

EE/CNS/CS 148 Lecture notes 12: How to utilize attention for object recognition and unsupervised learning

Ueli Rutishauser

Computational and Neural Systems, California Institute of Technology

May 8, 2003

1 Approach

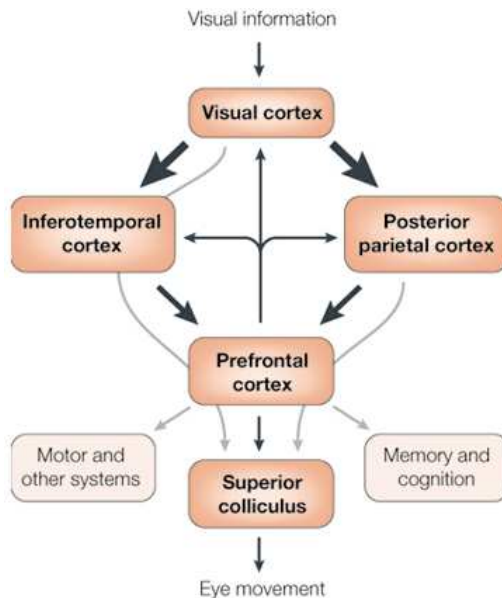


Figure 1: Neuronal mechanisms for the control of attention

Use principle from neuroscience: there is a dorsal and a ventral pathway in the visual system of primates (Figure 1, from [IK01]). The dorsal stream, which includes the posterior parietal cortex, are mainly concerned with spacial information (where pathway) whereas the ventral stream, which includes the inferotemporal cortex, is mainly concerned with recognition and identification.

From a computational viewpoint, the dorsal and ventral streams must interact, as scene understanding involves both recognition and spatial deployment of attention. One region where such interaction has been extensively studied is the prefrontal cortex (PFC). (from [IK01]).

Idea: Could this principle be used for computational vision?

2 Attention

The computational approach used for modelling visual attention is the approach of Itti et al. : [IKN98].

2.1 Gaussian pyramid

Gaussian pyramid: A number of sub-sampled images which are generated from a higher resolution by smoothening with a gaussian kernel. For a gaussian kernel of size 5x5:

$$Next(i, j) = \sum_{m=-2}^2 \sum_{n=-2}^2 w(m, n)prev(2i + m, 2j + n) \quad (1)$$

$Next(i, j)$ is the subsampled image (reduced in resolution and smoothened) and w is a 2-dimensional gaussian kernel of the form:

$$e^{-\frac{(x^2+y^2)}{2*\sigma^2}} \quad (2)$$

2.2 Center-surround differences

Center-surround (akin to visual receptive field) is implemented as differences between fine and coarse scale images in the gaussian pyramid. The center is a pixel at scale $c \in \{2, 3, 4\}$, and the surround is the corresponding pixel at scale $s = c + \delta$, $\delta \in \{3, 4\}$. This operation is denoted as Θ in the following description.

2.3 Extracting early visual features

Figure 2 illustrates the approach.

r , g and b are the red, green and blue channel of the input image. Obtain intensity image I by computing $I = (r + g + b)/3$. Use I to construct a gaussian pyramid $I(\sigma)$ with $\sigma = [0..8]$. This scales the image with the factors $2^0 \dots 2^8$.

Normalize r, g and b of the original image by $I(x, y)$ for every x, y . Only pixels with $\frac{1}{10}max(I()) < I(x, y)$ are normalized. For pixels where this condition does not hold set $r = g = b = 0$. This is to account for the fact that hue variations can not be perceived at low luminance. Define 4 color channels as :

