New Zealand Journal of Agricultural Research, 2007, Vol. 50: 781–787
0028–8233/07/5005–0781

A model for image categorisation based on a biological visual mechanism

HE DONGJIAN
SHAO JUNMING*
GEN NAN
College of Information Engineering
Northwest A&F University
Yangling, Shaanxi, 712100
China

YANG QINLI
College of Resources and Environment
Northwest A&F University
Yangling, Shaanxi, 712100
China

*Author for correspondence: xinyuanwo@yahoo.com.cn

Abstract For integrating a visual attention mechanism and object recognition in the visual cortex we propose a novel biologically-motivated computational model for image categorisation. We first extract the focus of attention using an image-driven, bottom-up attention model and then adjust it according to the principles of whole effect and centre preference. After that, we obtain the region of interest, depending on the characteristics of object spatial proximity and object similarity. Based on this we compute a set of position- and scale-invariant C2 features and finally pool them into the standard classifier to achieve image categorisation. We test our model on an image database used in SIMPLIcity. The results suggest that our model can not only classify images effectively under various complex "clutters" but also that it needs only a few training samples.

Keywords image categorisation; region of interest; visual attention; visual cortex

A07092; Online publication date 25 January 2008
Received and accepted 10 August 2007

INTRODUCTION

In the last two decades, along with an improvement in image segmentation technology and the appearance of new approaches for machine learning, image classification and recognition technology has made considerable progress. These traditional approaches mostly use local features of the image such as colour, shape and texture, which can perform well on a rigid image like a car registration plate (Parisi et al. 1998), a face (Rong et al. 2004) and medical images (Yezzi et al. 1997). However, it is difficult to classify the different kinds of images having various complex backgrounds using these local features. Meanwhile, the more complex the backgrounds become, the lower the performance of the image classification and recognition. At present, increasing numbers of researchers are focusing on generic image categorisation and recognition, which is crucial for image semantic retrieval, image understanding and computer vision. Compared with existing computer vision systems, humans and primates outperform them with respect to almost any measure. In particular, the high accuracy of primates in ultra-rapid object categorisation and rapid serial visual presentation tasks is remarkable (Serre 2007a).

Considerable numbers of experiments (Ishai et al. 1999; Kastner et al. 2000) suggest that visual information progresses along two parallel hierarchical streams when visual information enters the primary visual cortex via the lateral geniculate nucleus. One is the "dorsal stream", also named as the "what" stream, which is primarily concerned with spatial localisation and directing attention and gaze towards objects of interest in the scene. The other is the "ventral stream", also called the "where" stream, which is concerned mainly with recognition and identification of visual stimuli. The two streams are not independent but interact in several higher-function areas of the visual cortex such as the prefrontal cortex (PFC), which plays an important role in modulating, via a feedback, the two processing streams thereby helping humans to recognise objects reliably and quickly (Miller 2000; Itti & Koch 2001).

