach outperforms other assessed approaches and, although it has a different purpose, it has a resemblance with our model: it combines spectral 1D decorrelation with spatial 2D decorrelation.

The AWS model is compatible with both approaches to decorrelation, with and without dimensionality reduction. Computational complexity of spatial decorrelation is linear with the number of spectral bands, but spectral decorrelation has a cubic dependency with them. To alleviate this load when many spectral bands are involved –like in hyperspectral imagery– a low complexity version can be used, like those proposed by [PTMO07] or by [DF08] to decorrelate the spectral bands.

As it is shown in the figures 6.6 and 6.7 with five different examples, the whitened components extracted from the RGB representation span practically the same space that the first three whitened components do from the hyperspectral representation, and consequently saliency from the AWS model is practically the same. When comparing with the saliency that results from a space of 33 whitened spectral components, some differences arise, but not too many. These differences seem to be caused from compression to three components. Indeed, this compression appears to reduce noise, probably ought to

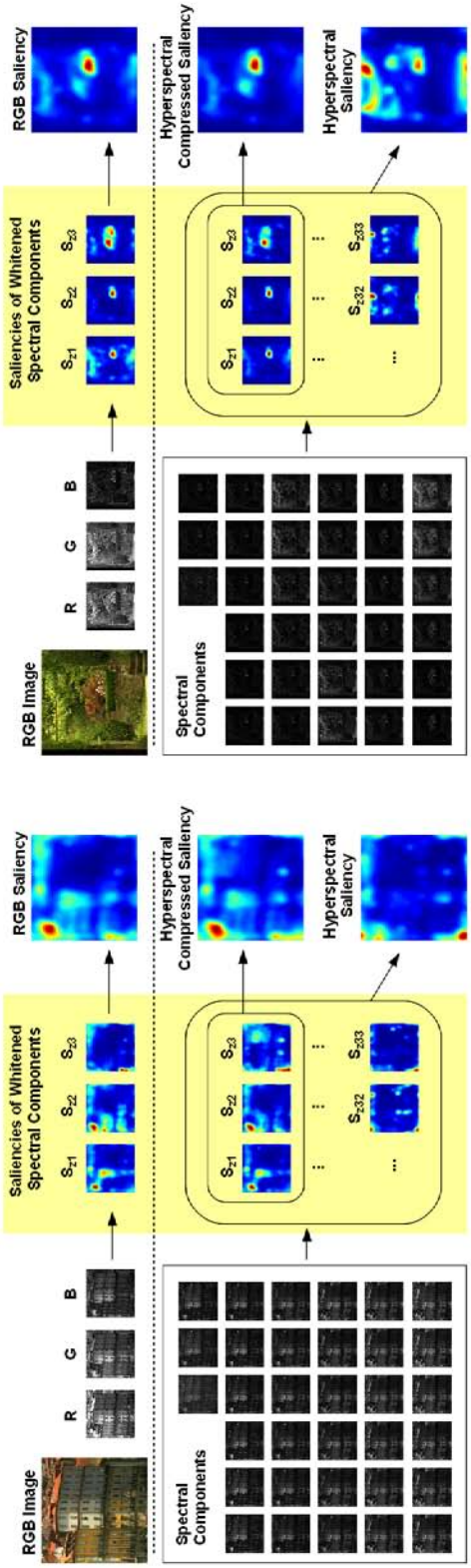

Figure 6.6: Example of saliency computation on two hyperspectral images obtained from [FNA05] with 33 spectral bands in the visible spectrum. Results of saliency from the first 3 whitened components are also shown, as well as saliency from an RGB image of the same scene.

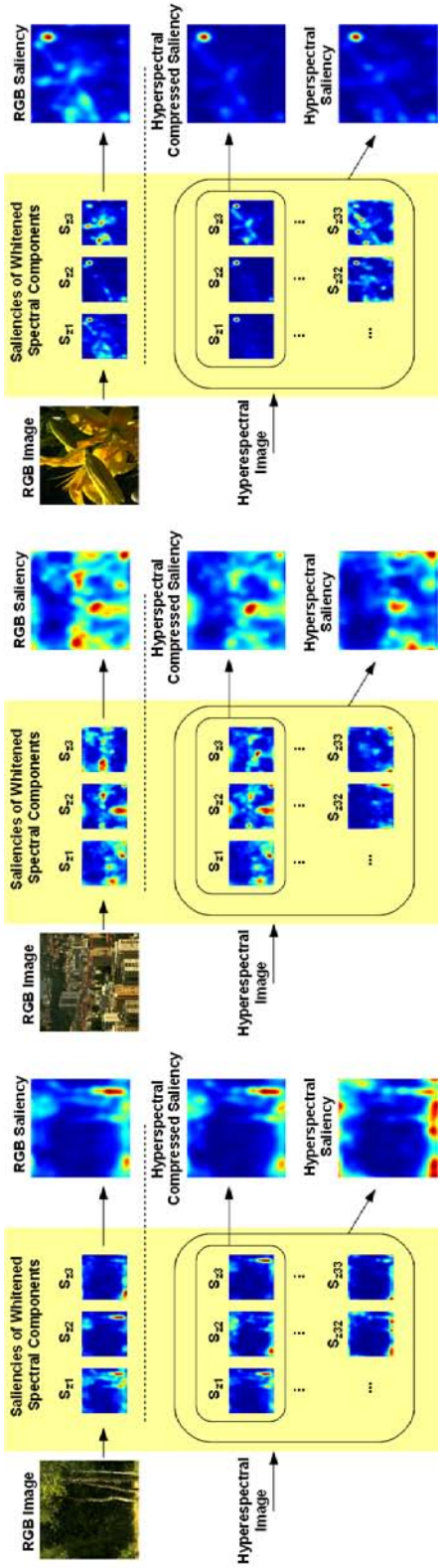

Figure 6.7: Example of saliency computation on three additional hyperspectral images obtained from [FNA05] with 33 spectral bands (omitted) in the visible spectrum. Results of saliency from the first 3 whitened components are also shown, as well as saliency from an RGB image of the same scene.

a reduction of the high redundancy in hyperspectral data. Besides, this is in fair agreement with the proposal of Lee et al., who state that *color opponency may in part be a result of the properties of natural spectra and not solely a consequence of the cone spectral sensitivities* [LWS02]. As a consequence of the adaptive whitening of color components, the HVS would very robustly compute the physical saliency that results from the spectral and the spatial structure of the image. Therefore, this measure of saliency would be highly invariant against details in spectral sensitivities of the sensors involved, be the cones of the HVS or any others.