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

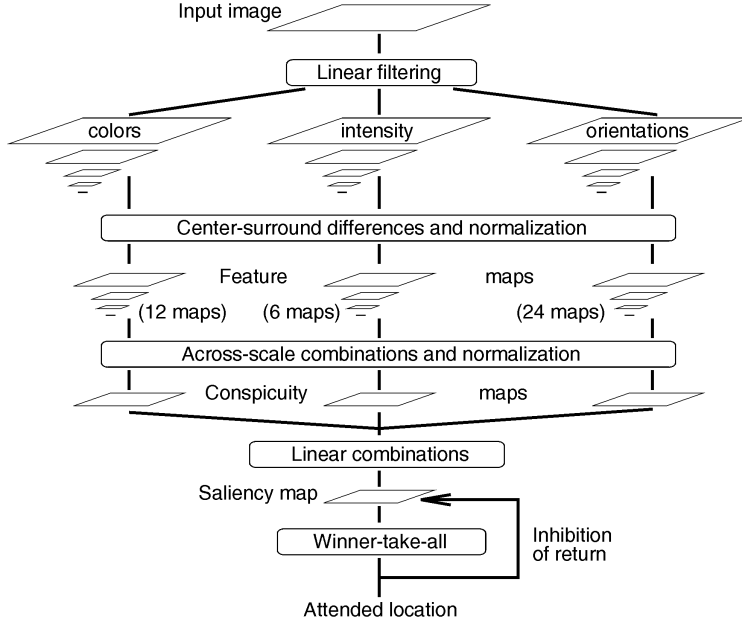


Figure 2: Attention model

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

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

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

Create gaussian pyramids $R(\sigma)$, $G(\sigma)$, $B(\sigma)$ and $Y(\sigma)$ from these tuned color channels.

2.4 Compute feature maps

Compute 6 intensity feature maps with $c \in \{2, 3, 4\}$ and $s = c + \delta$ with $\delta = \{3, 4\}$ as:

$$\mathcal{I}(c, s) = |I(c)\Theta I(s)| \quad (7)$$

Compute 12 color feature maps with $c \in \{2, 3, 4\}$ and $s = c + \delta$ with $\delta = \{3, 4\}$ as:

$$\mathcal{RG}(c, s) = |(R(c) - G(c))\Theta(G(s) - R(s))| \quad (8)$$

$$BY(c, s) = |B(c) - Y(c)\Theta(Y(s) - B(s))| \quad (9)$$

For the orientation feature maps oriented gabor pyramids $O(\sigma, \theta)$ with $\sigma \in [0..8]$ and $\theta \in \{0, 45, 90, 135\}$ are used.

Compute 24 orientation feature maps:

$$\mathcal{O}(c, s, \theta) = |O(c, \theta)\Theta O(s, \theta)| \quad (10)$$

These results in a total of 42 feature maps.

2.5 Compute saliency map

The saliency map is a joint representation of all feature maps. Because the different feature maps itself are not comparable they need to be normalized. This normalization operator is called $\mathcal{N}(\cdot)$. It's details are not shown here (refere to [IKN98]).

The feature maps are combined into 3 conspicuity maps I^* , C^* and O^* . These is achieved by across-scale addition of the feature maps at scale $\sigma = 4$. For this purpose, every feature map is reduced to scale $\sigma = 4$ and added point-by-point. Refere to [IKN98] for the exact formula.

Now the saliency map S can be computed as:

$$S = \frac{1}{3}(\mathcal{N}(I^*) + \mathcal{N}(C^*) + \mathcal{N}(O^*)) \quad (11)$$

The maximum of S defines the most salient point.

To prevent the saliency to focus on one single point only salient points are inhibited with an inhibition of return (IOR) method. IOR ([M.K00]) inhibits the currently most salient regions in the salient map such that approximatly 30-70ms (simulated time) after the jump to a salient point an other point becomes most salient (if there is one).

3 Shape estimation

Walther et al. ([WIR⁺02]) demonstrated how the saliency map can be used for shape estimation. The saliency map only tells which point in the image is salient, but not which object. Saliency map based shape estimation utilizes the information the pyramids used to construct the saliency map supply. This is done by modulating the original input image by a factor $M(x, y)$ which is in the range of $0 \leq M(x, y) \leq 1$. The values of M are extracted from the parts of the salient map whereas only the parts are used of the currently salient point (others are excluded). This is done by gradually lowering the threshold to follow the gradient down the pyramids that build the saliency map. The original image I is modulated as following:

$$I^*(x, y) = [(1 - m) + mM(x, y)]I(x, y) \quad (12)$$

m is a (static) modulation factor that can be modified to influence to global modulation strength.

