International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 8, Issue 5 (August 2013), PP.41-53

A Medical Image Retrieval Framework in Cognitive Processes in Eye Guidance Enhanced Visual Concept

Himadrinath Moulick[1], Moumita Ghosh[2]
[1]*CSE, Aryabhatta Institute of Engg& Management, Durgapur, PIN-713148, India*
[2]*CSE,University Institute Of Technology,(The University Of Burdwan) Pin -712104,India*

Abstract:-This paper presents a medical image retrieval frameworkthat uses visual concepts in a feature space employing statisticalmodels built using a probabilistic multi-class supportvector machine (SVM). The images are represented usingconcepts that comprise color and texture patches fromlocal image regions in a multi-dimensional feature space.A major limitation of concept feature representation is thatthe structural relationship or spatial ordering between conceptsare ignored. We present a feature representationscheme as visual concept structure descriptor (VCSD) thatovercomes this challenge and captures both the concept frequencysimilar to a color histogram and the local spatialrelationships of the concepts. A probabilistic frameworkmakes the descriptor robust against classification and quantizationerrors. Evaluation of the proposed image retrievalframework on a biomedical image dataset with differentimaging modalities validates its benefits.When inspecting an image for the first time,how does the viewer decide where to look next? The saliencymap hypothesis proposes that viewers initiallyanalyse the image for variations in low-level visual featuresincluding intensity, colour, and edge orientation, and thattheir eyes are guided towards the most salient region. Thesaliency of objects in scenes may provide an explanation ofwhy some experiments find that incongruent objects attractattention whilst other studies do not find this effect.Experiments that have monitored eye movements duringscene inspection have found some support for the saliencymap hypothesis, particularly when pictures are inspected inanticipation of a memory test. Under some circumstancesthe hypothesis fails to account for inspection patterns.When scenes are inspected to check the presence orabsence of a named object, or when two images are comparedto determine whether they are identical, or when theviewer has specialised domain knowledge of the scenedepicted, then saliency has little influence. This paperevaluates the saliency map hypothesis of scene perceptionusing evidence of eye movements made when images arefirst inspected, and concludes that visual saliency can beused by viewers, but that its use is both task-dependent andknowledge-dependent.

Keywords:- Content- Based Image Retrieval (CBIR),Attentio**n** _ Scene perception ,Saliency map models ,Eye movements ,Fixation scanpaths.

I. INTRODUCTION

Biomedical Images are commonly stored, retrieved andtransmitted in the DICOM (Digital Imaging and Communicationin Medicine) format 1 in a Picture Archiving andCommunications System (PACS) [2] and image search ison the textual attributes, such as person information, otherhealth meta data, often found in image headers. These attributesare often very brief, however, typically limited tothe diagnostic content. It is believed that while improvementsin medical image-based diagnoses could be effectedthrough efficient and accurate access to images and relatedinformation, their utilization may be limited due to the lackof effective image search methods [1]. Further, search resultsmay be improved by combining text attribute-basedsearch capability with low-level visual features computeddirectly on the image content commonly known as Content-Based Image Retrieval (CBIR) [3]. CBIR has the capabilityto identify visually similar images from a database,however, their relevance may be limited by the “semanticgap”. This gap is introduced due to the limited discriminativepower of low-level visual features that are used as descriptorsfor high-level semantic concepts expressed in animage. In an effort to minimize the semantic gap, some recentapproaches have used machine learning on image featuresextracted from local regions in a partitioned image ina *“bag of concepts”*-based image representation schemeby treating the features as visual concepts [3]. Such an imagerepresentation scheme is based on the *“bag of words”*representation commonly used in information retrieval fromtext documents [7]. In this approach, each word is consideredindependent of all other words and results in loss indocument structure. While it has proven effective for textretrieval, it suffers from loss of semantics expressed in adocument. This limitation also extends to image retrievaland is further exacerbated because often the correspondencebetween an image region and local concept is not alwaysalways direct [3]. Considering only a single concept perimage region while completely ignoring others may lead totwo regions matched to different concepts even though theymight be very similar or correlated with each other.This paper presents a spatial correlation-enhanced medicalimage representation and retrieval framework to addressthese limitations of the low-level and concept-level featurerepresentation schemes. The organization of the paper isas follows: Section 2 describes the visual concept-basedimage representation approach. Sections 3 and 4 presenta correlation enhanced probabilistic feature representationand structural relationship enhanced feature representationscheme respectively. The experiments and the analysis ofthe results are presented in Section 5 and Section 6 providesconclusions.When we first inspect a picture—a photograph, a drawing,or a painting—our eyes are attracted to some objects andfeatures in preference to others. We look at objects insuccession rather than holding our eyes in the centre of theimage. This is inevitable, given that our vision is mostacute at the point of fixation, and given that we can onlylook in one place at a time. We move our eyes around animage in order to give the components of the image fovealscrutiny. But what are the characteristics of images thatattract our attention and in what order should the picture'scomponents be inspected? Do we look predominantly thelow-level visual features defined most appropriately interms of contour, contrast and colour, or is the meaningfulconfiguration of the objects depicted by those featuresperceived quickly enough for eye guidance to be a topdownprocess? The argument presented here considers abottom-up saliency map hypothesis as a model of attentionalguidance, reviewing evidence from eye-trackingstudies of image processing, and concluding that the modelworks well in very specific circumstances, but that theeffects of visual saliency can be overridden by the cognitivedemands of the task. By way of introducing theattraction of the visual saliency map hypothesis, we firstconsider explanations for a long-standing controversy inthe psychology of picture perception—the issue of whetherobjects that violate the gist of a scene are perceived moreeasily than congruent objects, or with more difficulty.To illustrate the processes in scene inspection, take abrief look at Fig. 1, which is a photograph taken in akitchen. Close inspection will reveal the identities of severalobjects that seem to be in their place, but there is alsoan object that does not adhere to the scene gist—the tapemeasure on the lower left side of the picture. Is the tapemeasure easier to identify, as a result of being set in an incongruous context, or more difficult? A straightforwardanswer to this question comes from studies of objectnaming, in which the perceiver has the task of eitherdeciding whether a named object is present in a scene [1],or whether a member of a named category of objects ispresent [2], or of declaring the identity of an object in aspecific location [3]. It is more difficult to identify objectsthat violate the gist in these experiments. For example, identifying a fire hydrant in a living room, or a footballplayer in a church, would be more difficult in either form ofobject detection task. The pattern of results in these studiessupports an interactive model of scene perception in whichthe context and the component objects provide mutualfacilitation, with the scene gist aiding the identification ofother objects that contribute to this context. This resultlends support to the idea that we recognise scenes by theircomponents and that the overall scene helps in the identificationof its component objects. Any misfit object that is incongruent with the scene will be recognised with greaterdifficulty than objects that are usually associated with thatscene.It is important to note that in both of the object identificationtasks considered so far the viewer is required tomatch an object to a name, and this requirement may helpexplain why incongruous objects are sometimes seen earlierthan those that comply with the gist. The starting pointfor this debate is an experiment reported by Mackworthand Morandi [4] in which viewers tended to look first atthose parts of a picture that were judged by a set of independentviewers as being highly informative, suggestingthat salient meanings could be captured sufficiently early todirect eye movements during the first few seconds ofviewing. Instead of having a panel of judges rate theinformation values of zones within a picture, Loftus andMackworth [5] showed sketches of scenes with a recognizable gist (e.g., a farmyard scene comprising drawings ofa barn, farmhouse, fencing and a cart), and placed an objectin the drawing that was congruous (a tractor) or incongruous(an octopus). Incongruous objects were fixatedbefore their congruous counterparts, leading to the suggestionthat gist and violations of gist are detectedsufficiently early to guide the first few eye fixations, if notthe very first movement to an object in the scene. A similarresult is found with photographs of natural scenes in whichobjects are edited in to create new pictures that havecongruous or incongruous objects in them [6]. Again,objects that were not usually a part of the scene, such as acow grazing on a ski slope, were fixated earlier than congruousobjects that were edited into a similar place (a skierin this example). This is an interesting effect because itsuggests that we do not need to inspect each object in ascene to understand the gist or to identify an object thatviolates the gist. The effect, if it is robust, demonstratesthat parafoveal or peripheral vision can be used for objectidentification.When we ask whether incongruous objects are perceivedmore easily or less easily, two kinds of investigationsproduce very different conclusions. The object detectionstudies requiring viewers to say whether a named object ispresent, or to offer the name of an object, report that misfitobjects are more difficult than those that comply with thegist, but eye movement studies that call for free inspectionof a picture find that unusual objects are fixated early. Toresolve this inconsistency we first need to consider anotherinconsistency—one between the results of different investigationsof attentional capture by objects that violate thegist.