In an interesting work by Nieves et al. [NPVR08], it has been shown how images from an RGB digital camera were enough to estimate the spectral power distribution of the illuminant. Using a learning-based algorithm they demonstrated that acquiring spectral infomation suitable for both spectral and colorimetric analysis was possible, while avoiding the need of a spectroradiometer. Therefore, this observation points to the sufficiency of a trichromatic representation to code and process all the -redundant- optical information available in natural images. That is, in normal conditions, and in the visible spectrum it makes little sense to manage more information about visual objects, than that is provided by a trichromatic representation from broadband detectors, like the RGB commonly used in machine vision or the LMS detectors of human retina. This is in fair agreement with the observation that saliency measures with an adaptive whitening scheme are highly equivalent when using an RGB representation or the 3 principal components extracted from 33 narrow and non-overlaping spectral bands.

Ultimately, the previous observations support the validity of the approximation done in the chapter 3 when defining visual saliency as closely related to optical variability. Therefore, regarding the formal definition of optical variability formulated, it is legitimate to claim that the measure of relative variability in the visible spectrum is barely affected by the change of the narrow, monochromatic and non-overlaping detectors –tuned to monochromatic sets of plane waves– by three broadband detectors. At least for a measure of the compressed optical variability, that is, for the optical variability enclosed by the first three principal components of the spectrum.

In the figure 6.8 is shown an example of a satellite multispectral IR image of a wildfire, and the obtained saliency map. As can be seen, saliency conveniently captures regions with an outstanding spatial structure through the set of spectral components. Likewise, the few clouds in this terrain scene are salient, but also the most active fire fronts, as well as isolated vegetation regions surrounded by fire. On the other hand, a segmentation of the saliency map, done like in the first section of this chapter, does provide multispectral proto-objects and contextual information. Therefore, segmented

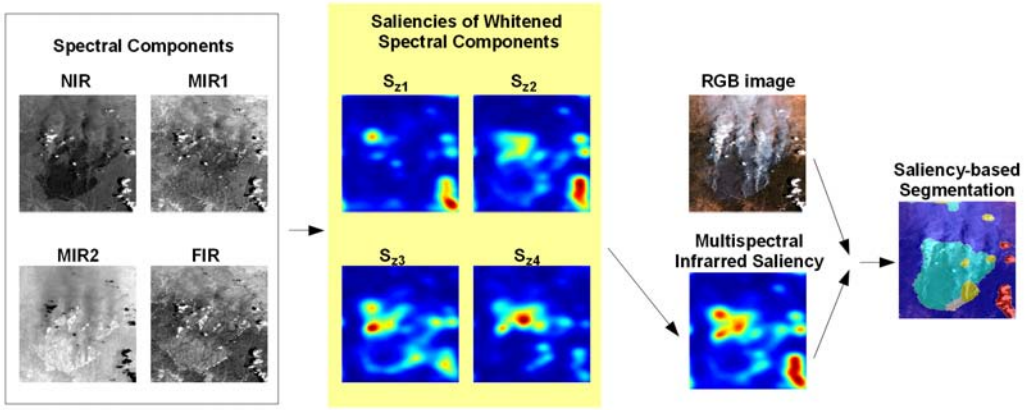

Figure 6.8: Example of saliency computation on a satellite multispectral image with 4 spectral bands in the infrarred: one in the near-infrared (NIR), two in the middle-infrared (MIR) and the fourth in the far-infrared (FIR). The segmentation of the saliency map is also shown superimposed on an RGB image of the same scene. (Downloaded from http://l7downloads.gsfc.nasa.gov/downloadP.html)

regions correspond to burned and unburned areas, fronts of fire, as well as clouds. Hence, it seems to be helpful to apply previous works on visible RGB images to multi- and hyperspectral images, for recognition and selective lossy compression based on saliency. But it could also be useful for anomaly detection, a problem of high interest in multispectral and hyperspectral imagery [CC02]. Additional information for segmentation or classification can be obtained again from whitened band partial saliencies. Indeed, previous specific works dealing with segmentation and classification on hyperespectral images have pointed the need of using spatial –and not only spectral– information [TCB10].

6.4. Saliency-based evaluation of sensor fusion and spatial visualization

In a recent work on visualization of hyperespectral data in RGB displays, Cui et al. observed that *existing methods map spectral samples to unbounded 3-D Euclidean space. After dimension reduction, they all use not only a second nonuniform mapping to color space that creates colorful images but also the illusion of salient features that are not present in the data.* They pointed that this problem arises from the sacrifice of the preservation of spectral distances to take advantage of the dynamic range of the display.

Consequently, they proposed a method to avoid this limitation, that aimed to preserve spectral distance in the projection of the n-dimensional space of n spectral components to the 3-dimensional chromatic space perceived by humans and spanned by a display. However the term salient was only treated in an intuitive -non formal- way. Because of this, no objective measure to evaluate results was provided, but only quality ranking by a set of subjects [CRHW09].

As advanced in chapter 3, the definition of saliency proposed as a measure of optical variability provides a suitable ground for saliency to be translated to and from other physical domains different of the visual window. In fact conservation of relative variability in the space can be used to this end. Indeed, the projection of other physical windows in the visual window under the constraint of conservation of relative variability in the space provides a general and objective criterion to evaluate sensor fusion techniques in a generic way not linked to specific purposes.

