SELECTIVE ATTENTION BASED OPTICAL RECOGNITION OF HANDWRITTEN DIGITS

A. A. Salah, E. Alpaydın, L. Akarun

Department of Computer Engineering
Boğaziçi University, Istanbul Turkey
{salah,alpaydin,akarun}@boun.edu.tr

ABSTRACT

Human vision is mainly a serial process, in that we do not absorb the whole information content of the visual field at once, but selectively attend to locations that have low probability, and thus contain more information. In this paper, we model this process in a biologically plausible framework to achieve some of the desired properties of natural visual systems, and test it on a well-studied handwritten numeral recognition problem. We simulate the primitive, bottom-up attentive level with a saliency scheme, and the more complex, top-down associative level with Markov models. Our results are promising and indicate that such an approach can also be applied to other vision applications like face recognition.

1. INTRODUCTION

Most of the vision tasks solved by computers are attacked in a parallel fashion. Parallel recognition requires great computational resources, and it is counter-intuitive from the point of human vision. Primates solve the problem of object recognition and scene analysis in a serial fashion [2, 7], which is slower but less costly.

The biological structure of the eye is such that a high-resolution fovea and its low-resolution periphery provide data for recognition purposes. The fovea is not static, but is moved around the visual field in saccades, but these sharp, directed movements of the fovea are not random. The periphery provides low-resolution information which is processed to reveal the targets for the fovea [6]. The eye selects locations which contain more information, and a sense of economy is dominant in the whole design.

The selection depends on both bottom-up and top-down elements. If a location contains a pattern which has a low probability of occurring, the existence of this pattern gives us valuable information. This is the bottom-up part of the information.

There is also a top-down part, which is analogous to the expectation of the individual. The context-dependent, high-level information is integrated to the system at this level [8]. It is often useful to visualize the top-down information flow as a decision oriented process: The search for particular features, the elimination of possibilities (mostly through inhibitory mechanisms), and deciding between similar classes are the basic functions to consider.

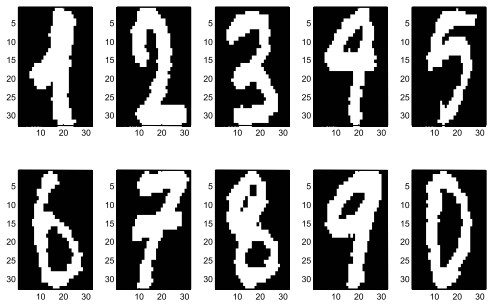

Figure 1: Sample Digits from Optdigits Database

Our previous work [1] used a recurrent multilayer perceptron to simultaneously learn both the foveal features and the class sequences. Other techniques are explored in the literature to apply the idea of selective attention to classification and analysis tasks citeitti, rimey. Our approach is to combine a feature integration scheme [10], with a Markov model [4, 9]. We first apply a saliency scheme to direct the fovea around the down-sampled input image. We call this the *attentive level*. The locations ('where' information), as well as the high-resolution fovea contents ('what' information) are put together in an associative level, where classification is performed.

Handwritten numeral recognition is a well-studied problem. In our database [11], there are ten classes with 3823 training and 1797 test cases. Each sample is a 32×32 binary image which is normalized to fit the bounding box (Fig. 1). Since these contain too much detail, we have down-sampled the digits to 8×8 and worked on this new set.

In the Section 2 and 3, the two major parts of our model, namely the attentive level and the associative level, will be explored, respectively. We present our results in Section 5, which is followed by our conclusions.

2. SALIENCY AND ATTENTIVE LEVEL

In the first step of the model, the bottom-up part of the vision system is simulated. The data provided by the bottom-up process is then used to train a system at the associative level, which does classification. The most important task

in designing the bottom-up system is to select the means of constructing a good saliency map. This map indicates the locations which need to be further inspected by the high-resolution fovea [5].

The saliency master map is constructed from various maps. Each map functions after the low-probability - high-saliency principle. For example edges are discontinuities in the texture, and therefore have low probability of occuring. We detect edge locations in the image and increase their saliency by using an edge map.

Another feature is the line orientations in the image, which are also detected by different primitive mechanisms, operating in coarse, intermediate and fine scales [3]. We have implemented twelve orientation detectors, in three levels and four angle orientations.

Since the orientation process also performs downsampling of the image, all levels of orientation maps are resized before adding them to the master orientation map. This has the effect of blurring on the whole map. In combining the feature maps, the orientation master map should have the largest influence.

We have used a gradient map to prevent the saliency conflicts in near-symmetric images. The map has a high intensity in its upper-left corner, and zero intensity in its lower-right corner. A similar phenomenon exists for the humans who are accustomed to reading from left to right, from top to bottom.