II. IMAGE REPRESENTATION ON LOCAL CONCEPT SPACE

In a heterogeneous collection of medical images, it ispossible to identify specific local patches that are perceptually and/or semantically distinguishable, such as homogeneoustexture patterns in grey level radiological images, differentialcolor and texture structures in microscopic pathologyand dermoscopic images. The variation in these localpatches can be effectively modeled by using supervisedlearning based classification techniques such as the SupportVector Machine (SVM) [8]. In its basic formulation, theSVM is a binary classification method that constructs a decisionsurface and maximizing the inter-class boundary betweenthe samples. However, a number of methods havebeen proposed for multi-class classification.For concept model generation, we utilize a voting-based multi-class SVM known as *one-against-one* or pairwisecoupling (PWC) [9]. In developing training samples for this SVM, only local image patches that map to visual conceptmodels are used. A fixed-partition based approach is used at first to divide the entire image space into a ($r \times r$) gridof non-overlapping regions. Manual selection is applied tolimit such patches in the training set to those that have amajority of their area (80%) covered by a single semanticconcept. In order to perform the multi-class SVMs trainingbased on the local concept categories, a set of L labels areassigned as $C = \{c1, \cdot\cdot\cdot, ci, \cdot\cdot\cdot, cL\}$, where each $ci \in C$characterizes a local concept category. Each patch is labeledwith only one local concept category and is represented bya combination of color and texture moment-based features.Images in the data set are annotated with local conceptlabels by partitioning each image Ijinto an equivalent $r\times r$grid of l region vectors $\{\mathbf{x}1j, \cdot\cdot\cdot, \mathbf{x}kj, \cdot\cdot\cdot, \mathbf{x}lj\}$, whereeach $\mathbf{x}kj \in \_d$ is a combined color and texture feature vector.For each $\mathbf{x}kj$, the local concept category probabilitiesare determined by the prediction of the multi-class SVMsas [9]

$$p_{ik_j} = P(y = i \mid \mathbf{x}_{k_j}),\ 1 \le i \le L. \quad (1)$$

Based on the probability scores, the category label of $xkj$is determined as cm as the label with the maximum probabilityscore. Hence, the entire image is thus represented asa two-dimensional index linked to the concept or localizedsemantic labels assigned for each region. Based on this encodingscheme, an image Ijcan be represented as a vectorin a local semantic concept space as

$$\mathbf{f}_j^{\text{Concept}} = [f_{1_j}, \cdots, f_{i_j}, \cdots f_{L_j}]^{\mathrm{T}} \quad (2)$$

Where each fijcorresponds to the normalized frequency ofa concept ci, $1 \le i \le L$ in image Ij. However, this representationcaptures only a coarse distribution of the concepts.It is very sensitive to quantization or classification

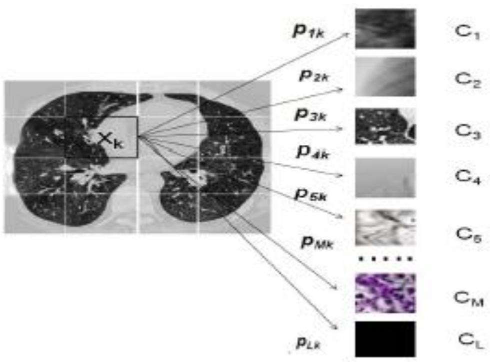

Fig. 1: Probabilistic membership scores errors and ignores correlations and structural relationshipsamong concepts.

A. Probabilistic Feature Representation

The feature vector **f** concept can be viewed as a local conceptdistribution from a probabilistic viewpoint. Given a set of concept categories of length L, each element $fij$of**f**concept$j$ for an image $Ij$is calculated as $fij= li/l$. It is the probability of a region in the image encoded with label i ofthe concept $ci \in C$, and li is the total number of regions thatmap to ci. According to the total probability theory [10], fijcan be defined as

$$f_{i_j} = \sum_{k_j=1}^{l} P_{i|k_j} P_k = \frac{1}{l} \sum_{k_j=1}^{l} P_{i|k_j} \quad (3)$$

Where$Pk$is the probability of a region selected from imageIjbeing the kjth region, which is $1/l$, and $Pi|kj$is theconditional probability that the selected kjth region in Ijmaps to the concept ci. In the context of the concept vector

fconcept$j$, the value of $Pi|kj$is 1 if the region $kj$is mapped tothe ci concept, or 0 otherwise. Due to the crisp membershipvalue, this feature representation is sensitive to quantizationerrors.We present a feature representation scheme based on theobservation that there are usually several concepts that arehighly similar or correlated to the best matching one fora particular image region. This scheme spreads each region'smembership values or confidence scores to all thelocal concept categories. During the image encoding process,the probabilistic membership values of each regionto all concept prototypes are computed for an image Ij.For example, Figure 1 shows a particular region in a segmentedimage and its probabilistic membership scores todifferent local concept categories. Based on the probabilisticvalues of each region, an image $Ij$is represented as