Moreover, in the case of spectral components or other kinds of sensors, the proposed measure of saliency takes into account the existence of different constraints in the spatial characteristics of the corresponding physical window. As well, it is robust against different bias in the spatial statistics of different types of sensors and scenes. This is important for instance when dealing with aerial images, with spatial statistics clearly different from normal natural images that allow the use of a single predefined scale for robust matching [Gil98].

To compare the saliency maps of the visualization results to the variability map from the original sensor data, different standard measures for the comparison of probability distributions can be employed, like for example the ROC analysis employed in chapter 4 or different implementations of the Kullback-leibler divergence.

Of course, the quantitative evaluation proposed does not give a straight method to develop visualization techniques, since different approaches to visualization can produce the same *quality* in terms of conservation of relative variability and its translation into visual saliency. However it supports some guidelines for tasks of dimensionality reduction and projection, like decorrelation of the original components and conservation of the maximum amount of variance, that are indeed in good agreement with the main trends in a number of works in this field.

A complete quality evaluation procedure of sensor fusion techniques for spatial visualization -with and without data compression- could combine the proposed measures of preservation of variability in terms of visual saliency, with other existing quantitative methods of image quality. Such a combined measurement would virtually allow the optimization of visualization quality,

while retaining the actual spatial variability present in the physical window sensed.

Conclusions

In this dissertation, an explanation has been given to a variety of phenomena commonly associated to early vision. It has been done in a simple scheme of recoding of few simple optical dimensions like wavelength, spatial frequency, and intensity. It makes a step further in the trend to simplification from the use of the Treisman's *primitives* –as found in early approaches to visual attention –to simple computational mechanisms that is implicit in the most recent models of saliency and early coding, mostly based on a probabilistic foundation. The major contributions of this thesis are the following:

- A novel functional framework for early visual coding has been proposed that is based on a simple mechanism of adaptive whitening and that is biologically plausible. The whitened responses obtained in this framework are shown to be suitable to explain figure-ground segmentation and to reproduce several visual illusions closely related to it, usually associated to contextual adaptation. The initial decomposition used classical receptive fields that have been interpreted in terms of independent components of natural images, as arising from a kind of long-term whitening of responses. Adaptive whitening aims to catch the overall context-driven shift observed in these receptive fields under particular natural stimulations. From our viewpoint, such a simple approach, combining long-term and short-term adaptive whitening, could represent a bridge between local and collective responses observed in the visual cortex. Indeed, this model is compatible with population coding strategies supposed to take place in the cortex, as well as with several existing mechanistic approaches, defined at the level of single neurons. In sum, this functional framework is in agreement, not only with center-surround competition phenomena, but also with the contextual influences that modulate these phenomena.

- From the optical representation of an image as a superposition of plane waves, a definition of optical variability has been derived. In the context of the previous framework for early visual coding, saliency –defined

as a simple modulus–takes an equivalent form, except for the domain of application. Referring the visual domain of physical magnitudes as the optical visual window, both definitions are shown to be closely related. Therefore, it is proposed that saliency encodes the optical variability present in the optical visual window. The proposal is ultimately grounded on the classical efficient coding hypothesys, upheld by Barlow [Bar61, BF89]. Besides, regarding the coding catastrophe underlying many visual illusions and assumed –but not explained–by such hypothesis, this thesis yields an explanatory goal for it: the invariance of the HVS in the perception of saliency. That is, neural adaptation ensures the invariance of the HVS to cope with the optical variability present in the visual window.

- A particular implementation of the model of saliency has been proposed that is simple and light. It outperforms other important models of the state of the art in widely used comparisons with human behavior, while keeping a low computational complexity.

- Regarding the comparison with human fixations, it clearly exceeds the best reported results in two open access datasets with a standard measure based on ROC analysis [TBG05, ZTM$^+$08]. Likewise, we point out a clear incompatibility between the huge variation in the results depending on the used dataset and the very tight values of uncertainty delivered by such procedure. To overcome this problem, a comparison is proposed with the predictive capability shown by humans themselves. Hence, results with two datasets, originally very different, become compatible. With this procedure, it has been found out that AWS shows the same predictive capability than an average human. Moreover, it still clearly outperforms all other models that show an evident lack of robustness –for instance, against salient symmetries or high frequency textures–, most probably associated to different design biases. This measure allows us to hold that in these two datasets, bottom-up saliency completely explains inter-subject consistency. It does it without the need of any top-down mechanism or any other kind of bias, different to the well known center bias. We believe that this appraisal is extensible to other image datasets, under conditions of free surveillance, and in the absence of a high level task.

- The AWS model reproduces a wide variety of relevant psychophysical results related to both covert and overt attention. Some of these assessment procedures are not invariant to monotonic transformations, unlike the measures of comparison with human fixations. Because of

this fact, they pose a further challenge to model saliency. The same battery of tests delivers failures in other state-of-the-art models, using the publicly available code for them. To our knowledge no other model has claimed to be able to reproduce the complete ensemble of tests.

- Regarding the direct applications of the model proposed, its usefulness is shown in the seggregation of proto-objects from context by means of a simple segmentation of the saliency maps. Likewise, its application for the selection of landmarks for scene recognition is demonstrated.

- Additionally, a novel application of bottom-up saliency to multispectral and hyperspectral images is presented. This application is in the line of a number of works in that field that share the need to manage spatial –and not only spectral– information. AWS is to our knowledge the first bio-inspired model of saliency to be applied with this kind of images. From the results with hyperespectral images in the visible spectrum using z-scores in the whitening procedure, it follows an interpretation for the discrete overcomplete representation that springs up from the visual window. It is equivalent to a lossy compression of information, retaining the most of variance in the original data. Moreover, this singular feature of the AWS is applicable to other multisensor spatial representations, as well as to spatial visualizations from sensor fusion techniques, providing this way a tool to quantitatively assess the capability of a given visualization technique of projecting physical variability on the visual window to translate it in visual saliency.