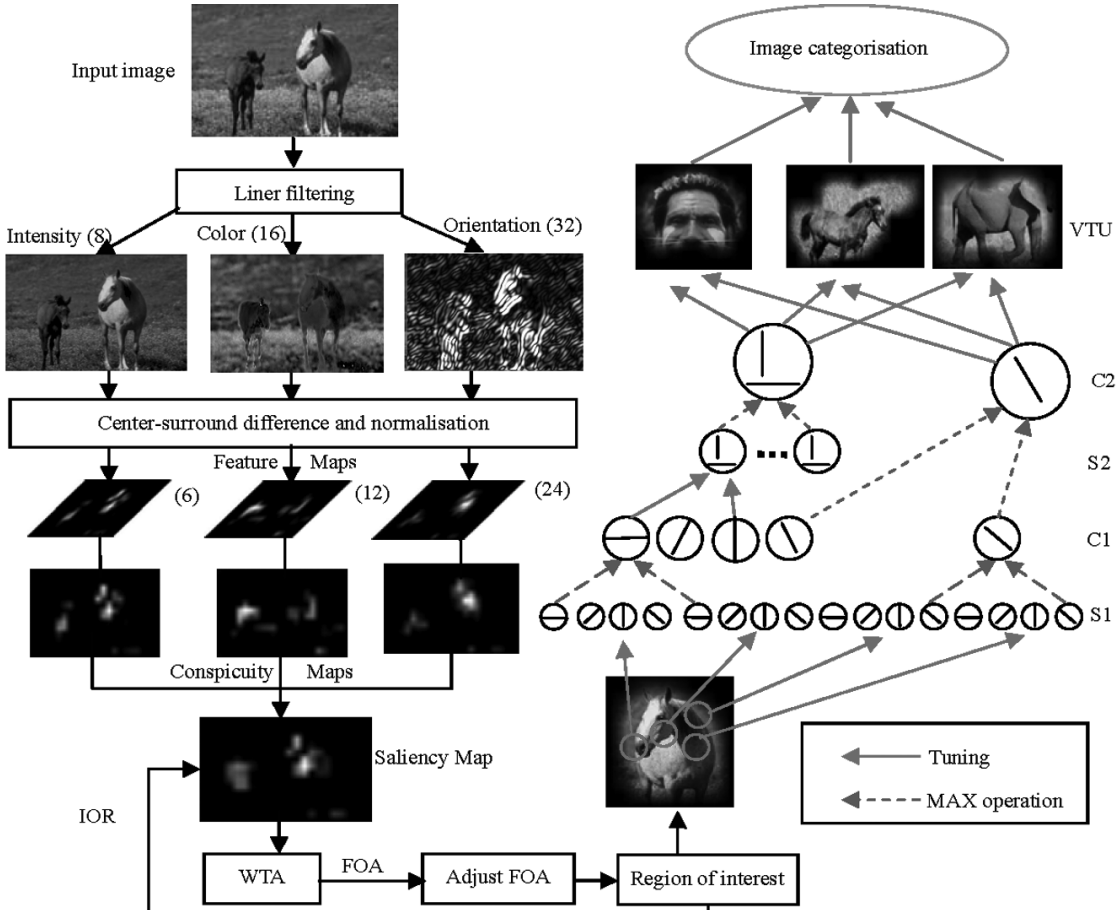

Fig. 1 Model for image categorisation based on biological visual mechanism.

Inspired by biology, we present here a model that combines a visual attention mechanism and object recognition in the visual cortex for image categorisation which imitates human and primate visual systems based on related achievements in visual psychology, nerve science and cognitive science.

MODEL

Considering the diversity and complexity of natural image backgrounds, effective image categorisation comes down to two key questions. The first challenge is how to extract objects from the images effectively and the other is how to obtain effective feature sets of image to achieve correct classification which can also perform well with different sizes, localisations and illuminations of the object in the image.

We propose a novel model framework for image categorisation as shown in Fig. 1. We first extract the focus of attention using a saliency map, image-based visual attention model and then adjust it according to the principles of whole effect and centre preference. The region of interest is obtained based on the focus of attention and the intrinsic characteristics of the object using the classical region grow approach. Based on this, the complex C2 features were computed by combining scales- and position-tolerant edge-detectors, which is motivated by a quantitative model of visual cortex proposed by Serre et al. (2007a). In the feedforward model, different levels correspond to different biological visual cortex areas. The first two layers correspond to the primary visual cortex (V1), the first visual cortical stage, which contains simple (S1) and complex (C1) cells. The S2 units located in V4 and C2 units in the inferotemporal (IT) cortex. From S1 units to C2 units, there is an increase

in invariance to position and scale and, in parallel an increase in the size of the receptive fields as well as in the complexity of the optimal stimuli for the neurons. Finally, the set of C2 features were put into a standard classifier to achieve classification in the view tuned unit (VTU) layer which corresponds to the prefrontal cortex (PFC) area.

DETAILED IMPLEMENTATION

Extraction of region of interest

In view of a lack of universal features for all kinds of natural images, most classical image segmentation or machine learning approaches do not perform well when extracting generic objects with various complex clutters. Inspired by human visual systems, we are always attracted by the visual saliency of the image when we first gaze upon an image without any pre-knowledge. The visual saliency is always presented by the contrast with the intrinsic features of the image, such as intensity, as well as colour and orientation. In this paper, we extract the region of interest to act as the object using a bottom-up attention model based on serially scanning a saliency map (computed from local feature contrasts) for salient locations in the order of decreasing saliency.

Focus of attention

We adopt the classic Itti's algorithm (Itti et al. 1998; Itti & Koch 2001) to extract the visual focus of attention, as the left half of Fig. 1 describes. The input image is first decomposed using Gaussian pyramids and early visual features like intensity, colour, and orientation are then extracted at each scale. The feature maps are obtained using Centre-Surround differences operation and Normalisation. These are combined into three conspicuity maps: intensity conspicuity map (FMI), colour conspicuity map (FMC) orientation conspicuity map (FMO) through across-scale addition. Finally, the conspicuity maps are normalised and summed into the saliency map (SM). The focus of attention is defined as the maximum of the saliency map, which is obtained by using a WTA (winner-takes-all) network.