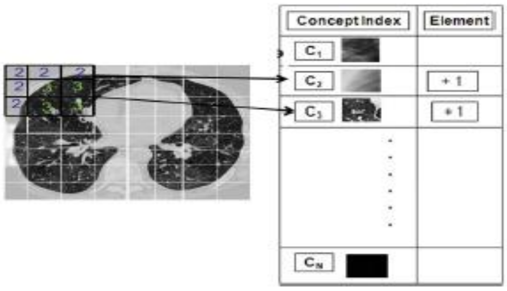

Fig. 2: Visual concept structure descriptor

$$\mathbf{f}_j^{\mathrm{PVCV}} = [\hat{f}_{1_j} \cdots \hat{f}_{i_j} \cdots \hat{f}_{L_j}]^{\mathrm{T}}, \text{ where}$$

$$\hat{f}_{i_j} = \sum_{k=1}^{l} p_{ik_j} P_k = \frac{1}{l}\sum_{k=1}^{l} p_{ik_j}; \quad \text{for } i = 1, 2, \cdots, L \qquad (4)$$

Where$pikj$is determined based on (1). Here, we considereach of the regions in an image being related to all the conceptsvia the membership values such that the degree of associationof the kj-th region in $Ij$to the ci concept is determinedby distributing the membership values to the correspondingindex of the vector. In contrast to the simple concept vector **f** concept, this vector representation considersnot only the similarity of different region vectors from different concepts but also the dissimilarity of those regionvectors mapped to the same concepts.

B. Structural Feature Representation

A major limitation of concept feature representation isthat the structural relationship or spatial ordering betweenconcepts are ignored. This representation can not distinguishbetween two images in which a given concept ispresent in identical numbers but where the structure of thegroups of regions having that concept is different. Wepresent a feature representation scheme as *visual conceptstructure descriptor* (VCSD) that overcomes this challenge and captures both the concept frequency similar to a colorhistogram and the local spatial relationships of the concepts.Specifically, it is a vector **f**VCSDj = [$fv1j$ ・ ・ ・$fvij$ ・ ・ ・$fvLj$]T,where each element $fvij$represents the number of times avisual concept label is present in a windowed neighborhooddetermined by a small square structuring element. The sizeof the structuring element is ($b \times b, b < r$) units. This isillustrated in Figure 2 where an image is partitioned into 64blocks ($r = 8$). A 9-element ($b = 3$) structuring elemententables distinction between images with the same conceptsthat are in equal proportions on their distribution. The structuringelement is moved over the image in an overlappingfashion and accumulates the visual concept labels. This processis also illustrated in the figure. For each unique conceptat a particular position in the image within the structuringelement, the corresponding element of the feature vector isincremented. Upon completion, the concept vector is normalizedby the number of positions of the structuring element.

C. Experiments and Results

The image collection for experiment comprises of 5000bio-medical images of 30 manually assigned disjoint global categories, which is a subset of a larger collection of sixdifferent data sets used for medical image retrieval task inImageCLEFmed 2007 [5]. In our collection, the images areclassified into three levels as modalities, body parts, orientationsor distinct visual observation. For the SVM training,30 local concept categories, such as tissues of lung or brainof CT or MRI, bone of chest, hand, or knee X-ray, microscopicblood or muscle cells, dark or white background, etc.are manually defined. The training set used for this purposeconsist of only 5% images of all global categories of the entiredata set. To generate the local patches, each image inthe training set is at first partitioned into an 8 × 8 grid generating64 non-overlapping regions. Only the regions thatconform to at least 80% of a particular concept categoryare selected and labeled with the corresponding category label.For the SVM training, a 10-fold cross-validation (CV)is conducted to find the best values of tunable parameters$C$ = 200 and γ = 0.02 of the radial basis function (RBF)kernel with a CV accuracy of 81.01%. We utilized the *LIBSVM*2 software package for implementing the multi-classSVM classifier.For a quantitative evaluation of the retrieval results, weselected all the images in the collection as query imagesand used *query-by-example (QBE)* as the search method.Figure 3 shows the precision-recall curves based on theEuclidean similarity matching in different feature spaces.The performance was compared to the low-level MPEG-7based color layout descriptor (CLD) and edge histogramdescriptor (EHD) [11]. By analyzing the Figure 3, wecan observe that the proposed concept-based feature representationschemes performed much better compared tothe low-level MPEG-7 (e.g., CLD and EHD) based featuresin terms of precision at each recall level. The betterperformances are expected as the concept features aremore semantically oriented that exploits the domain knowledgeof the collections at a local level. It is also noticeablethat, the performances of both the probabilistic visualconcept vector (PVCV) and visual concept structure descriptor(VCSD) increase at a lower recall level (e.g., upto 0.6) when compared to the normalized frequency based feature vector (e.g.,Concept). These results areencouraging enough as users are mainly interested to find relevant images in only few retrieved images (e.g., at a low recall level).