Open paths for future work

An obvious and major direction of future work is related to the extension of the model to dynamic and stereoscopic scenes, in order to reproduce visual saliency in an actually unconstrained visual experience. There are a variety of possibilities to extend the adaptive whitening strategy of short-term adaptation to temporal and depth cues. Therefore a detailed investigation should be still dealt. Even other perceptual modalities like the perception of sound can be good candidates for an adaptive whitening approach to early perceptual coding.

Other important directions for further research, already pointed along this dissertation are the following:

- The implementation of the proposed framework using plausible mechanistic models of neural networks for the computation of whitening

can yield new results and insights in the understanding of early visual coding and many visual illusions.

- The adapted representation proposed has been observed to show a remarkable ability for object discrimination. From this observation we expect very good possibilities for the use of whitened components in a generic object learning and recognition strategy. This has been already done with succes using center-surround features related to models of saliency outperformed by AWS, for instance in [SM10].

- It has been pointed how saliency can be used for the quantitative assessment of sensor fussion techniques for spatial visualization of the responses of non-optical sensors. Since it does not exist, to our knowledge, a generic objective approach to the evaluation of such visualization techniques, further in depth investigation in this respect poses a major interest.

- Although the perception of saliency has been shown to be quite robust against spectral sensitivities and spatial acuity, the limits of such a robustness remain unclear. Correspondingly, several interesting questions remain unanswered. For instance, to which extent could fixation patterns be used to detect alterations of the visual window?. Is it possible to estimate the kind and amount of alteration of the visual window (loss of visual acuity, loss of color sensitivity) from a pattern of fixations using an specifically tuned version of the AWS or other measure of saliency?. To which extent can be suitable eye-tracking data from the observation of natural images, combined with especific modelling of saliency and comparative measures of distribution of fixations, to detect and characterize visual impairments?. Otherwise the interplay between saliency and relevance deserves much more research related to a variety of questions. For instance, are there age-related differences of developmental nature that modify the relative strength of saliency versus relevance in different types of scenes?. We believe that its demonstrated robustness and unbiased behavior makes AWS a suitable measure of saliency to tackle such studies on biological vision.

Bibliography

[ADF10] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, page 73–80, 2010.

[AEWS08] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk. Salient region detection and segmentation. *Computer Vision Systems*, page 66–75, 2008.

[AHES09] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk. Frequency-tuned salient region detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[AL10] T. Avraham and M. Lindenbaum. Esaliency (extended saliency): Meaningful attention using stochastic image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):693–708, 2010.

[ALP06] B. B Averbeck, P. E Latham, and A. Pouget. Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5):358–366, 2006.

[ALR93] J. J Atick, Z. Li, and A. N Redlich. What does post-adaptation color appearance reveal about cortical color representation? *Vision Research*, 33(1):123–129, 1993.

[AMT06] R. A Applegate, J. D Marsack, and L. N Thibos. Metrics of retinal image quality predict visual performance in eyes with 20/17 or better visual acuity. *Optometry and Vision Science*, 83(9):635, 2006.

[Att54] F. Attneave. Some informational aspects of visual perception. *Psychological Review*, 61(3):183–193, 1954.

[AvEO05] C.H. Anderson, D.C. van Essen, and B.A. Olshausen. Directed visual attention and the dynamic control of information flow. *Neurobiology of Attention*, pages 11–17, 2005.

[B+06] C. M Bishop et al. *Pattern recognition and machine learning*. Springer New York, 2006.

[Bar61] H. B. Barlow. *Possible principles underlying the transformation of sensory messages Sensory Communication*. MIT Press, 1961.

[BBK09] E. Birmingham, W. F Bischof, and A. Kingstone. Saliency does not account for fixations to eyes within social scenes. *Vision Research*, 49(24):2992–3000, 2009.

[BETG08] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[BF89] H. B Barlow and P. Foldiak. Adaptation and decorrelation in the cortex. *The Computing Neuron*, page 54–72, 1989.

[BM07] T. J Buschman and E. K Miller. Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, 315(5820):1860, 2007.

[BMwHC08] L. Barrington, T. K Marks, J. Hui wen Hsiao, and G. W Cottrell. NIMBLE: a kernel density model of saccade-based visual memory. *Journal of Vision*, 8(14):17, 2008.

[BS97] A. J Bell and T. J Sejnowski. Theindependent components' of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[BSP07] J. Bauer, N. Sünderhauf, and P. Protzel. Comparing several implementations of two recently published feature detectors. In *Proc. of the International Conference on Intelligent and Autonomous Systems*, 2007.

[BT06a] R. J Baddeley and B. W Tatler. High frequency edges (but not contrast) predict where we fixate: A bayesian system identification analysis. *Vision Research*, 46(18):2824–2833, 2006.

[BT06b] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *Advances in Neural Information Processing Systems (NIPS)*, volume 18, page 155, 2006.

[BT09] N. D.B Bruce and J. K Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):5, 2009.

[CA02] A. Cichocki and S. Amari. *Adaptive blind signal and image processing*. John Wiley and Sons, 2002.

[Cav99] K. R Cave. The FeatureGate model of visual selection. *Psychological Research*, 62(2):182–194, 1999.

[CBM02] J. R Cavanaugh, W. Bair, and J. A Movshon. Selectivity and spatial distribution of signals from the receptive field surround in macaque v1 neurons. *Journal of Neurophysiology*, 88(5):2547, 2002.

[CBVG10] C. Clopath, L. Büsing, E. Vasilaki, and W. Gerstner. Connectivity reflects coding: a model of voltage-based STDP with homeostasis. *Nature Neuroscience*, 13(3):344–352, March 2010.

[CC02] C. I Chang and S. S Chiang. Anomaly detection and classification for hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 40(6):1314–1325, 2002.

[CG03] G. A Carpenter and S. Grossberg. Adaptive resonance theory. *The Handbook of Brain Theory and Neural Networks*, 2:87–90, 2003.

[CGS06] Y. Chen, W. S Geisler, and E. Seidemann. Optimal decoding of correlated neural population responses in the primate visual cortex. *Nature Neuroscience*, 9(11):1412–1420, 2006.