Figure 3 illustrates the principle. In this figure, the image on the upper right is I^* , the one on the lower right is I . M is the 3'rd image in the upper row. The difference between M and I^* outside of the salient patch is accounted for by m .

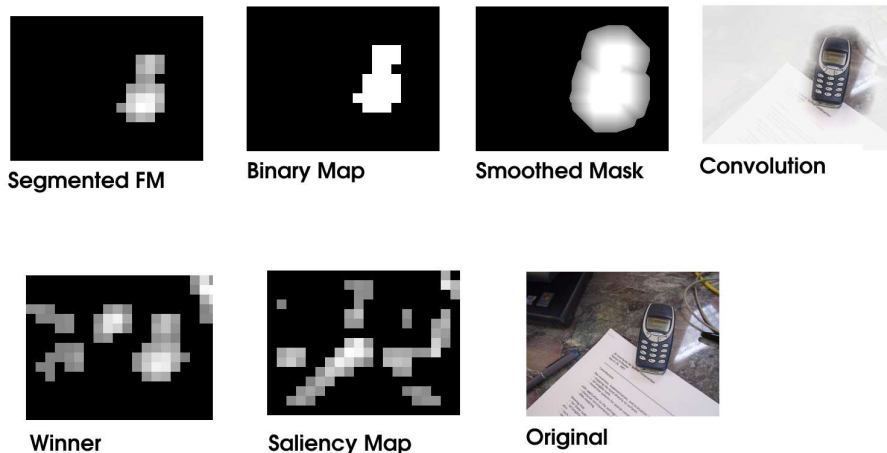


Figure 3: Saliency-map based shape estimation

4 Object recognition

Figure 4 illustrates the idea of using attention for object recognition. Both during training and recognition only the most salient points (most salient 5 in our case) are attended and used for training or recognition. This enables us to i) train unsupervised on complex scenes ii) train on objects in front of complex background (unsupervised) iii) reduce computational complexity during training and recognition. The main contribution of this approach is that it enables unsupervised learning of multiple objects from complex scenes. Experimental results as shown in [RWKP03] show significant increases in recognition rate and significant lower false positive rates with attention compared to without attention. Figure 5 shows an ROC curve to demonstrate the shift observed due to attention. Note that performance with a reasonable low false positives rate is below chance in the case without attention.