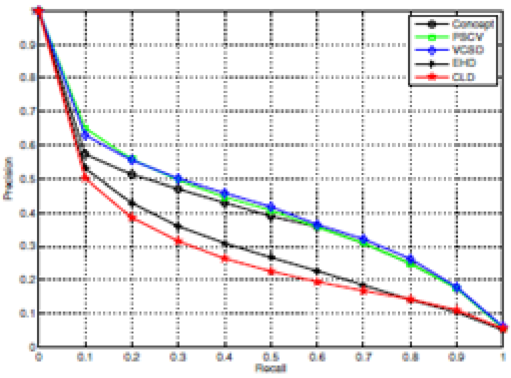

Fig. 3: Precision-recall curves in differentfeature spaces**.**

From the results, we can conjecture that there isalways enough correlation and structural relationships betweenthe local concepts, which can be exploited in the featurerepresentation schemes.Saliency Maps in Scene PerceptionAs part of a model of saccadic programming, Findlay andWalker [10] identified two separate pathways for eyemovement control. These two mechanisms essentiallycontrol the when and the where of saccadic movement, andthe decision about where the next fixation should be targetedis made with the aid of a saliency map. (Note:Findlay and Walker used the term ‘‘salience map’’ but forconsistence with other descriptions the term ‘‘saliencymap’’ will be used here, and it will be assumed that the twoterms refer to the same idea.) The map is a topographicdescription of points of interest, enabling the spatial pathway(the ‘‘where pathway’’ in their model) to select asaccadic target and to controlling the decision where tomove. One source of input to the saliency map is visualcontours and another is contrast. We can anticipatedevelopments of the model by suggesting that regions ofimage that have distinctive colours would also be input tothe map. Identification of these low-level visual characteristicswould together provide a description of thefeatures of an image, and would influence decisions aboutsaccadic programming. Henderson et al. [8] outlined theprocess whereby the saliency map is used to guide successivefixations. The map itself is generated by an earlyparsing of the scene into visually differentiated regions ofinterest plus an undifferentiated background with a fastanalysis of low spatial frequency information. Regions ofinterest can then be assigned weights that also reflect theirpotential to attract fixations. The low-level factors thatcontribute to the weightings are luminance, contrast, texture,colour, and contour density, with regions of greatervariance having larger weightings in the map. When aviewer first looks at an image, their attention is allocated tothe region with the greatest weightings, and saccades areprogrammed to move their eyes to an attended region. Theinitial fixations on a picture are therefore determined bylow-level visual factors, according to the Henderson et al.model, and this accounts for the absence of semanticeffects in their experiments with incongruous objects. Aftera perceptual and cognitive analysis of the region, whichresult in the contribution of semantic information to thesaliency map, attention shifts to the region with the nexthighest weighting. Over a series of fixations the mapchanges, with saliency weights initially determined by lowlevelvisual features, and eventually modified to represent asemantic description of the picture. The important pointabout this description is that early fixations are determined by low-level visual features, and it is only after makingseveral fixations on a picture that the viewer with have asemantic interpretation. Only when a region has received adirect or near fixation (within 3_ or 4_) can its saliency weight be determined by its semantic content, and until it isfixated the representation of a region in the map will be dominantly low level. This version of the model has noplace for global scene semantics—the gist of the scene— butTorralba et al. [11] have developed a more powerfulversion in which local visual features are analysed in parallelwith global scene-level features and fixationsdetermined in a ‘‘contextual guidance’’ model. Navalpakkam andItti [12] have also integrated top-down cognitiveinfluences into a revised version of the saliency mapmodel.The early versions of the saliency map model areinformal sketches of the factors that determine where aviewer will look when first inspecting an image, and it wasfor Itti and Koch [13] to make available a fully implementedmodel that could generate specific predictionsabout images that could in turn be tested against humaninspection behaviour. In effect their model formalises thesame principles outlined in the Henderson et al. [8]description, with an early analysis of the distribution ofintensity, colour, and of the orientation of edges, based onKoch and Ullman’s [14] initial formulation of a saliencymap that enables the preattentive selection of regions. Theprocess is essentially competitive, to generate a singleregion that corresponds to the most salient object in thedisplay, the next most salient, and so on. Variations in thevisual characteristics of regions are identified with centresurroundfiltering that operates with several spatial scales,and these analyses result first in feature maps that aredescriptions of the distributions of specific features. Thefiltering of these features results in conspicuity maps foreach characteristic that is analysed. Three characteristicsare appropriate for two-dimensional pictures, but the modelhas been extended to take motion into account with a fourthconspicuity map [15]. The three conspicuity maps forintensity, colour, and orientation are then combined into asingle topographic saliency map. The relationship betweenthese maps is illustrated in Fig. 2.The top panel of the figure shows the original imageprior to processing, and the central panel of three imagesshows the intensity, colour, and orientation conspicuitymaps (from left to right) taken from the original. Note howthe intensity map highlights the brightness of the whiteclothing of the people on the quayside,how the colour mapidentifies the only red and yellow objects in the scene, andhow the orientation map picks out

Fig. 2: A colour image (top) processed through the saliency mapalgorithm developed by Itti and Koch

[13]. The centre panel showsthe three conspicuity maps obtained by identifying variations inintensity, colour, and orientation, respectively. The lower imagerepresents the overall saliency map, using a combination of the threeconspicuity maps (refer to online version for colour figures) the density of contourchanges on the right of the picture. The intensity and orientationmaps are related, but with the boat identified moreclearly in the intensity map, which has picked out the lightcanopy and edging to the deck. The colour map pinpointsthe yellow fishing nets and the boat’s red tiller as the mostconspicuous regions because these are the only objects inthe scene that have these colours. The bottom panel showsthe derived saliency map, which is formed by combiningthe three conspicuity maps. Dark areas indicate lowsaliency.Elazary and Itti [16] evaluated the saliency model usinga dataset of 25,000 photographs of real-world scenes inwhich objects of interest had been previously identified.They used the LabelMe collection of images [17] in whichthe objects in scenes have been outlined on the basis oftheir subjective interest. There is an average of approximatelythree objects of interest in each image in the dataset.When this process of outlining is applied to an image suchas Fig. 1, the areas of interest might be identified as shownin Fig. 3, but the identification of interesting objects isentirely subjective,

and different perceivers might outlinedifferent objects (the labels on the food packages, perhaps,or the title of the book, or the individual grapes). Themodel tended to identify these outlined areas as being themost salient. In 76% of the images, at least one of thethreemost salient regions corresponded to an object of interest,and in 43% of the pictures the most salient region waswithin an outlined area. Both of these percentages are wellabove what would be expected by chance. The techniquesuggests an overlap between the subjective identification ofa ‘‘region of interest’’ and an objective analysis of lowlevelvisual properties. Elazary and Itti’s result gives somesupport to the idea that we might use saliency maps whenidentifying objects in scenes, but this does not tell us howpeople inspect pictures when they first encounter them. The model makes strong predictions about the allocation of attention to objects during the early stages of inspection, and while the correspondence between salient points and interesting objects is supportive, the real test of the model is with the eye fixations of naı¨ve observers.

Fig. 3: A version of the scene from Fig. 1, with important objects identified by outlining