[CRHW09] M. Cui, A. Razdan, J. Hu, and P. Wonka. Interactive hyperspectral image visualization using convex optimization. *IEEE Transactions on Geoscience and Remote Sensing*, 47(6):1673–1684, 2009.

[CS02] M. Corbetta and G. L Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3):201–215, 2002.

[CSP93] J. F. Cardoso, A. Souloumiac, and T. Paris. Blind beamforming for non-Gaussian signals. In *IEE Proceedings on Radar and Signal Processing*, volume 140, page 362–370, 1993.

[CWS+07] C. W.G Clifford, M. A Webster, G. B Stanley, A. A Stocker, A. Kohn, T. O Sharpee, and O. Schwartz. Visual adaptation: neural, psychological and computational aspects. *Vision Research*, 47(25):3125–3131, 2007.

[DF08] Q. Du and J. E Fowler. Low-complexity principal component analysis for hyperspectral image compression. *International Journal of High Performance Computing Applications*, 22(4):438, 2008.

[DK96] K. I. Diamantaras and S. Y. Kung. *Principal component neural networks: theory and applications*. John Wiley & Sons, Inc. New York, NY, USA, 1996.

[DR04] G. Deco and E. T Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, 44(6):621–642, 2004.

[DS10] M. Donk and L. Soesman. Salience is only briefly represented: Evidence from probe-detection performance. *Journal of Experimental Psychology: Human Perception and Performance*, 36(2):286–302, 2010.

[DZ01] G. Deco and J. Zihl. A neurodynamical model of visual attention: Feedback enhancement of spatial resolution in a hierarchical system. *Journal of Computational Neuroscience*, 10(3):231–253, 2001.

[EBK+10] A. S Ecker, P. Berens, G. A Keliris, M. Bethge, N. K Logothetis, and A. S Tolias. Decorrelated neuronal firing in cortical microcircuits. *Science*, 327(5965):584, 2010.

[ESP08] W. Einhäuser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):18, 2008.

[FBR05] S. Frintrop, G. Backer, and E. Rome. Goal-directed search with a top-down modulated computational attention system. In *Pattern Recognition (DAGM Symp.)*, page 117–124, 2005.

[Fie87] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987.

[FJC07] S. Frintrop, P. Jensfelt, and H. Christensen. Simultaneous robot localization and mapping based on a visual attention system. In *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint (WAPCV)*, page 417–430, 2007.

[FKR07] S. Frintrop, M. Klodt, and E. Rome. A real-time visual attention system using integral images. In *Proc. of the Int. Conf. on Computer Vision Systems*, 2007.

[FM06] J. H Fecteau and D. P Munoz. Salience, relevance, and firing: a priority map for target selection. *Trends in Cognitive Sciences*, 10(8):382–390, 2006.

[FNA05] D. H. Foster, S. M.C Nascimento, and K. Amano. Information limits on neural identification of colored surfaces in natural scenes. *Visual Neuroscience*, 21(03):331–336, 2005.

[FNSH] S. Frintrop, A. Nuchter, H. Surmann, and J. Hertzberg. Saliency-based object recognition in 3D data. In *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, volume 3.

[FU08] T. Foulsham and G. Underwood. What can saliency models predict about eye movements? spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, 8(2):6, 2008.

[Gas78] J. D Gaskill. *Linear systems, Fourier transforms, and optics.* Wiley, 1978.

[GHV09] D. Gao, S. Han, and N. Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6):989, 2009.

[Gil98] S. Gilles. *Robust description and matching of images.* University of Oxford, 1998.

[GKG98] J. P Gottlieb, M. Kusunoki, and M. E Goldberg. The representation of visual salience in monkey parietal cortex. *Nature*, 391(6666):481–484, 1998.

[GM04] K. Grill-Spector and R. Malach. The human visual cortex. *Neuroscience*, 27(1):649, 2004.

[GMR97] S. Grossberg, E. Mingolla, and W. D Ross. Visual brain and visual perception: How does the cortex do perceptual grouping? *Trends in Neurosciences*, 20(3):106–111, 1997.

[GMV07] D. Gao, V. Mahadevan, and N. Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, 2007.

[GMV08] D. Gao, V. Mahadevan, and N. Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, 8(7):13, 2008.

[GMZ08] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *IEEE Conf on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[Goo05] J. W Goodman. *Introduction to Fourier optics*. Roberts & Company Publishers, 2005.

[Gro76] S. Grossberg. Adaptive pattern classification and universal recoding: I. parallel development and coding of neural feature detectors. *Biological Cybernetics*, 23(3):121–134, 1976.

[GV09] D. Gao and N. Vasconcelos. Decision-theoretic saliency: Computational principles, biological plausibility, and implications for neurophysiology and psychophysics. *Neural Computation*, 21(1):239–271, 2009.

[Hay09] S. S Haykin. *Neural networks and learning machines*. Prentice Hall, 2009.

[HB05] M. Hayhoe and D. Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4):188–194, 2005.

[HC08] D. S Hwang and S. Y Chien. Content-aware image resizing using perceptual seam carving with human attention model. In *Proc. of the IEEE ICME*, page 1029–1032, 2008.

[HH00] P. O Hoyer and A. Hyvärinen. Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, 11(3):191–210, 2000.

[HK09] J. Harel and C. Koch. On the optimality of spatial attention for object detection. In *Attention in Cognitive Systems*, page 1–14, 2009.

[HKP07] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems (NIPS)*, volume 19, page 545, 2007.

[HO97] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[HS03] P. Hao and Q. Shi. Reversible integer KLT for progressive-to-lossless compression of multiple component images. In *Int. Conf. on Image Processing (ICIP)*, volume 1, 2003.

[HV10] S. Han and N. Vasconcelos. Biologically plausible saliency mechanisms improve feedforward object recognition. *Vision Research*, 50(22):2295–2307, 2010.

[HW59] D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3):574, 1959.

[HW68] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215, 1968.

[HYZF10] J. Huang, X. Yang, R. Zhang, and X. Fang. Re-ranking image search results by multiscale visual saliency model. In *IEEE Int. Symp. on Broadband Multimedia Systems and Broadcasting (BMSB)*, page 1–4, 2010.