After the individual saliency maps are prepared, they are combined in a *saliency master map* to guide the attention process. The master map is constructed by passing a 3×3 low-pass Gaussian filter over the individual maps and taking the weighted sum as the saliency value at that point. The maximum of these values at any point is passed on to the master map.

Once a salient location is visited, its saliency is decreased, so that it is not visited repeatedly. This process has a biological counterpart: Once neurons attuned to detect a specific feature fire in the brain, they are temporarily inhibited. The saliency decrease is most intense at the visited location, somewhat lesser in the periphery of the fovea.

3. ASSOCIATIVE LEVEL AND CLASSIFICATION

The pre-attentive level, simulated by the movements of the fovea, provides us with two types of information, the position of the fovea with respect to the whole image and the content of the fovea. These are called the 'where' and 'what' information, respectively.

What we observe, and where we observe it are determined by a sequential process. This suggests the use of a Markovian Model in the associative level. To see whether the 'where' or 'what' information contain the means for a good classification, we have used Hidden Markov Models with different number of states. To compare their information content, visited locations and fovea contents were quantized separately using k-means clustering and used as observation vectors in different runs. Two types of information were later combined with a Markov model for the measurement of their joint performance (Fig. 2).

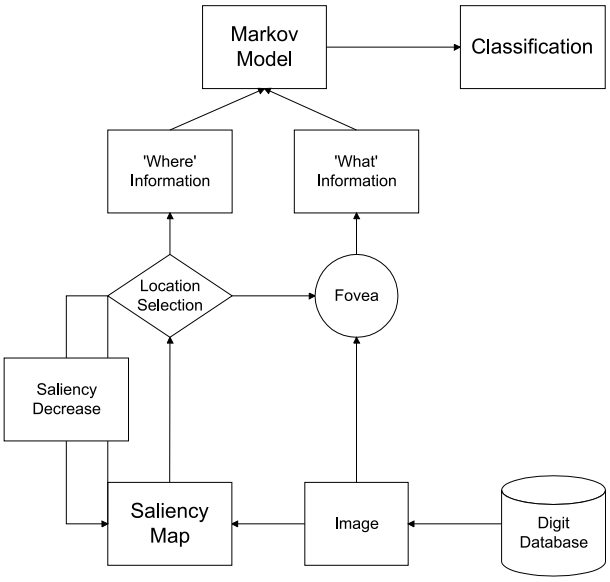

Figure 2: Classification with a Markov Model

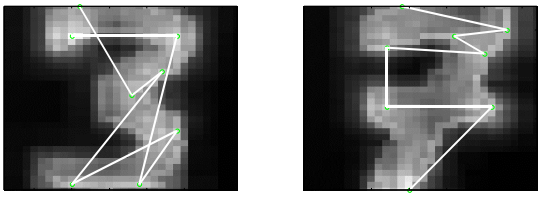

Figure 3: Saliency map and visited locations

The number of locations attended depends on the application. If a level of certainty is reached for classification, there remains no further need to attend the image. Furthermore, since the HMM makes use of the sequence information, attending more locations than necessary would increase the error and reduce the performance. Testing a number of values, we have decided that eight locations are enough for this application (Fig. 3).

The visited locations are not uniformly distributed over the image. This is to be expected, since the digits reside at the centre of the image, and the periphery usually consists of background. For this reason, a k-means clustering on 'where' information performs better than a uniform division of the image space into regions. However, regarding performance vs. biological plausibility (and therefore simplicity), we have consistently chosen simpler implementations. The results presented are obtained with 16 uniform regions.

In the case of fovea contents, the fovea size is a parameter that needs to be determined. Three different sizes are inspected: 5×5, 7×7 and 9×9. Note that of the 1024-pixel image, with 8 fixations, we need to visit at most 200, 392 and 648 pixels, respectively. Even with the small-

est fovea, the domain of possible observations contains 2^{25} possibilities, thus necessitating a quantization. We compare k-means clustering (Fig. 4) with a set of pre-assigned means that contain various features like corners, line-ends and junctions, that we believe are relevant in classification. The pre-assigned fovea contents perform slightly better than the k-means clustered means, which suggests some overlearning of the training data in the latter case.

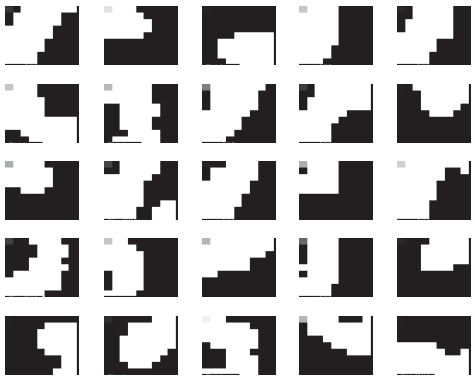

Figure 4: Sample means from k-means clustering on 9×9 fovea contents

Combining the 'where' and 'what' information can be done in various ways. We employ a Markov Model, where we have two choices of assigning the states and observations. Either the fovea positions are treated as states of the model and the fovea contents as observations, or vice versa. We test both schemes with all three fovea sizes.