When attention is first allocated to an image such as the weightings of the regions in the saliency mapdetermine the locations of fixations. The single most salientregion in the image is indicated in the top panel of the hat worn by the woman standing on the extreme right ofthe picture. This region is weighted highly in the intensityand orientation maps. The next most salient region isslightly to the right of the centre of the picture, where lightclothing is adjacent to dark shadow. The weights predictthe locations of eye fixations and their sequence and inFig. 4 they are indicated by the ranks of the six most salientregions. The first fixation is predicted to be upon the mostsalient region (the white hat of the woman on the right, inour example), and once this is processed then attentionmoves to the next most salient region, with an inhibitionof-return mechanism suppressing the saliency weighting offirst location in the map. This is necessary in order toprevent attention moving back and forth between the firstand second weights in the saliency map. The inhibition-ofreturnmechanism allows attention to move around theimage without being captured by two points.Evaluating Saliency Maps with Behavioural DataThe saliency map model provides firm predictions aboutthe locations of fixations, and for simple displays andsimple tasks it performs very well. Itti and Koch [13] testedthe model with displays of coloured bars against darkbackgrounds, and the model very readily identified a singlered bar among an array of green bars, and a bar rotatedthrough 90_ in an otherwise homogenous array. This isexactly how human observers perform, displaying the socalledpop-out effect that is central to feature-integrationtheory [18]. The model also performs well with naturalimages shown to participants in a free-viewing task [19]. Inthis task a range of images were shown—indoor and outdoorscenes, as well as computer-generated fractals—andviewers given a few seconds to inspect them while theireye fixations were recorded. The first few fixations tendedto be upon more salient regions. It is difficult to imaginewhat the participants thought they should be doing in thistask, however, given that they were told to look at a seriesof pictures, and nothing more. They might have anticipateda surprise test of recognition at the end of the study period,or some questions about aesthetic preference, but lookingat picture with no purpose might introduce unwanted variancebetween individuals who imagined different purposesto their viewings. When participants are given a specifictask to perform, they behave according to the predictions ofthe model or not, depending on the task. In two memoryexperiments we instructed viewers to inspect photographsof natural scenes in preparation for a memory test, andwere given a few seconds to look at each picture [20, 21].As in the Parkhurst study, their eye movements wererecorded while they looked at the pictures, and as in thatstudy, fixations were located on the regions identified asbeing salient by the Itti and Koch [13] algorithm. Highersaliency objects were fixated earlier than less salientobjects when viewers were attempting to encode the picturein preparation for a task in which they would have todiscriminate between new pictures and those presented forencoding. However, when the same pictures were used in adifferent task, a different result was obtained. In eachpicture there was an object of particular interest—it did notstand out as being of any interest for purposes of thememory test, but it was useful in a search task. In Underwoodet al. [20] the object was a piece of fruit thatappeared in some of the pictures, and in Underwood andFoulsham [21] it was a small grey ball. When viewerssearch for this object in order to declare whether it waspresent or absent in each picture, they successfully avoidedlooking at highly salient distractors. In the search task thesaliency of regions does not attract fixations.A similar result is obtained if the viewer inspects apicture in preparation to answer a question about a specificaspect of the scene. The bottom panel of Fig. 4 shows asequence of fixations recorded from one viewer who wasasked to reply true/false to the statement ''The fisherman isselling his catch at the quayside''. Although there is somecorrespondence between fixations predicted on the basis ofsaliency peaks (top panel of Fig. 4) and the observed fixations(bottom panel), the match is not good for the firstfew fixations. This sentence verification task is perhapsmore similar to an object search task than to an encodingtask, and when comparing a grossly simple measure suchas the number of fixations made, or the overall inspectiontime, this is borne out. Memory tasks elicit longer and moredetailed inspections than object search (e.g., Refs. 20, 21),and the same pattern is seen with sentence verificationbetween presentations where the picture is presented beforethe sentence, and therefore requires encoding into memory,versus presentations where the sentence is read first and thepicture shown afterwards. The picture-first inspectionswere associated with detailed scrutiny of most of theobjects displayed, with an average of more than 14 fixationson each picture, but when the picture was shown afterthe sentence there were less than 7 fixations per picture[22]. In the sentence-first cases, the viewer knew what tolook for in order to verify the sentence, and was able toguide the search to the relevant parts of the scene. Thepicture-first inspections were similar to a short-termmemory test, with encoding in preparation for a singlespecific question about a display that was no longer visible.When viewers inspect pictures in preparation for amemory test, they are attracted to the visually salient areasof the image, but when searching for a named object theyare not so influenced. This distinction helps us to understandthe object congruency effect that started this iscussion. Byconsidering the images used in the different experimentsthat have investigated the congruency effect, the possibilityemerged that inconsistencies in the pattern of results wereattributable to differences in the visual saliency of theincongruous objects used. Perhaps Loftus and Mackworth[5] and others have found that incongruous objects arefixated early because their incongruous objects were visuallymore salient than the objects used by Henderson et al.and others [7–9], who did not find an effect. This suggestionis certainly consistent with the examples of drawings publishedby these authors, but when we investigate the effectwith saliency controlled, in two different paradigms, itemerges that saliency is not the confounding factor.Underwood, Humphreys and Cross [6] photo-editedcongruent and incongruent objects into pictures presentedas part of a recognition memory task. The objects werematched for saliency based on estimates derived fromanalyses of the pictures using the Itti and Koch [13]algorithm. In the first experiment the congruent objects hada mean saliency rank of 3.65 (counting the most salientregion of the picture as rank 1, the second most salientregion as rank 2, and so on) and there was a mean rank of3.55 for the incongruent objects. Congruency was manipulatedin this experiment by exchanging indoor andoutdoor objects between indoor and outdoor scenes. Thesecond experiment used congruent objects (e.g., a skier ona snowy slope, with other skiers in the background),incongruent objects (a snowman edited into the picture, inplace of the skier), and bizarre objects (a cow on the skislope). The mean ranks were 3.07 (congruent), 2.80(incongruent), and 2.77 (bizarre). In neither experiment didthe difference between the ranks approach being a statisticallyreliable difference. In both experiments, however,there were more saccades prior to fixation on a congruousobject than on objects that did not naturally belong in thescene. The incongruent objects were fixated earlier thancongruent objects, and in the second experiment the bizarreobjects were fixated earliest of all. The early fixation ofincongruent objects is consistent with the Loftus andMackworth [5] result, but in conflict with the results fromother experiments that have used line drawings [7–9].Before considering explanations of the inconsistency, weshould establish the robustness of the incongruency effectwith a demonstration from a totally different paradigm.The pattern of inspection was interesting, and is illustratedin the bottom pair of pictures in Fig. 5. Objects arecompared in serial order, first identified in one of thepictures and then matched against the object in the correspondinglocation in the other picture. In this case (a pair ofidentical pictures), the first saccade takes the viewer's eyesto the cola can (the incongruous object) in the right-sidepicture and then to the equivalent location in the left-sidepicture. From there the eyes go to another object in the leftsidepicture (a shampoo bottle), and then to the shampoobottle in the right-side picture, and so on. The viewermakes four of these comparisons before deciding that thepictures are the same. This strategy, which we have seenwhen arrays of individual objects are used rather thancomposed scenes [26], suggests that viewers do not encodea whole scene unless they need to, and will rely on theirvisual memories of individual objects when they can.Saliency differences explain the inconsistency of earlierfixation of incongruent objects in some experiments but notin others. When we control the visual saliency of theobjects the effect remains, whatever the task. So why dosome experiments find an effect of congruency and othersnot? Saliency is not the answer, but the difficulty of objectidentification may be. Consider the two images in Fig. 6,one of which is a colour photograph similar to those used in our experiment, and shows a scene from the corner of aroom that is being decorated. There is an incongruous garden trowel in this picture. The other is a processedversion that identifies the edges, without colour, and which is somewhat similar to the drawings used inexperiments that have failed to find a congruency effect.