Adjust the focus of attention

As a result of the various distracting backgrounds and noises in natural images, it is possible that the focus of attention we obtain is an isolated point. To extract the object effectively in a complex scene, it is necessary to adjust the focus of attention. We propose two simple principles to restrict the focus of attention. These are whole effect and centre preference. The whole effect refers to whether the object is popped out from the background, the global positions of the object should be salient on the salience map. The centre preference means that people tend to pay more attention to the central region of an image (Privitera & Stark 2000). Therefore, we use these two principles to filter the focus of attention.

(1) Whole effect: Calculating the mean of the salience value *avg* in the neighborhood $n \times n$ of the focus of attention on the salience map, if $avg < \varepsilon * Value_{FOA}$ (ε is a constant, $0 < \varepsilon < 1/5$, $Value_{FOA}$ is the salience value of the focus of attention), then it demonstrates that the neighbourhood pixels of the focus of attention are not salient. Therefore, we view it as isolated point and discard it.

(2) Centre preference: We define the shortest distance between the focus of attention and the border of the image as d. If $d < \alpha$ * (width + height) then the focus of attention locates near the margin of the image and we discard it (α is a constant, $0 < \alpha < 1/20$, width and height are the image width and height respectively).

Shifting the attention

Shifting the attention is implemented by the inhibition of a return mechanism. This has been widely observed in human psychophysical experiments. The traditional approach for implementation of inhabitation of return (IOR) inhibits just the single neuron in the saliency map at the currently attended location. Actually, however, the biological IOR has been shown to be object-bound (Gibson & Egeth 1994; Ro & Rafal 1999). This means that when we shift the attention, it should track and follow moving objects. Therefore, it seems feasible to obtain an interesting target when we are shifting to another different and distant salient location. According to two intrinsic characteristics of the object (spatial proximity and similarity preference (Koch & Ullman 1985)), we extract the region of interest by using the focus of attention and the feature map. We then implement the attention shift by inhibiting these neurons among the object-bound. The algorithm of extracting the region of interest is as follows:

(1) After adjusting the focus of attention, we find the feature map, which has the maximum salient value of the location:

$$FM_{\mathrm{T}} = \max_{\Omega \in \{I,C,O\}}^{Loc=FOA} FM_{\Omega}$$

(2) We convert the feature map FM_T into a binary image, and then carry on the close morphology operation. After that, we take the focus of attention as the seed point to make region growth so as to form the marked map of the region of interest M_B.

(3) We convolve the mark map with the three channels of original image respectively, extracting the region of interest: $ROI = M_{SRC} * M_B$

Image categorisation based on the set of ROI-C2 features

It is important for us to extract effective features to continue image categorisation after we have obtained the region of interest. Selective and invariance are two key criteria for this. Selective means these features can distinguish different kinds of images, meanwhile invariance means they excel in recognising the same kind of image with respect to image transformation. Risesenhuber & Poggio (1999) proposed a hierarchical model (HMAX) in 1999, which imitates object recognition in the cortex. Recently, Serre et al. (2007a,b) extracted the scale- and position-invariant C2 features (SMF) by extending the HMAX and applied them to natural images. These results demonstrate that the C2 feature set achieves good performance, which surpasses the best categorisation system currently available. In this paper, we extracted the C2 feature in the region of interest, and view it as the object. This hierarchical model is divided into five layers, S1, C1, S2, C2 and VTU. The first four layers are used to extract the C2 feature while the fifth, the VTU layer is responsible for image classification. The implementation is as follows.

S1 Applying to the input image a battery of Gabor filters, we adjusted these filter parameters to match the V1 parafoveal simple cells. Finally we obtained 16 × 4 activation maps in four orientations $\theta \in \{0°,45°,90°,145°\}$and on eight bands (16 scales).

C1 The S1 activation maps were sub-sampled by taking the maximum over a grid with cells of size K × K first and the maximum between the two scale members second. The process is repeated independently for each of the four orientations and each scale band. The maximum response extracts the optimised features to achieve invariance.