4. RESULTS

We apply the saliency map to the training samples, record the visited locations and the fovea contents. With the clustered 'where' and 'what' information, we test three different Hidden Markov Models. The HMMs do not perform as good as expected, and we switch to a Markov model with visible states. Using both 'where' and 'what' information, we assume that the states and observations are both known to us.

The Markov model, where the fovea position and content are combined, results in a better performance than the HMM (Table 1). Two schemes are employed to minimize the variation due to the model. In the first scheme, the fovea contents are treated as states, and the fovea positions as observation vectors. The second scheme does it the other way around. Although the second scheme is more intuitive, and has a higher performance, the difference is not statistically significant.

The models are iteratively tested to observe the performance on training and test sets. The reported values are peak performances on the test set, just before over-learning starts. In each case, the correct classification percentages

Table 1: Accuracy(±standard deviation) on training/test sets

METHOD	PERFORMANCE
Only 'what' HMM	43.32(±8.51) / 36.34(±11.87)
Only 'where' HMM	49.59(±15.39) / 47.80(±17.34)
'Where' and 'what' HMM	57.49(±11.76) / 49.72(±17.11)
'What'-state, 'where'-observation Markov model	86.20(±7.11) / 63.66(±12.61)
'Where'-state, 'what'-observation Markov model	82.43(±8.82) / 68.05(±13.54)
Nearest mean classifier	87.63(±7.77) / 86.19(±9.09)

are given for both training and test sets. This value is calculated over the entire set, because there is a lot of variation among the classification of individual digits. For example digit '0' is classified easily, whereas '5' and '9' are frequently mixed up.

5. CONCLUSION

The selective attention mechanism exploits the fact that real images often contain vast areas of data that is insignificant from the perspective of recognition. A low-resolution, down-sampled image is scanned in parallel to find interesting locations through a saliency map, and complex features are detected at those locations by means of a high-resolution fovea. Recognition is done serially as the location and feature information is combined in time. This cuts down on complexity though it increases recognition time.

Our attempt to classify digits may be seen as a toy problem, since the ratio of the fovea area to the image is not high enough to demonstrate the benefits of the mechanism. We are planning to test our model on the problem of face recognition, where parallel processing is too cumbersome to use in a real-time application, and necessitates great computing power.

6. ACKNOWLEDGEMENTS

This work is supported by Boğaziçi University Research Funds 00A101D.

REFERENCES

[1] Alpaydın , E., "Selective Attention for Handwritten Digit Recognition," Advances in Neural Information Processing Systems 8, ed. D.S. Touretzky, M.C. Mozer, M.E. Hasselmo MIT Press, 771-777, 1996.

[2] Crick, F., and C. Koch, "Towards A Neurobiological Theory Of Consciousness," Seminars in the Neurosciences, 2, 263-275, 1990.

[3] Foster, D.H. and S. Westland, "Multiple Groups of Orientation-selective Visual Mechanisms Underlying

Rapid Oriented-line Detection," Proc. Royal Society London B 265, 1605-1613, 1998.

[4] Hacisalihzade, S. S., L. W. Stark and J. S. Allen, "Visual Perception and Sequences of Eye Movement Fixations: A Stochastic Modeling Approach," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 22, No. 3, May-June 1992.

[5] Itti, L., 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, No. 11, November 1998.

[6] Koch, C., and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," Human Neurobiology, 4:219-227, 1985.

[7] Noton, D. and L. Stark, "Eye Movements and Visual Perception," Scientific American, 224:34-43, 1971.

[8] Olshausen, B., C. Andersen, and D. Van Essen, "A Neural Model of Visual Attention and Invariant Pattern Recognition,' California Institute of Technology, Computation and Neural Systems Program, CNS Memo 18, August 6, 1992.

[9] Rimey, R. D. and C. M. Brown, "Selective Attention as Sequential Behavior: Modeling Eye Movements with an Augmented Hidden Markov Model," TR-327, Computer Science, University of Rochester, February 1990.

[10] Treisman, A. M., and G. Gelade, "A Feature Integration Theory of Attention," Cognitive Pscychology, vol.12, no.1, pp.97-136, Jan.1980.

[11] UCI Machine Learning Repository, Optdigits Database, prepared by E. Alpaydın and C. Kaynak, ftp://ftp.ics.uci.edu/pub/machine-learning-databases/optdigits.