Fig. 4:A real-world scene with a readily identifiable gist and a singleobject that is incongruous, represented as a colour photograph and asa version processed through an algorithm that identifies edges andlines (refer to online version for colour figures)

With conducting alaboratory experiment to answer this question, it looks asif the original photograph objects can be recognised moreeasily, and if this is generally the case, then we may havethe basis for an explanation. If each object in the scenehas overlapping edges with other objects, and needs to befirst isolated from its background, then attention isrequired for object recognition. By this process, objectsare constructed from their features, rather than recognized as wholes without attention. If we construct objects inorder to recognise them, they cannot be recognised preattentively,as they must be if we are to identify themwith peripheral vision and move our eyes to them early inthe process of scene inspection. This is the distinctionbetween single feature recognition and feature conjunctionrecognition that forms the basis of the featureintegrationmodel of recognition [18], which argues that attention is the necessary component when we need tocombine features into objects. In the Loftus and Mackworth line drawings, the incongruous objects wereisolated from their backgrounds and could be recognized readily—pre-attentively—but in the studies that used theLeuven library of drawings the objects could not besegregated from their backgrounds without attention andthey had to be inspected in order to enable recognition.Although our experiments with colour photographs usedobjects against rich backgrounds, their segregation ismade possible pre-attentively by virtue of their naturaltexture and colouring, as is apparent in Fig. 6. This is atentative account of differences between experiments, inorder to explain differences in patterns of results, andthere may be other explanations. The appropriate studywould be to use photographs and line drawings in thesame experiment, aiming to demonstrate an incongruencyeffect with one type of stimulus but not the other. Garezeand Findlay [9] did just that, comparing the eye movementsmade with line drawings and greyscalephotographs. A toaster (or a teddy bear) appeared in akitchen or in a child’s playroom, but there was no differencein the number of saccades made prior to fixationof the toaster or the teddy bear. There was no incongruencyeffect in this experiment. On the basis of theexamples presented in their paper, this is unsurprisingbecause object discrimination is still a problem. It isdifficult to identify many of the objects in the photographsor the line drawings, and even when told that theincongruous object in the playroom photograph is atoaster it is not clear where it is (their Figure 4d). Thepossibility remains that the congruency effect dependsupon easy object recognition, and that this emergesonly with a clear separation of the objects from theirbackground. In a free-viewing experiment in which participantsexpected a memory test, the congruency effectemerged with colour photographs [27]. The photographswere edited to introduce anomalous changes (such as aperson’s hand painted green), and these changes werefixated earlier than with the unchanged equivalents. Whenneutral objects were painted—objects that could reasonablyappear in green (a coffee mug)—then fixation wasno earlier in the changed than in the unchanged versions.If we can assume that the congruency effect is real, thenwe still have the problem of explaining why misfit objectscan sometimes attract early fixations. For an incongruentobject to attract an early fixation, both the gist of the sceneand the offending object must be recognised prior toinspection of the object. The simplest explanation is that allobjects in the scene are recognised to the extent that theyform a gist, and that the incongruent object is identifiedincompletely, but to the extent that the viewer becomesaware that there is a problem. This is a perturbation modelof scene recognition that suggests that object recognition isnot all-or-none but is interactive, and that we can know thatsomething is a certain type of object without knowingexactly what it is. The cow on the ski slope in our earlierexperiment, for example, may be identified as an animal orperhaps just as a non-skier, before foveal scrutiny reveals itto be a cow. Partial identification of any object in the scenewould contribute to the development of the scene gist, andonce this context is available it will facilitate the recognitionof additional objects. A misfit object that is partiallyrecognised would attract an eye fixation in order to give itthe attention required to resolve the conflict between objectand context.

III. SCENE PERCEPTION,SALIENCYAND EYE FIXATION SCANPATHS

The experiments with incongruous objects did not resolvethe problem of why some studies find that misfits attract attention early while others do not, but they did eliminatevisual saliency as the explanation. Saliency maps do providea good fit for the data on the early fixations on realworldscenes in some tasks, however, and in this part of thediscussion the extent of the model's explanatory power isconsidered.When viewers look at scenes with no purpose other thanto comply with an researcher's request to do so, the earlyfixations tend to land upon regions identified as highlysalient by the Itti and Koch [13] model [19]. However,salient objects are more likely to fixated when viewersinspect a scene with the intention of encoding it in preparationfor a later memory test than when the same imagesare used in a search task [20, 21]. As we have just seen,saliency plays no role in a comparative visual search task inwhich two pictures are compared for differences. Thepurpose of inspection is important here, implying that topdowncognitive factors can override the attractive powersof visually salient regions. When we know what we arelooking for—a bunch of keys on a desktop, for instance—we are not distracted by a brightly coloured coffee mug.However, when attempting to memorise the scene, thecoffee mug gets our full attention, possibly because it couldbe used as a discriminating feature when making judgementsabout pictures in a recognition test. The brightest,most colourful objects serve a valuable role in memorytests because they can be used as the basis for a decision asto whether the image has been seen previously. Salientregions may be sought in memory experiments, but thisdoes not mean that saliency has a role to play in imageinspection generally. This caveat does not mean that saliencyhas no value to our understanding of sceneperception, only that its potency is specific to the task setfor the viewer. Tatler et al. [28] have raised other objectionsto the saliency map model, arguing that the pattern ofresults in scene perception experiments can just as easily beexplained by habitual tendencies for saccadic eye movements,especially the tendency to fixate objects in thecentre of the screen [29].Rather than comparing the fixation probabilities ofindividual objects in memory and search tasks, Foulshamand Underwood [30] looked at the first five fixations onreal-world scenes, relative to the saliency map. How welldoes the saliency map model predict the locations of thefirst few fixations and particularly the sequence of thosefixations? The purpose of viewing was to prepare for amemory test, and fixations during encoding and recognition were compared against model-predicted fixation locations.With a 2_ radius around each saliency peak, an area of approximately 10% of each picture was defined, andaround 20% of fixations during each phase of the tasklanded on these salient regions: the model performs betterthan chance at predicting the locations of fixations. Analternative way of looking at these data is to calculate thesaliency values of the regions that are actually fixated. We found that the mean saliency values of fixation locations atencoding and during the recognition test were higher thanwould be expected by chance. Estimates of chance werecalculated by three methods: by assuming that the five fixations would be located randomly, with a biased randommodel that uses only actual fixation locations, and with a transitional model that assumed that any fixation woulddepend upon the location of the previous fixation. All three estimates of chance gave mean saliency values lower thanthose observed when actual eye movements were recorded.When the sequence of fixations was taken into account,the model continued to perform well against the eyemovement data. To calculate a five-fixation scanpath, weused a string-editing procedure with fixation locations converted into letters that corresponded to grid locations.Regions of the image were classified according to a 5 *9* 5 grid, with each cell of the grid coded with a letter of thealphabet. The first fixation (centre screen) was eliminated from the string, and repeated fixations on the cell werecondensed into one ''gaze''. Two strings could then becompared using the edit method that calculates the numberof editing operations necessary to convert one string into the other. Insertions, deletions, and substitutions each carrya levy of one edit, using the somewhat dubious assumptionthat all operations have equal value. When the string-editmethod is compared against other string-based methodsthat use the linear distance between fixations, however,very similar estimates of string similarity are obtained. Wecompared actual scanpaths recorded during encoding andduring test against each other and also against fixationsequences predicted by the Itti and Koch [13] saliency mapmodel. The similarity between scanpaths on the samepicture at encoding and at test was reliably better than thesimilarity score for a viewer's scanpaths on two differentpictures, whichever method of quantifying a fixationsequence was used. To predict a scanpath with the saliencymodel, we calculated the five most salient non-contiguousregions, and assumed that the sequence of fixations shouldfollow this rank ordering. The string similarity scores werecalculated for the model against encoding and against therecognition test, and in both comparisons the string similarityscores were lower than when we compared the actualeye fixations made during two viewings of the same picture.The model did not perform as well as humanparticipants looking at a picture the second time, but forboth comparisons with the model the scores were betterthan would be expected by chance, suggesting that thesaliency map model accounts for a significant amount ofthe variance in similarity scores.The Foulsham and Underwood [30] comparison ofobserved fixations against model-predicted fixation locationsestablished that there was a tendency for fixations tooccur in salient regions of the images, that the saliency offixated regions was higher than would be expected bychance, that five-fixation scanpaths were consistentbetween the first and second viewings of a picture, and thatalthough actual fixation sequences were more similar toeach other than to model-predicted sequences, the modeldid perform better than chance. The model is good but notperfect, and we have now started to explain some of thevariability in performance by taking into account the priorknowledge of the observer who is inspecting the images.Humphrey and Underwood [31] compared viewers withspecialist domain knowledge inspecting images fromwithin their area of interest against viewers with a verydifferent area of interest. They were undergraduatesenrolled on specific courses. We recruited engineers andhistorians and presented all participants with the same setof images, some of which showed engineering plant, withmotors, pipes, valves, etc., and others that showed artefactsof the American Civil War such as uniforms and insignia,military equipment, domestic tools from the era, etc. (thesestudents had recently completed a module on the CivilWar). Both groups of domain experts saw both groups ofimages in an eye-tracking experiment with a similar designto that used by Foulsham and Underwood [30]. Accuracyscores on the recognition test confirmed the special interestsof the two groups of viewers—engineers performedbest with engineering pictures and historians performedbest with the Civil War pictures. As well as comparingindividual fixation locations against those predicted by thesaliency map model, we again compared scanpaths atencoding against those recorded at recognition, and againstthose predicted by the model on the basis of the five mostsalient locations. The model predicted the locations offixations, but only for viewers looking at pictures in theother domain of interest. When engineers looked at engineeringpictures, salient objects did not attract theirfixations, but when they looked at Civil War pictures theybehaved as the model predicted. The same pattern held forthe historians: within-domain they were resistant to theeffects of visual saliency, but when looking at picturesfrom another specialist domain they looked at the bright,coloured objects. Neutral pictures formed a third set ofimages, and showed outdoor and indoor scenes, and fixationson these images were similar to those on otherdomainimages. Both groups were more likely to look at asalient region of a neutral scene than at a salient region in apicture from their own domain. We also tested a group ofviewers from a third domain of interest—individuals withno special knowledge of engineering or the American CivilWar—and their fixations on all three types of pictures wereuniform and resembled the fixations of specialists lookingat pictures from the domain of the other specialists.