[HZ07a] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, page 1–8, 2007.

[HZ07b] X. Hou and L. Zhang. Thumbnail generation based on global saliency. In *Advances in Cognitive Neurodynamics (ICCN)*, page 999–1003, 2007.

[HZ08] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. In *Advances in Neural Information Processing Systems (NIPS)*, volume 21, page 681–688, 2008.

[HZL+09] G. Hua, C. Zhang, Z. Liu, Z. Zhang, and Y. Shan. Efficient scale-space spatiotemporal saliency tracking for distortion-free video retargeting. In *Asian Conf. on Computer Vision (ACCV)*, page 182–192, 2009.

[IK00] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, 2000.

[IKN98] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[Itt04] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, 13(10):1304–1318, 2004.

[KB01] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[KC10] C. Kanan and G. Cottrell. Robust classification of objects, faces, and flowers using natural image statistics. In *IEEE int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[KNdB08] G. Kootstra, A. Nederveen, and B. de Boer. Paying attention to symmetry. In *Proc. of the British Machine Vision Conference (BMVC)*, page 1115–1125, 2008.

[Koh07] A. Kohn. Visual adaptation: physiology, mechanisms, and functional benefits. *Journal of Neurophysiology*, 97(5):3155, 2007.

[Kov96] P. Kovesi. *Invariant measures of image features from phase information*. PhD thesis, Department of Psychology, University of Western Australia, 1996.

[KS09] G. Kootstra and L. R.B Schomaker. Prediction of human eye fixations using symmetry. In *Cognitive Science Conference (CogSci), Amsterdam, NL*, 2009.

[KU85] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–27, 1985.

[KZ07] A. R Koene and L. Zhaoping. Feature-specific interactions in salience from combined feature contrasts: Evidence for a bottom-up saliency map in v1. *Journal of Vision*, 7(7):6, 2007.

[LER+07] A. N Landau, M. Esterman, L. C Robertson, S. Bentin, and W. Prinzmetal. Different effects of voluntary and involuntary attention on EEG activity in the gamma band. *Journal of Neuroscience*, 27(44):11986, 2007.

[LH84] M. S. Livingstone and D. H. Hubel. Anatomy and physiology of a color system in the primate visual cortex. *Journal of Neuroscience*, 4(1):309, 1984.

[LH87] M. S Livingstone and D. H Hubel. Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *Journal of Neuroscience*, 7(11):3416, 1987.

[LL10] M. Loog and F. Lauze. The improbability of harris interest points. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[Low04] D. G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[LSL01] S. Lim, K. H Sohn, and C. Lee. Principal component analysis for compression of hyperspectralimages. In *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*, volume 1, 2001.

[LWS02] T. W Lee, T. Wachtler, and T. J Sejnowski. Color opponency is an efficient representation of spectral properties in natural scenes. *Vision Research*, 42(17):2095–2103, 2002.

[LYS+10] Z. Liu, H. Yan, L. Shen, K. N Ngan, and Z. Zhang. Adaptive image retargeting using saliency-based continuous seam carving. *Optical Engineering*, 49:017002, 2010.

[MAH10] E. Mavritsaki, H. A Allen, and G. W Humphreys. Decomposing the neural mechanisms of visual search through model-based analysis of fMRI: top-down excitation, active ignoring and the use of saliency by the right TPJ. *NeuroImage*, 52(3):934–946, 2010.

[MB88] M. C Morrone and D. C. Burr. Feature detection in human vision: A phase-dependent energy model. In *Proc. of the Royal Society of London. Series B, Biological Sciences*, page 221–245, 1988.

[MC10] O. Le Meur and P. Le Callet. What we see is most likely to be what matters: visual attention and applications. In *IEEE Int. Conf. on Image Processing (ICIP)*, page 3085–3088, 2010.

[MCB07] O. Le Meur, P. Le Callet, and D. Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47(19):2483–2498, 2007.

[MCBT06] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):802–817, 2006.

[MFL+08] D. Meger, P. E Forssén, K. Lai, S. Helmer, S. McCann, T. Southey, M. Baumann, J. J Little, and D. G Lowe. Curious george: An attentive semantic robot. *Robotics and Autonomous Systems*, 56(6):503–511, 2008.

[MGS+07] T. Michalke, A. Gepperth, M. Schneider, J. Fritsch, and C. Goerick. Towards a human-like vision system for resource-constrained intelligent cars. In *Int. Conf. on Computer Vision Systems, Bielefeld*, 2007.

[Mil93] R. Milanese. *Detecting salient regions in an image: from biological evidence to computer implementation.* PhD thesis, University of Geneva, 1993.

[MK01] D. Melcher and E. Kowler. Visual scene memory and the guidance of saccadic eye movements. *Vision Research*, 41(25-26):3597–3611, 2001.

[MKH09] S. K Mannan, C. Kennard, and M. Husain. The role of visual salience in directing eye movements in visual object agnosia. *Current Biology*, 19(6):R247–R248, 2009.

[MLBV10] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[MLHL07] L. Montaser-Kouhsari, M. S Landy, D. J Heeger, and J. Larsson. Orientation-selective adaptation to illusory contours in human visual cortex. *Journal of Neuroscience*, 27(9):2186, 2007.

[MS05] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1615–1630, 2005.

[MS10] S. Montabone and A. Soto. Human detection using a mobile platform and novel features derived from a visual saliency mechanism. *Image and Vision Computing*, 28(3):391–402, 2010.

[MTA04] J. D Marsack, L. N Thibos, and R. A Applegate. Metrics of optical quality derived from wave aberrations predict visual performance. *Journal of Vision*, 4(4), 2004.

[MWG+94] R. Milanese, H. Wechsler, S. Gill, J. M Bost, and T. Pun. Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).*, page 781–785, 1994.

[NI05] V. Navalpakkam and L. Itti. Modeling the influence of task on attention. *Vision Research*, 45(2):205–231, 2005.