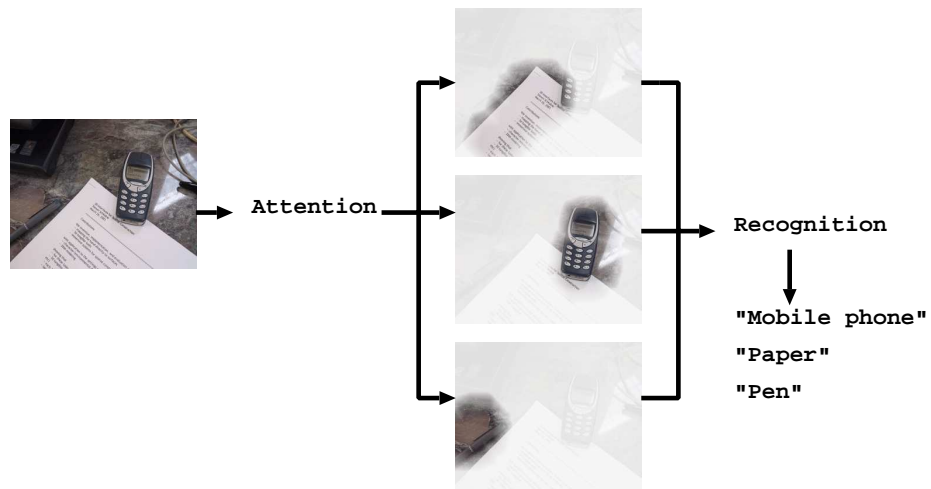


Figure 4: Utilizing attention for object recognition

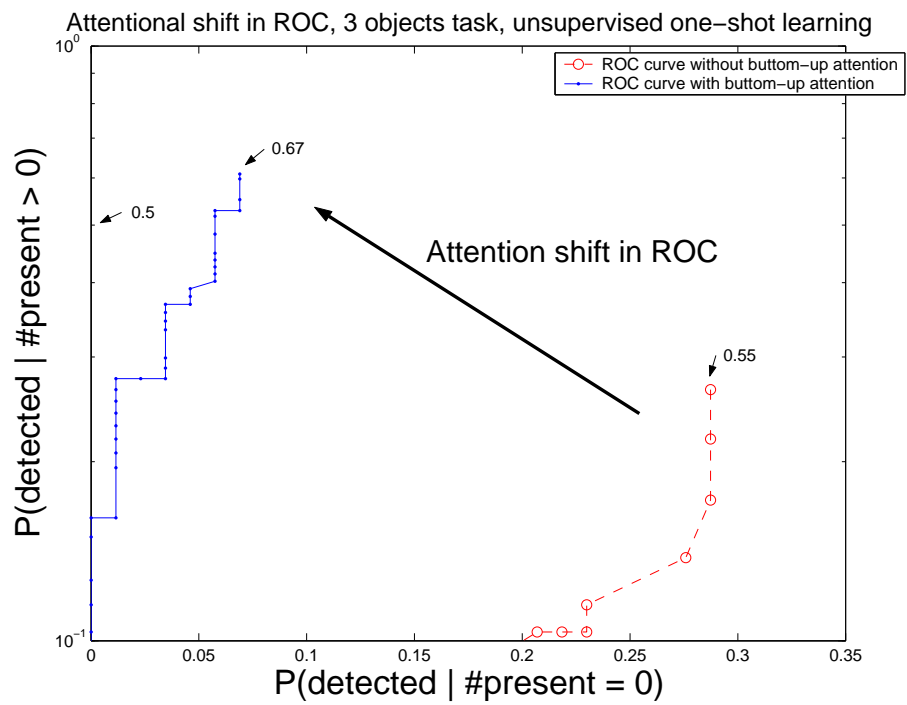


Figure 5: ROC that demonstrates left-shift due to attention

5 Literature

A nice overview of how visual attention can be modelled computationally is [IK01]. Rotation invariant pyramids, gaussian pyramids: [HSR⁺94]. An implementation of the bottom-up saliency as published by Itti et al. ([IKN98]) is available as GPL: [ILA]. The idea of using the saliency map for shape estimation was first published by Walther et al.: [WIR⁺02] Object recognition algorithm of Lowe: [Low99], [Low00]. The approach as shown in this notes will be published in [RWKP03].

References

- [HSR⁺94] H.Greenspan, S.Belongie, R.Goodman, P.Perona, S.Rakshit, and C.H. Anderson. Overcomplete steerable pyramid filters and rotation invariance. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 222–228, 1994.
- [IK01] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, vol. 2(No. 3):194–203, 2001.
- [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*, vol.20:1254–1259, 1998.
- [ILA] ilab c++ neuromorphic vision toolkit. <http://ilab.usc.edu/toolkit>.
- [Low99] David G. Lowe. Object recognition from local scale-invariant features. In *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume vol.2, pages 1150–1157, 1999.
- [Low00] David G. Lowe. Towards a computational model for object recognition in IT cortex. In *Biologically Motivated Computer Vision*, pages 20–31, 2000.
- [M.K00] Raymond M.Klein. Inhibition of return. *Trends in Cognitive Sciences*, vol. 4(No. 4):138–147, 2000.
- [RWKP03] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Utilizing attentional selection for object recognition in complex scenes. In *in press*, 2003.
- [WIR⁺02] Dirk Walther, Laurent Itti, Maximilian Riesenhuber, Tomaso Poggio, and Christof Koch. Attentional selection for object recognition – a gentle way. In H.H. Bülthoff, S.-W. Lee, T.A. Poggio, and C. Wallraven, editors, *Biologically Motivated Computer Vision. Second International Workshop, BMCV 2002, Tübingen, Germany, November 22-24, 2002. Proceedings. LNCS2525*, pages 472–479. Springer, 2002.