IV. CONCLUSIONS

This paper proposes new techniques for improving accuracyof medical image retrieval by representing image contentat an intermediate level local visual concept level. Theintermediate level is higher than low-level visual featuresthat are traditionally used and a step closer to the high-levelsemantics in the image content. A visual concept is definedfor local image regions and an image may comprise of severalconcepts. The feature space is enhanced by exploitingthe correlations and structural relationships among thethese visual concepts. Using SVM-based training, the proposedimage representation schemes realize semantic abstractionvia prior learning when compared to the representationsbased on the low-level features. Experimental resultsvalidate the hypothesis and shows that the proposedrepresentation schemes improve overall retrieval accuracy.The saliency map model of attention predicts that whenviewers first inspect a picture it is predominantly the bottom-up visual characteristics of the image that guide theireye movements [8, 10, 13, 14]. The initial parsing of thescene is conducted in terms of variations in intensity,colour, and the orientation of edges, resulting in a saliencymap that identifies the regions that have maximum variationof these characteristics. Before there is any analysis ofthe meaning of the scene, the viewers' eyes are attracted tothe single most salient region. As the viewers' eyes moveto the second most salient region, a process of inhibition ofreturn suppress the high saliency weight of the first region,to prevent an immediate return to an already inspectedobject. The model accounts for some of the variation in thelocation of eye fixations [13, 15, 19–21, 30, 31], and so is aviable model of scene inspection. The model does notaccount for some patterns of eye fixations, however [6, 20,21, 23–25], and it is appropriate to review the circumstancesunder which the low-level purely visualcharacteristics of an image dominate eye guidance.The saliency map hypothesis was introduced here as apossible explanation of an inconsistency in laboratoryreports of the inspection of images containing unexpectedobjects. Incongruous objects attract early attention,implying that they have been at least partially recognized prior to fixation, but not in experiments where objectidentification is difficult. There are reports of an incongruencyeffect from studies where objects are isolated fromtheir backgrounds [5] and where objects are otherwisereadily discriminated from their backgrounds in colourphotographs [6, 21, 23, 27], but not when densely packedline drawings or greyscale photographs are used [7–9]. Thesaliency values of objects do not provide good discriminationbetween these groups of experiments, however,because highly salient objects do not attract attention anyfaster than inconspicuous objects [21]. Perhaps the problemhere is that in this experiment with colour photographs allobjects were easily identified. They did not need to becarefully scrutinised to determine what they were, and themore appropriate study would be to use greyscale photographs(with difficult object identification) and with highand low saliency target objects. Perhaps the objects incolour photographs are identified simply too easily for theirsaliency values to have any influence on their detectability.At the present time we do not have a good understanding ofwhy the incongruency effect appears in some experimentbut not others.Saliency does have an effect upon the inspection ofpictures of real-world scenes, with fixations tending to landon salient regions and with objects of interest tending tohave higher saliency values. The effect upon eye fixationshas been reported in experiments in which participants aregiven ‘‘free viewing’’ instructions, in which the purpose ofinspection is to look at the image to comply with therequest from the experimenter [19], and in experiments inwhich the participants inspect images in preparation for arecognition memory test in which they will later declarewhether other pictures have previously been seen in theexperiment [20, 21, 30, 31]. There are circumstances inwhich visual saliency has little or no influence in theinspection of these pictures. First, if the viewer is searchinga picture to determine whether a specified target object ispresent [20, 21]; second, if the viewer is comparing twoimages to determine whether there are any differencesbetween them [23]; and third, if the viewer has specialized knowledge of the scene being shown [31]. There are twodistinct alternative explanations of this inconsistency, onewhich regards the effects of saliency as being a product ofthe task demands in the free-viewing and memory experiments,and one which regards saliency as being irrelevantto the task of viewers who know what they are looking for.These alternatives will now be considered briefly.The memory task requires viewers to look at a set ofpictures knowing that they will have to perform a discriminationtask. In the recognition test they see another setof pictures and they have to say whether each is ‘‘old’’ or‘‘new’’ according to whether it appeared in the first part ofthe experiment. One way to succeed in this task is to lookfor distinguishing features in each picture—something thatwould help identify it during test—and these features arelikely to be the bright, colourful objects, the salient objects.If a viewer adopts this strategy then it is the salient objectsthat will attract attention. A memory task performed in thisway would show effects of the saliency variations in animage not because the saliency map is used to guideattention in picture perception, but because the viewers arelooking for some features that would help them discriminatebetween pictures in a laboratory recognition test.