[NI06] V. Navalpakkam and L. Itti. Optimal cue selection strategy. In *Advances in Neural Information Processing Systems (NIPS)*, volume 18, page 987, 2006.

[NM89] K. Nakayama and M. Mackeben. Sustained and transient components of focal visual attention. *Vision Research*, 29(11):1631–1647, 1989.

[Not93] H. C Nothdurft. The conspicuousness of orientation and motion contrast. *Spatial Vision*, 7(4):341–363, 1993.

[NPVR08] J. L Nieves, C. Plata, E. M Valero, and J. Romero. Unsupervised illuminant estimation from natural scenes: an RGB digital camera suffices. *Applied Optics*, 47(20):3574–3584, 2008.

[O+96] B. A Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[OAE93] B. A Olshausen, C. H Anderson, and D. C Van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13(11):4700, 1993.

[OBH+01] N. Ouerhani, J. Bracamonte, H. Hugli, M. Ansorge, and F. Pellandini. Adaptive color image compression based on visual attention. In *Int. Conf. on Image Analysis and Processing*, page 416–421, 2001.

[OF05] B. A Olshausen and D. J Field. How close are we to understanding v1? *Neural Computation*, 17(8):1665–1699, 2005.

[Oli05] A. Oliva. Gist of the scene. *Neurobiology of Attention*, 17:251–257, 2005.

[OT01] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[OT06] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 155:23–36, 2006.

[OTM+08] J. Otero-Millan, X. G Troncoso, S. L Macknik, I. Serrano-Pedraza, and S. Martinez-Conde. Saccades and microsaccades during visual fixation, exploration and search: Foundations for a common saccadic generator. *Journal of Vision*, 8(14):21, 2008.

[Oue04] N. Ouerhani. *Visual Attention: From Bio-Inspired Modelling to Real-Time Implementation*. PhD thesis, University of Neuchatel, 2004.

[PDZ00] A. Pouget, P. Dayan, and R. Zemel. Information processing with population codes. *Nature Reviews Neuroscience*, 1(2):125–132, 2000.

[PIW10] N. Parikh, L. Itti, and J. Weiland. Saliency-based image processing for retinal prostheses. *Journal of Neural Engineering*, 7:016006, 2010.

[PLN02] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, 2002.

[PTMO07] B. Penna, T. Tillo, E. Magli, and G. Olmo. Transform coding techniques for lossy hyperspectral data compression. *IEEE Transactions on Geoscience and Remote Sensing*, 45(5 Part 2):1408–1421, 2007.

[PVV09] C. A Parraga, M. J Vazquez-Corral, and M. Vanrell. A new cone activation-based natural images dataset. In *European Conference on Visual Perception (ECVP)*, 2009.

[RFV10] C. Roggeman, W. Fias, and T. Verguts. Salience maps in parietal cortex: Imaging and computational modeling. *NeuroImage*, 52(3):1005–1014, 2010.

[RLB+08] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer. Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub. In *Int. Conf. on Robotics and Automation (ICRA)*, page 962–967, 2008.

[RNB04] R. Rosenholtz, A. L Nagy, and N. R Bell. The effect of background color on asymmetries in color search. *Journal of Vision*, 4(3):224–240, 2004.

[RR09] F. Rieke and M. E Rudd. The challenges natural images pose for visual adaptation. *Neuron*, 64(5):605–616, 2009.

[RvdLBC08] U. Rajashekar, I. van der Linde, A. C Bovik, and L. K Cormack. GAFFE: a gaze-attentive fixation finding engine. *IEEE Transactions on Image Processing*, 17(4):564, 2008.

[SACB10] P. Santana, N. Alves, L. Correia, and J. Barata. A Saliency-Based approach to boost trail detection. In *Int. Conf. on Robotics and Automation (ICRA)*, 2010.

[SBR+10] O. Soceanu, G. Berdugo, D. Rudoy, Y. Moshe, and I. Dvir. Where's waldo? human figure segmentation using saliency maps. In *Proc. of the Int. Symp. on Communications, Control and Signal Processing (ISCCSP)*, 2010.

[SGJ+95] A. M Sillito, K. L Grieve, H. E Jones, J. Cudeiro, et al. Visual cortical mechanisms detecting focal orientation discontinuities. *Nature*, 378(6556):492–496, 1995.

[SHD07] O. Schwartz, A. Hsu, and P. Dayan. Space and time in visual context. *Nature Reviews Neuroscience*, 8(7):522–535, 2007.

[SI07] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 300–312, 2007.

[SJT10] C. Savin, P. Joshi, and J. Triesch. Independent component analysis in spiking neurons. *PLoS Computational Biology*, 6(4):e1000757, 2010.

[SLF03] P. Series, J. Lorenceau, and Y. Frégnac. The "silent" surround of v1 receptive fields: theory and experiments. *Journal of Physiology*, 97(4):453–474, 2003.

[SM09] H. J Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9:12–15, 2009.

[SM10] H. J Seo and P. Milanfar. Training-free, generic object detection using locally adaptive regression kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1688–1704, 2010.

[SO01] E. P Simoncelli and B. A Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216, 2001.

[SP75] V. C Smith and J. Pokorny. Spectral sensitivity of the foveal cone photopigments between 400 and 500 nm. *Vision Research*, 15(2):161–171, 1975.

[SPV07] Y. B Saalmann, I. N Pigarev, and T. R Vidyasagar. Neural mechanisms of visual attention: how top-down feedback highlights relevant locations. *Science*, 316(5831):1612, 2007.

[SR10] M. W Self and P. R Roelfsema. A monocular, unconscious form of visual attention. *Journal of Vision*, 10(4):17, 2010.

[SS00] A. Stockman and L. T Sharpe. The spectral sensitivities of the middle-and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision Research*, 40(13):1711–1737, 2000.

[ST91] E. A. B. Saleh and M. C Teich. *Fundamentals of Photonics*. John Wiley & Sons, New York, Chichester, Brisbane, Toronto, Singapore, 1991.