S2 We first obtain the image prototype by unsupervised learning from the C1 activation maps. The training stage in this paper is as follows: First N patches of various sizes are extracted at random positions in each of the C1 activation maps for all orientations and stored as S2 initial prototypes. We assign a representative function $T(p_i)$ to each prototype, the initial value is set at 1. We update the representative function $T(p_i)$ by computing the extent of matching between each prototype and patch extracted from the C1 activation maps for each training image. Each patch is then assigned to the prototype with the highest match and the number of patches assigned to each prototype is counted as $n(p_i)$. The update of the representative function's formula is Eqn (1):

$$T_{t+1}(p_i) = \begin{cases} \alpha * T_t(p_i) & n(p_i) = 0 \\ \beta^{n(p_i)} * T_t(p_i) & n(p_i) > 0 \end{cases} \quad (1)$$

where $0 < \alpha < 1$ and $\beta > 1$. If $T_{t+1}(p_i)$ becomes lower than a threshold λ, then the image prototype is discarded and re-initialised to a new random patch and reset its representative function of 1. The ultimate prototypes are achieved through several iterations over all training images and finally view them as radical basis function (RBF) units of S2.

During the test stage For each of prototype P_i, extract all image patches x (at all positions) from C1 maps at a particular scale. The response r of the corresponding S2 unit is computed by:

$$r = \exp(-\frac{1}{2\sigma^2}\|x - p_i\|) \quad (2)$$

where σ defines the sharpness of the tuning.

C2 We finally obtain the scale- and position invariant C2 features by taking the maximum over all scales and positions for each S2 activation map. The result is a vector of N C2 values.

VTU The set of C2 features is put into a standard classifier to achieve image categorisation.

RESULTS AND ANALYSIS

We evaluated our model for image categorisation on the image data set using SIMPLIcity, which can be downloaded at http://wang.ist.psu.edu/docs/related/. The image dataset includes 10 distinct semantic categories such as horse, beach, food and flower, each with 100 images.

Experiment to extract ROI

We tested our algorithm for extracting the region of interest using an Intel Pentium IV 2.4 GHz CPU, memory 512 MB PC under the Matlab7.0 environment and obtained results as shown in Fig. 2. We tested all 10 kinds of images in the database and the accuracy of extracting the region of interest is shown in Table 1.

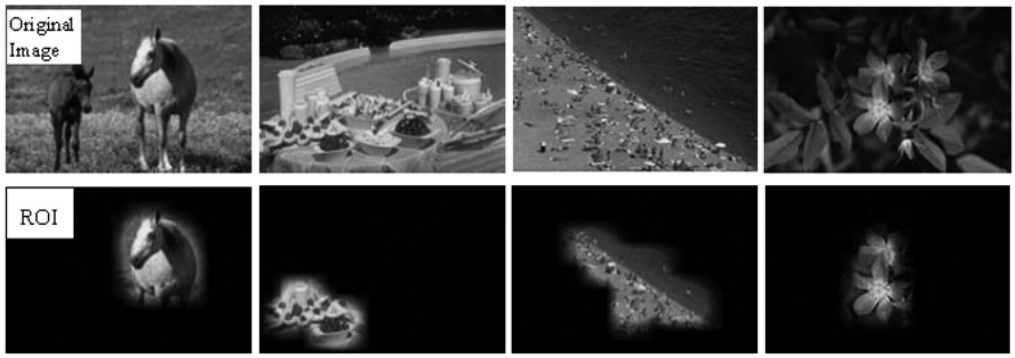

Fig. 2 Extract of region of interest (ROI).