REFERENCES

[1] H. M¨uller, N. Michoux, D. Bandon, and A. Geissbuhler,“A Review of Content-Based Image Retrieval Systemsin Medical Applications Clinical Benefits andFuture Directions,” *International Journal of MedicalInformatics*, vol. 73, pp. 1–23, 2004.

[2] R.H. Choplin, J.M. Boehme, and C.D. Maynard,“Picture archiving and communication systems: anoverview,” *RadioGraphics*, vol. 12, pp. 127–129,1992.

[3] Y. Liua, D. Zhang, G. Lu, and W.Y. Ma, “A survey ofcontent-based image retrieval with high-level semantics,”*Pattern Recogntion*, vol. 40, pp. 262–282, 2007.

[4] W. Hsu, S. Antani, L.R. Long, L. Neve, and G.R.Thoma, “SPIRS: a Web-based Image Retrieval SystemFor Large Biomedical Databases,” *InternationalJournal of Medical Informatics*, vol. 78, pp. 13–24,2008.

[5] H. M¨uller, T. Deselaers, E. Kim, C. Kalpathy, D.Jayashree, M. Thomas, P. Clough, and W. Hersh,“Overview of the ImageCLEFmed 2007 Medical Retrievaland Annotation Tasks,” *8th Workshop of theCross-Language Evaluation Forum (CLEF 2007),Proceedings of LNCS*, vol. 5152, 2008.

[6] T.M. Lehmann, B.B. Wein, J. Dahmen, J. Bredno, F.Vogelsang, andM. Kohnen, “Content–based image retrieval in medical applications-A novel multi-step approach,”*Proc SPIE*, vol. 3972, pp. 312-320, 2000.

[7] R.B. Yates and B.R. Neto, *Modern Information Retrieval*,1st ed., AddisonWesley, 1999.

[8] V. Vapnik, *Statistical Learning Theory*, New York,NY, Wiley; 1998.

[9] T.F. Wu, C.J. Lin, and R.C. Weng, “Probability Estimatesfor Multi-class Classification by Pairwise Coupling,”*Journal of Machine Learning Research*, vol. 5,pp. 975–1005, 2004.

[10] K. Fukunaga, *Introduction to Statistical PatternRecognition*, Boston, 2nd edition: Academic Press;1990.

[11] S.F. Chang, T. Sikora, and A. Puri, "Overview of theMPEG-7 standard," *IEEE Trans CircSyst Video Technology*,vol. 11, pp. 688–695, 2001.

[12] Murphy GL, Wisniewski EJ. Categorizing objects in isolation andin scenes: what a superordinate is good for. J ExpPsychol LearnMemCogn. 1989;15:572–86.

[13] Davenport JL, Potter MC. Scene consistency in object andbackground perception. Psychol Sci. 2004;15:559–64.

[14] Mackworth NH, Morandi AJ. The gaze selects informative detailswithin pictures. Percept Psychophys.1967;2:547–52.

[15] Loftus GR, Mackworth NH.Cognitive determinants of fixationlocation during picture viewing. J ExpPsychol Hum PerceptPerform. 1978;4:565–72.

[16] Underwood G, Humphreys L, Cross E. Congruency, saliencyand gist in the inspection of objects in natural scenes. In: vanGompel RPG, Fischer MH, Murray WS, Hill RL, editors. Eyemovements: a window on mind and brain. Oxford: Elsevier;2007. p. 561–77.

[17] De Graef P, Christiaens D, d'Ydewalle G. Perceptual effects ofscene context on object identification. Psychol Res. 1990;52:317–29.

[18] Henderson JM, Weeks PA, Hollingworth A. The effects ofsemantic consistency on eye movements during scene viewing. JExpPsychol Hum Percept Perform. 1999;25:210–28.

[19] Gareze L, Findlay JM. In: van Gompel RPG, Fischer MH,Murray WS, Hill RL, editors. Eye movements: a window on mindand brain. Oxford: Elsevier; 2007. p. 617–37.

[20] Findlay JM, Walker R. A model of saccade generation base onparallel processing and competitive inhibition.Behav Brain Sci.1999;4:661–721.

[21] Torralba A, Castelhano MS, Oliva A, Henderson JM. Contextualguidance of eye movements and attention in real-world scenes:the role of global features on object search. Psychol Rev.2006;113:766–86.

[22] Navalpakkam V, Itti L. Modeling the influence of task onattention. Vision Res. 2005;45:205–31.

[23] Itti L, Koch C. A saliency-based search mechanism for overtand covert shifts of visual attention. Vision Res. 2000;40:1489–506.

[24] Koch C, Ullman S. Shifts in selective visual attention: towardsthe underlying neural circuitry. Hum Neurobiol. 1985;4:219–27.

[25] Itti L. Quantitative modelling of perceptual salience at human eyeposition.Vis Cogn. 2006;14:959–84.

[26] Elazary L, Itti L. Interesting objects are visually salient. J Vis.2008;8(3):3.1–15.

[27] Russell BC, Torralba A, Murphy KP, Freeman WT. LabelMe: adatabase and a web-based tool for image annotation. Int J ComputVis. 2008;77:157–73.

[28] Treisman AM, Gelade G. A feature-integration theory of attention.Cognit Psychol. 1980;12:97–136.

[29] Parkhurst D, Law K, Niebur E. Modelling the role of salience in theallocation of overt visual attention. Vision Res. 2002;42:107–23.

[30] Underwood G, Foulsham T, van Loon E, Humphreys L, Bloyce J.Eye movements during scene inspection: a test of the saliencymap hypothesis. Eur J Cogn Psychol. 2006;18:321–42.

[31] Underwood G, Foulsham T. Visual saliency and semantic incongruencyinfluence eye movements when inspecting pictures.Q J Exp Psychol. 2006;59:1931–49.

[32] Underwood G, Jebbett L, Roberts K. Inspecting pictures forinformation to verify a sentence: eye movements in generalencoding and in focused search. Q J Exp Psychol. 2004;57A:165–82.

[33] Underwood G, Templeman E, Lamming L, Foulsham T. Isattention necessary for object identification? Evidence from eyemovements during the inspection of real-world scenes.ConsciousCogn. 2008;17:159–70.

[34] Stirk JA, Underwood G. Low-level visual saliency does not predictchange detection in natural scenes. J Vis. 2007;7(10):3.1–10.

[35] Underwood J, Templeman E, Underwood G. Conspicuity andcongruity in change detection. Lect Notes Comput Sci. 2008.