[SVP10] S. M Stuit, F. A.J Verstraten, and C. L.E Paffen. Saliency in a suppressed image affects the spatial origin of perceptual alternations during binocular rivalry. *Vision Research*, 50(19):1913 – 1921, 2010.

[SVQO01] A. Schoups, R. Vogels, N. Qian, and G. Orban. Practising orientation identification improves orientation coding in v1 neurons. *Nature*, 412(6846):549–553, 2001.

[SWZW10] J. Sun, Y. Wang, Z. Zhang, and Y. Wang. Salient region detection in high resolution remote sensing images. In *The Annual Wireless and Optical Communications Conference (WOCC)*, page 1–4, Shanghai, China, 2010.

[TBG05] B. W Tatler, R. J Baddeley, and I. D Gilchrist. Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45(5):643–659, 2005.

[TBV01] A. Toet, P. Bijl, and J.M. Valeton. Image dataset for testing search and detection models. *Optical Engineering*, 40:1760, 2001.

[TCB10] Y. Tarabalka, J. Chanussot, and J. A. Benediktsson. Segmentation and classification of hyperspectral images using watershed transformation. *Pattern Recognition*, 43:2367–2379, 2010.

[TCW+95] J. K Tsotsos, S. M Culhane, W. Y Kei Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545, 1995.

[TG80] A. M Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

[TG88] A. Treisman and S. Gormican. Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, 95(1):15–48, 1988.

[THBA04] L. N Thibos, X. Hong, A. Bradley, and R. A Applegate. Accuracy and precision of objective refraction from wavefront aberrations. *Journal of Vision*, 4(4):9, 2004.

[TM04] J. J Todd and R. Marois. Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature*, 428(6984):751–754, 2004.

[TMMC05] X. G Troncoso, S. L Macknik, and S. Martinez-Conde. Novel visual illusions related to Vasarely's nested squares show that corner salience varies with corner angle. *Perception*, 34:409–420, 2005.

[TOCH06] A. Torralba, A. Oliva, M. S Castelhano, and J. M Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113(4):766–786, 2006.

[TS85] A. Treisman and J. Souther. Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General*, 114(3):285–310, 1985.

[TWY07] M. Tian, S. Wan, and L. Yue. A novel approach for change detection in remote sensing image based on saliency map. In *Computer Graphics, Imaging and Visualisation (CGIV)*, page 397–402, 2007.

[TZKM11] T. Töllner, M. Zehetleitner, J. Krummenacher, and H. J Müller. Perceptual basis of redundancy gains in visual pop-out search. *Journal of Cognitive Neuroscience*, 23(1):137–150, 2011.

[vdWGB06] J. van de Weijer, T. Gevers, and A. D. Bagdanov. Boosting color saliency in image feature detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):150–156, 2006.

[VG00] W. E Vinje and J. L Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273, 2000.

[VG02] W. E Vinje and J. L Gallant. Natural stimulation of the nonclassical receptive field increases information transmission efficiency in v1. *Journal of Neuroscience*, 22(7):2904, 2002.

[vKGS+10] M. G van Köningsbruggen, S. Gabay, A. Sapir, A. Henik, and R. D Rafal. Hemispheric asymmetry in the remapping and maintenance of visual saliency maps: A TMS study. *Journal of Cognitive Neuroscience*, 22(8):1730–1738, 2010.

[VM10] M. Verma and P. W McOwan. A semi-automated approach to balancing of bottom-up salience for predicting change detection performance. *Journal of Vision*, 10(6), 2010.

[vZD06] W. van Zoest and M. Donk. Saccadic target selection as a function of time. *Spatial Vision*, 19(1):61–76, 2006.

[WC06] J. Wang and C. I Chang. Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 44(6):1586–1600, 2006.

[WH04] J. M Wolfe and T. S Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5(6):495–501, 2004.

[WK06] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006.

[WM97] M. A Webster and J. D. Mollon. Adaptation and the color statistics of natural images. *Vision Research*, 37(23):3283–3298, 1997.

[WMBW02] M. A Webster, G. Malkoc, A. C Bilson, and S. M Webster. Color contrast and contextual influences on color appearance. *Journal of Vision*, 2(6):7, 2002.

[Wol94] J. M Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, 1:202–202, 1994.

[WPPF10] G. L. West, C. Pun, J. Pratt, and S. Ferber. Capacity limits during perceptual encoding. *Journal of Vision*, 10(2):14, 2010.

[WRKP05] D. Walther, U. Rutishauser, C. Koch, and P. Perona. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding*, 100(1-2):41–63, 2005.

[WSA03] T. Wachtler, T. J Sejnowski, and T. D Albright. Representation of color stimuli in awake macaque primary visual cortex. *Neuron*, 37(4):681–691, 2003.

[WTSL08] Y. S Wang, C. L Tai, O. Sorkine, and T. Y Lee. Optimized scale-and-stretch for image resizing. In *ACM Trans. on Graphics*, volume 27, page 118, 2008.

[Yar67] A. L. Yarbus. *Eye movements and vision.* Plenum Press, 1967.

[Zha98] L. Zhaoping. A neural model of contour integration in the primary visual cortex. *Neural Computation*, 10(4):903–940, 1998.

[Zha02] L. Zhaoping. A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6(1):9–16, 2002.

[Zha08] L. Zhaoping. Attention capture by eye of origin singletons even without awareness—A hallmark of a bottom-up saliency map in the primary visual cortex. *Journal of Vision*, 8(5):1, 2008.

[ZM07] L. Zhaoping and K. A May. Psychophysical tests of the hypothesis of a bottom-up saliency map in primary visual cortex. *PLoS Computational Biology*, 3(4):e62, 2007.

[ZS06] L. Zhaoping and R. J Snowden. A theory of a saliency map in primary visual cortex (V1) tested by psychophysics of colour–orientation interference in texture segmentation. *Visual Cognition*, 14(4):911–933, 2006.

[ZTM+08] L. Zhang, M. H Tong, T. K Marks, H. Shan, and G. W Cottrell. SUN: a bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32, 2008.