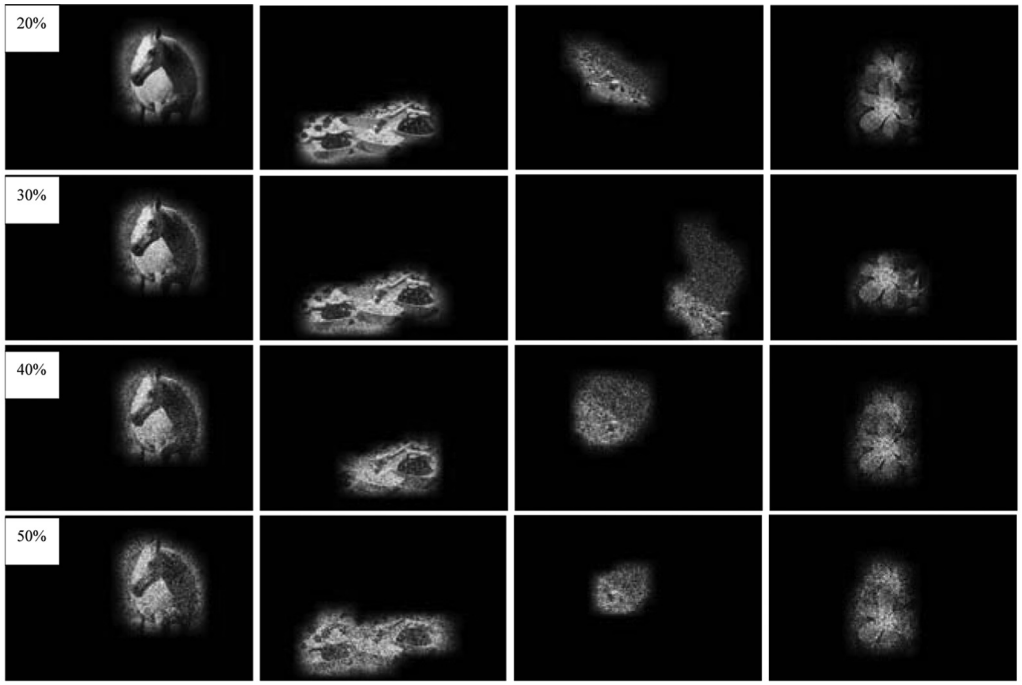

Fig. 3 Result of extracting region of interest based on different noise.

Table 1 Result of extracting region of interest.

Categories	Aboriginal	Beach	Building	Bus	Dinosaur	Elephant	Flower	House	Mountain	Food
Accuracy (%)	96	96	97	98	100	96	98	99	98	97

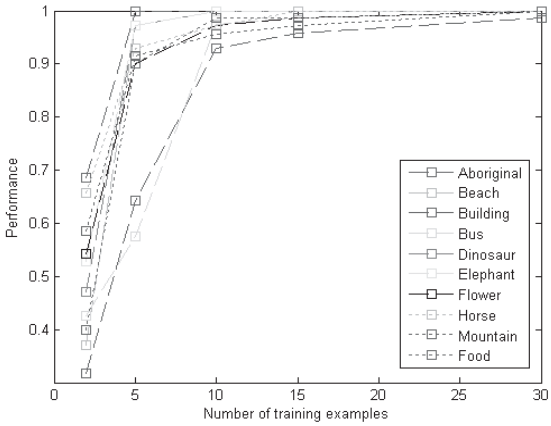

Fig. 4 Comparison between different numbers of training samples.

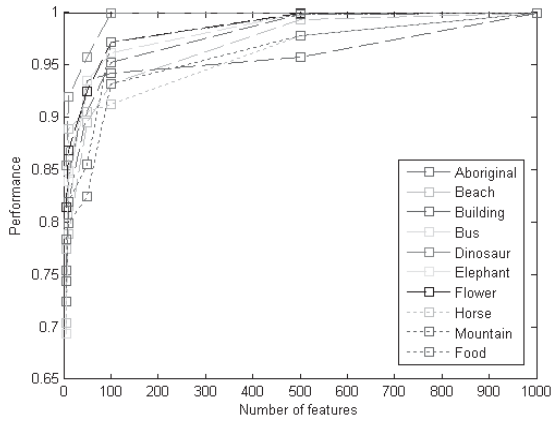

Fig. 5 Comparison between different numbers of ROI-C2 features.

For the sake of evaluating the influence of noise imposed on the algorithm of extracting the region of interest, we added 20–50% Gaussian noises to the distinct categories. The partial results of the region of interest extraction are shown in Fig. 3.

These results indicate that region of interest extraction using our algorithm performs well obtaining the interesting object effectively. Meanwhile, it can also extract generic objects under different clutters effectively and shows very strong robustness to noise.

Experiment of image categorisation based on ROI-C2 features

To evaluate the performance of our model effectively, we searched the same 10 categories of images (total 4211) in the website besides the SIMPLIcity image dataset. After we obtained the region of interest as described above, we converted all into grayscale images. In this experiment, we selected 30 images of each category at random to act as training images while the other images are test images. In addition, we selected 1000 as the number of features and used the K-nearest neighbour (KNN) algorithm and multiple-class standards support vector machine (SVM) to achieve image categorisation. Meanwhile, we compared with the classification results between our approach and directly extracting the C2 features from the original image. These results are shown in Table 2.

The results show that our model performed well and achieved higher accuracy of image categorisation than the system based on the C2 features.

Table 2 The result and comparison of categorisation using the region of interest (ROI-C2) feature. KNN, K-nearest neighbour; SVM, support vector machine.

			Accuracy of categorisation based on C2 features (%)		Accuracy of categorisation based on ROI-C2 features (%)	
Serial no.	Category	No.	KNN	SVM	KNN	SVM
1	Aboriginal	350	93.41	96.54	97.32	98.12
2	Beach	600	89.33	90.31	94.14	95.32
3	Building	700	91.36	90.12	95.57	96.45
4	Bus	500	91.14	94.32	96.55	96.25
5	Dinosaur	250	98.33	100.00	100.00	100.00
6	Elephant	500	94.23	97.34	98.14	99.03
7	Flower	800	93.33	91.67	98.57	98.24
8	Horse	500	94.33	95.45	96.45	96.53
9	Mountain	400	86.67	88.97	90.35	91.12
10	Food	311	91.14	93.12	94.14	95.47
Average accuracy of categorisation (%)			92.33	93.78	96.12	96.65

To evaluate our model for sensitivity to the number of training samples and features, we experimented on different number of training samples, num = {2, 5, 10, 15, 30} and different number of features N = {5, 10, 50, 500, 1000} respectively. The performances of the image categorisation are shown in Fig. 4, 5. We see clearly here that our model requires just a few training samples and it can perform very well.

CONCLUSIONS

In this paper, we describe a general model for image categorisation, which combines visual attention and object recognition in the visual cortex. The main contributions are:

(1) We propose a novel biologically-motivated model for robust automatic image categorisation, which imitates the human and primate visual system.
(2) Extending Itti's classic algorithm to extract focus of attention, we adjust the focus of attention according to the principles of whole effect and centre preference which improve the accuracy of extracting FOA by discarding some isolated points.
(3) We present an approach to obtain a region of interest depending on the characteristics of object spatial proximity and object similarity.

The proposed model has been validated by testing 5211 images with 10 semantically distinct categories. The experimental results indicate that our model performs well, with an average accuracy of image categorisation higher than 92.33% and this outperforms peer systems which use just the C2 features extracted directly from the image. How to integrate this model into the image semantic retrieval and scene understanding will be our main research objects for the future.

REFERENCES

Gibson BS, Egeth H 1994. Inhibition of return to object-based and environment-based locations. Perception and Psychophysics 55: 323–339.

Itti L, Koch C 2001. Computational modeling of visual attention. Nature Reviews Neuroscience 2(3): 194–203.

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

Ishai A, Ungerleider LG, Martin A, Schouten JL, Haxby AV 1999. Distributed representation of object in the human ventral visual pathway. Proceedings of the National Academy of Sciences 96: 9379–9384.

Kastner S, Ungerleider LG 2000. Mechanisms of visual attention in the human cortex. Annual Reviews Neuroscience 23: 315–341.

Koch C, Ullman S 1985. Shifts in selective visual attention: towards to underlying neural circuitry. Human Neurobiology 4: 219–227.

Miller EK 2000. The prefrontal cortex and cognitive control. Nature Reviews Neuroscience 1: 59–65.

Parisi R, Claudio EDD, Lucarelli G, Orlandi G 1998. Car plate recognition by neural networks and image processing Proceedings of the IEEE International Symposium on Circuits and Systems 3: 195–198.

Privitera CM, Stark LW 2000. Algorithms for defining visual regions-of-interest: comparison with eye fixations. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(19): 970–982.

Rong Xiao, Mingjing Li, Hongjiang Zhang 2004. Robust multi-pose face detection in images. IEEE Transactions on Circuits and Systems for Video Technology 14: 31–41.

Ro T, Rafal RD 1999. Components of reflexive visual orienting to moving objects. Perception and Psychophysics 61: 826–836.

Riesenhuber M, Poggio T 1999. Hierarchical models of object recognition in cortex. Nature Neuroscience 2(11): 1019–1025.

Serre T, Oliva A, Poggio T 2007a. A feedforward architecture accounts for rapid categorisation. Proceedings of the National Academy of Science 104(15): 6424–6429.

Serre T, Wolf L, Poggio T 2007b. Robust object recognition with cortex-like mechanisms. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(3): 411–426.

Yezzi A, Kichenassamy S, Kumar A, Olver P, Tannenbaum A 1997. A geometric snake model for segmentation of medical imagery. IEEE Transactions on Medical Imaging 16(2): 199–209.